Biostatisztika és informatika

5. előadás:
Becslés és megbízhatóság
2018. október 11.

Agócs Gergely

Források: – **Herényi L (2016): Statisztika és Informatika: 14. fejezet**
– Reiczigel J, Harnos A, Solymosi N (2014): Biostatisztika nem statisztikusoknak: 5. fejezet
– WolframMathWorld: Probability and Statistics: http://mathworld.wolfram.com/topics/ProbabilityandStatistics.html

Az előadás céljai

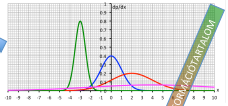
- A **becslés** célja, típusai és folyamata
  - minta és alapsokaság más néven populáció
  - **pontbecslés** és **intervallumbecslés**
  - minta, becslés, becsült és becslő érték
- Mik a „**jó becslés**" tulajdonságai?
  - torzítatlan, hatásos, konzisztens, elégséges
- **Standard hiba** (*SE*) megértése
- K**onfidenciaintervallum** (CI) megértése
  - konfidenciaintervallumok helyes értelmezése
  - konfidenciaintervallumok helyes feltüntetése
- Megtanulni egyes paraméterek **becslését**:
  - valószínűség (röv.: **valség**) más néven alapsokaságbeli arány
  - **várható érték** más néven alapsokaságbeli átlag
  - elméleti **variancia** (más néven szórásnégyzet) és elméleti **szórás**
- Adott SE-hoz **szükséges mintaelemszám** számítása
- **Excel**-függvények használata becsléshez

A becslés célja

Alapsokaság:
az összes lehetséges megfigyelés (általában végtelennek tekintett) halmaza. Minden vizsgálatunk erre a halmazra irányul.

Pl.: - testmagasság cm-ben
- szemszín
- terhességek száma
- sugárkárosodás foka
- ...

Eloszlásfüggvények (CDF, PDF, PMF)



Paraméterek (μ, σ^2, σ...)

Egyes **Valószínűségek**

De hogyan határozhatjuk meg ezeket?

A becslés célja

De hogyan határozhatjuk meg ezeket?

Példa: érmefeldobás

Modellalkotás
(pusztán elméleti megfontolások)

$P(\text{fej}) = 1/2$ & $P(\text{írás}) = 1/2$

- két elemi esemény, melyek
- azonos valségűek (= *feltesszük*, hogy az érme szabályos) és
- egyetemesek

Ezt a módszert csak **nagyon korlátozottan** tudjuk használni

Végezzünk megfigyeléseket

Ez azt jelenti, hogy az érmét ténylegesen feldobjuk, de persze nem végezhetjük el az összes lehetséges megfigyelést, mert:
- eleve lehetetlen (végtelen számú ismétlés)
- nem célszerű (pl. töröl teszt)
- sokba kerül (időbe, pénzbe...)

A becslés folyamata

Becslések típusai

Az emberek hány százaléka tartozik az Rh+ vércsoportba?

Nem vizsgálhatunk meg minden embert.

Vegyük az emberek egy részhalmazát (**minta**) és vizsgáljuk meg ezek elemeit.

Pontbecslés:

kb. 85%

csak egyetlen érték, melyet **becslő értéknek** nevezünk

Mennyire bízhatunk meg ebben a *számban*? **Fogalmunk sincs!**

Intervallumbecslés:

83% – 87% között

95%-os *megbízhatósággal*

tartomány (becslő érték és hiba) **+ valószínűség**

A **bizalom szintje**, melyet ebbe a *tartományba* vethetünk, kifejezve egy **valószínűséggel**

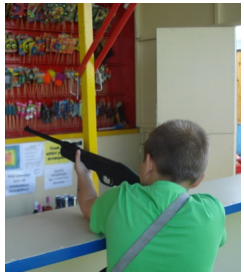
Pontbecslés

Elméleti (alapsokaság) értékek:

„CÉL"

becslő é.:

"LÖVÉS"



Pontbecslés

Elméleti (alapsokaság) értékek:

(diszkrét valségi változóra)

„CÉL"

- valség vagy arány (p_i)

- várható érték („alapsokaság átlaga") ($E(\xi)$ vagy μ)

$$E(\xi) = \mu = \sum_{i=1}^{n} p_i \cdot x_i$$

- elméleti variancia ($Var(\xi)$ vagy σ^2)

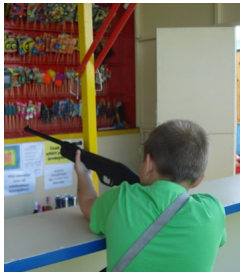
$$Var(\xi) = \sigma^2 = E\left[\left(\xi - E(\xi)\right)^2\right] = \sum_{i=1}^{n} p_i \cdot (x_i - \mu)^2$$

- elméleti szórás ($SD(\xi)$ vagy σ)

$$SD(\xi) = \sigma = \sqrt{Var(\xi)} = \sqrt{\sum_{i=1}^{n} p_i \cdot (x_i - \mu)^2}$$

becslő é.:

"LÖVÉS"



Pontbecslés

Elméleti (alapsokaság) értékek:
(diszkrét valószínűségi változóra)
„CÉL”

- valség vagy arány (p_i)

- várható érték („alapsokaság átlaga”) ($E(\xi)$ vagy μ)

$$E(\xi) = \mu = \sum_{i=1}^{n} p_i \cdot x_i$$

- elméleti variancia ($Var(\xi)$ vagy σ^2)

$$Var(\xi) = \sigma^2 = E\left[\left(\xi - E(\xi)\right)^2\right] = \sum_{i=1}^{n} p_i \cdot \left(x_i - \mu\right)^2$$

- elméleti szórás ($SD(\xi)$ vagy σ)

$$SD(\xi) = \sigma = \sqrt{Var(\xi)} = \sqrt{\sum_{i=1}^{n} p_i \cdot \left(x_i - \mu\right)^2}$$

Behelyettesítéses („plug-in”) becslő é.:
“LÖVÉS”

- relatív gyakoriság $\hat{p}_i = (k_i/n)$ ✓
  Excel: =DARABHATÖBB(*adat*)/DARAB2(*adat*)

- minta átlaga
  Excel: =ÁTLAG(*adat*)

$$\bar{x} = \sum_{i=1}^{n} \frac{k_i}{n} \cdot x_i = \frac{1}{n} \sum_{i=1}^{n} k_i \cdot x_i$$ ✓

- „plug-in” variancia (s^{**2})

$$s^{**2} = \sum_{i=1}^{n} \frac{k_i}{n} \cdot \left(x_i - \mu\right)^2 = \frac{1}{n} \sum_{i=1}^{n} k_i \cdot \left(x_i - \mu\right)^2$$ ✗

- „plug-in” szórás (s^{**}, *SD*)

$$s^{**} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} k_i \cdot \left(x_i - \mu\right)^2}$$ ✗

A „jó becslés” …

… torzítatlan

A valószínűség (p_i) pontbecslése relatív gyakorisággal ($\hat{p}_i$)
Tekintsük az imént tanult „plug in” becslő képletét:

$\hat{p}_i = (k_i/n)$ ✓

E becslés végtelen számú megismétlésével kapott becslő értékek átlaga maga a becsült érték (valség).

Excel: =DARABHATÖBB(*adat*)/DARAB2(*adat*)

A várható érték (μ) pontbecslése a mintaátlaggal ($\bar{x}$)

$$\bar{x} = \sum_{i=1}^{n} \frac{k_i}{n} \cdot x_i = \frac{1}{n} \sum_{i=1}^{n} k_i \cdot x_i$$ ✓

Excel: =ÁTLAG(*adat*)

Egy becslés torzítatlan, ha a végtelenszer megismételt becslésekkel kapott becslő értékek várható értéke megegyezik a becsült értékkel.

A „jó becslés” …

… torzítatlan

Az elméleti variancia (σ^2) pontbecslése

Tekintsük az imént tanult „plug-in” becslő (s^{**2}) képletét:

$$s^{**2} = \frac{1}{n} \sum_{i=1}^{n} k_i \cdot \left(x_i - \mu\right)^2$$ ✗

μ egy elméleti érték, melyet nem ismerünk. Helyette a **mintaátlagot** kell használnunk.

Friedrich **Bessel**
1784–1846

Helyettesítsük be a várható érték helyére a mintaátlagot:

$$s^{*2} = \frac{1}{n} \sum_{i=1}^{n} k_i \cdot \left(x_i - \bar{x}\right)^2$$ ✗

Ez a képlet a mintaátlagra minimális, nem a várható értékre. Ez **torzítást** okoz, a képlet **alábecsli** az elméleti szórást.

Az **n/(n–1)** korrekciós szorzóval (**Bessel’s korrekciónak** hívják) megszüntethető a torzítás:
(bizonyítás a jegyzetben, nem kell tudni)

$$s^2 = \frac{n}{n-1} \cdot \frac{1}{n} \sum_{i=1}^{n} k_i \cdot \left(x_i - \bar{x}\right)^2 = \frac{1}{n-1} \sum_{i=1}^{n} k_i \cdot \left(x_i - \bar{x}\right)^2$$ ✓

Excel: =VAR.M(*adat*)

A „jó becslés” …

… kevéssé torzít

Az elméleti szórás (σ) pontbecslése

A variancia becslő értékéhez hasonlóan a szórás “plug in”-képlete is torzít:

$$s^* = \sqrt{\frac{1}{n} \sum_{i=1}^{n} k_i \cdot \left(x_i - \bar{x}\right)^2}$$ ✗

Itt a Bessel-korrekciós faktor **n/(n–1)** használata
csak csökkenti de teljesen nem szünteti meg a torzítást:
(csak aszimptotikusan, vagyis végtelen nagy minta esetén;
oka: a gyökfüggvény aszimmetrikussága)

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

Excel: =SZÓR.M(*adat*)

Ezt a kevéssé torzító (de nem teljesen torzítatlan) becslést használjuk.

Pontbecslés

Elméleti (alapsokaság) értékek:
(diszkrét valségi változóra)
„CÉL”

- valség vagy arány (p_i)

- várható érték („alapsokaság átlaga”) ($E(\xi)$ vagy μ)

$$E(\xi) = \mu = \sum_{i=1}^{n} p_i \cdot x_i$$

- elméleti variancia ($Var(\xi)$ vagy σ^2)

$$Var(\xi) = \sigma^2 = E\left[\left(\xi - E(\xi)\right)^2\right] = \sum_{i=1}^{n} p_i \cdot \left(x_i - \mu\right)^2$$

- elméleti szórás ($SD(\xi)$ vagy σ)

$$SD(\xi) = \sigma = \sqrt{Var(\xi)} = \sqrt{\sum_{i=1}^{n} p_i \cdot \left(x_i - \mu\right)^2}$$

Behelyettesítéses („plug-in”) becslő é.:
“LÖVÉS”

- relatív gyakoriság $\hat{p}_i = (k_i/n)$ ✓
  Excel: =DARABHATÖBB(*adat*)/DARAB2(*adat*)

- minta átlaga
  Excel: =ÁTLAG(*adat*)

$$\bar{x} = \sum_{i=1}^{n} \frac{k_i}{n} \cdot x_i = \frac{1}{n} \sum_{i=1}^{n} k_i \cdot x_i$$ ✓

- tapasztalati variancia (s^2)
  Excel: =VAR.M(*adat*)

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} k_i \cdot \left(x_i - \bar{x}\right)^2$$ ✓

- tapasztalati szórás (s, SD)
  Excel: =SZÓR.M(*adat*)

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

Pontbecslés

1. feladat: Meg szeretnénk becsülni a kékszeműek arányát, valamint a testmagasság várható értékét, elméleti varianciáját és elméleti szórását a SOTÉ-s gólyák (elsőéves hallgatók) körében.

Egy 15-elemű mintát vettünk (lásd a táblázatot). Használjuk az Excel-t az elméleti paramterek pontbecslésre!

p(kék szem) ≈ $\hat{p}$ =4/15 = 0.2667 = 26,67%

μ(testmagasság) ≈ $\bar{x}$ =ÁTLAG(*adat*) = 170 cm

σ^2(testmagasság) ≈ s^2 =VAR.M(*adat*) = 107,5 cm^2

σ(testmagasság) ≈ s =SZÓR.M(*adat*) = 10,4 cm

Megfigyelés sorszáma	Szemszín kék?	Magasság (cm)
1	HAMIS	163
2	HAMIS	153
3	HAMIS	152
4	HAMIS	158
5	HAMIS	167
6	IGAZ	184
7	IGAZ	165
8	HAMIS	184
9	IGAZ	175
10	HAMIS	167
11	HAMIS	178
12	HAMIS	168
13	HAMIS	173
14	IGAZ	178
15	HAMIS	180

A „jó becslés” …

… hatásos

Egy becslés hatásos, ha szórása (az ún. *standard hiba*, *SE*) minimális.

- A megismételt becslések egy sor becslő értéket eredményeznek, amelyek különböznek egymástól a **mintavétel véletlenszerűsége** miatt (**véletlen hiba**)
- Tehát **a becslő érték maga is egy véletlen változó**, melynek van elméleti eloszlása, várható értéke, elméleti szórása stb.
- A becslő érték elméleti szórását standard hibának (*standard error*, *SE*) nevezzük

A következő diákon megtanuljuk hogyan kell kiszámolni az alábbi két becslő érték standard hibáját.

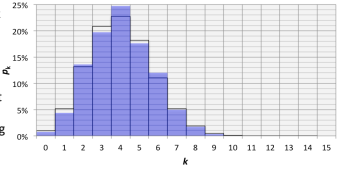
Arányok elméleti eloszlása:
Binomiális eloszlás

Átlagok elméleti eloszlása:
Student-féle t-eloszlás (*df* = *n* – 1)

Arány standard hibája

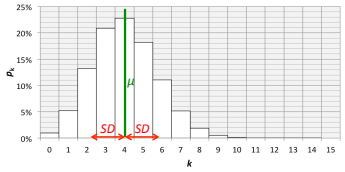
1. példa: Meg szeretnénk becsülni a kékszemű hallgatók arányát a gólyák mint alapsokaság között. Egy n = 15 elemű mintát veszünk, melyben a kékszemek száma k = 4-nek adódik.Mármegtanultuk, hogy az alapsokaságon belüli arány (azaz valség, p) becslő értéke a relatív gyakoriság ($\hat{p} = k/n$), mely esetünkben 4/15 = 0,2667 . Mekkora a becslés hibája?

Tegyük fel, hogy a becsült érték megegyezik a becslő értékkel: p = 4/15. Számítógépes szimulációt használva vegyünk 800 ilyen mintát és készítsünk relatív gyakorisági hisztogramot a kékszeműek mintánkénti k számából (kék oszlopok). Ez elég jól közelíti a megfelelő elméleti eloszlást: a binomiális eloszlást (fekete vonalrajz).



Arány standard hibája

A korábbi előadásokban már megtanult képletekkel számoljuk ki ezen binomiális eloszlás paramétereit (μ, σ^2 és σ).



Várható érték:
$\mu = np = 4$

Elméleti variancia:
$\sigma^2 = np(1-p) = 44/15 = 2{,}933$

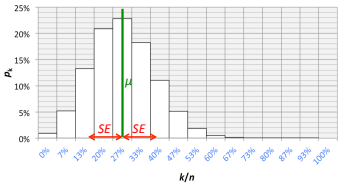
Elméleti szórás:

$SD = \sqrt{np(1-p)} = 1{,}713$

Arány standard hibája

Mivel arányokat becslünk, ezért a k változót k/n aránnyá kell alakítanunk, vagyis a vízszintes tengelyt át kell skáláznunk n-nel való osztással.

Ugyanezt kell tennünk a binomiális eloszlás μ és σ paramétereivel is.



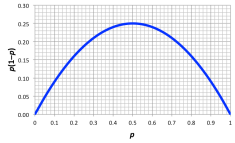
Várható érték
$\mu = p = 4/15$

Elméleti szórás = az arány standard hibája

$SE = \frac{\sqrt{np(1-p)}}{n} = \sqrt{\frac{p(1-p)}{n}} = 0{,}1142$

$SE_{arány} = \sqrt{\frac{p(1-p)}{n}} \approx \sqrt{\frac{\frac{k}{n}\left(1-\frac{k}{n}\right)}{n}}$

Arány standard hibája



Theoretical Distribution of Proportions:
Binomial Distribution
(normalized to sample size)

$\max\left(SE_{arány}\right) = \frac{1}{\sqrt{4n}}$

Mekkora az n-elemű mintából számolható arányok lehető legmagasabb standard hibája?

A $p(1-p)$ szorzat akkor maximális, ha $p = 1-p = 0{,}5$. Ez esetben $p(1-p) = 0{,}25$ így a SE :

$\max\left(SE_{arány}\right) = \sqrt{\frac{0{,}5(1-0{,}5)}{n}} = \sqrt{\frac{0{,}25}{n}} = \frac{1}{\sqrt{4n}}$

Ha egy tanulmány ugyanazon mintából (vagy azonos elemszámú mintákból) becsült több valséget is tartalmaz, akkor gyakran csak a maximális standard hibát adják meg a fenti képlet segítségével.

Pl. egy n = 100-elemű mintából számolt arány esetén a SE legfeljebb 0,05 lehet.

Arány standard hibája

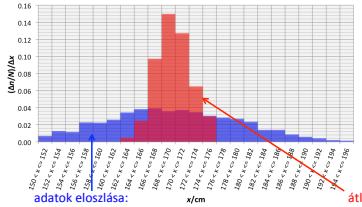
2. feladat: Mi a sarlósejtes anémia (SC) génhordozóinak *prevalenciája* (népességen belüli aránya) és a becslés standard hibája Nigériában, ha 172 vizsgált emberből 43 hordozta a betegség génjét?
Hordozók mintabeli aránya: $\hat{p}$ (SC) = k(SC)/n = 43/172 = 0,25 . A hiba az imént tanult képlettel számolható: $SE_{arány} = \sqrt{\frac{p(1-p)}{n}} \approx \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0{,}25\cdot(1-0{,}25)}{172}} = 0{,}033 = \underline{3{,}3\%}$

3. feladat: Szeretnénk egy vizsgálatban megbecsülni egyes krónikus betegségek budapesti prevalenciáját. Milyen mintaméretet javasolnál, ha nem akarjuk, hogy a becslés hibája az 1%-ot meghaladja?
Mivel konkrét valséget vagy arányt nem ismerünk, a „legrosszab esetnek" tekinthető 0,5-ös valséggel kell számolnunk, melynek adott mintaméret esetén a legnagyobb a standard hibája: $SE_{arány} \le \frac{1}{\sqrt{4n}}$
A képletet átrendezve kapjuk n-t: $n \le \frac{1}{4\left(SE_{arány}\right)^2} = \frac{1}{4\cdot 0{,}01^2} = \underline{2500}$

Vagyis egy 2500-as mintaelemszám garantálja, hogy egy esetleges 0,5-ös prevalencia becslési hibája sem haladja meg az 1%-ot. (A ≤ jel azt jelenti, hgoy 0,5 alatti vagy feletti prevalenciák esetén kisebb minták is elégségesek lennének az SE 1% alatt tartására.)

Átlag standard hibája

2. példa: Meg szeretnénk becsülni a gólyák testmagasságát. Vettünk egy n = 15 hallgatóból álló mintát, melyből az átlag $\bar{x}$ = 170 cm-nek, korrigált szórás s = 10 cm-nek adódott. Számítógépes szimulációval állítsuk elő az adateloszlást és a mintaátlagok eloszlását n = 15-elemű mintákra.



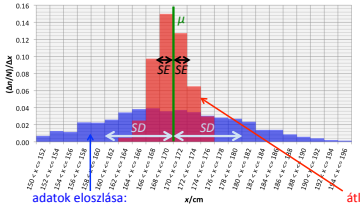
Kétszáz mintavétel (összesen 3000 elem) szimulálása tisztán mutatja, hogy az átlagok eloszlása lényegesen „keskenyebb", mint az adatoké. De mennyivel?

adatok eloszlása:
$\bar{x}$ = 170,01 cm
SD = 9,947 cm

átlagok eloszlása:
$\bar{x}$ = 170,01 cm
szórás = 2,507 cm

Átlag standard hibája

Ha az adatok szórását osztjuk a mintaelemszám (n) gyökével, megkapjuk az átlagok szórását. Ez utóbbit nevezzük az **átlag standard hibájának** ($SE_{átlag}$). Az átlagok eloszlása (legalábbis közelítőleg) normális a **centrális határeloszlás tétele** miatt.



$$SE_{átlag} = s_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

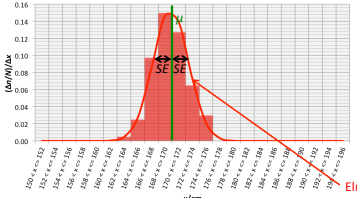
adatok eloszlása:
$\bar{x}$ = 170,01 cm
SD = 9,947 cm

átlagok eloszlása:
$\bar{x}$ = 170,01 cm
$SD/n^{0,5}$ = 9,947 cm/3,872 = 2,568 ≈ $SE_{átlag}$ = 2,507 cm

Átlag standard hibája

Azonban az átlag hibáját ($SE_{átlag}$) többnyire a minta szórásából számoljuk, így eloszlása n–1 szabadsági fokú (*degrees of freedom*, *df*) **Student t-eloszlás** lesz. Ez az eloszlás nagy *df* esetén hasonlít a normális eloszláshoz, kis *df* esetén azonban jelentősen „nehezebbek" a szélei (leptokurtotikus).

William S. **Gosset**
1876–1937
"Student"



$$SE_{átlag} = s_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \approx \frac{s}{\sqrt{n}}$$

Elméleti modell: t-eloszlás
$\mu \rightarrow \bar{x}$ = 170,01 cm
$\sigma \rightarrow$ ***SE*** = 2,507 cm
df = n–1 = 14

Átlag standard hibája

4. feladat: Meg szeretnénk becsülni a banán átlagos tömegét (várható értékét). Lemértünk öt banánt, az eredmények: 134 g, 152 g, 158 g, 141 g, 170 g. Add meg a becslés hibáját.
n = 5
$\bar{x}$ = 151 g
SD = 14,14 g
$SE_{átlag}$ = **6,32 g**

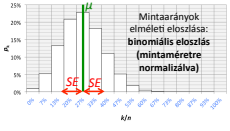
5. feladat: egy tudományos cikkben a következő mondat szerepelt: "*... a kísérletben használt patkányok átlagtömege 420 g (SE = 20 g), míg átlagos életkora 5 hónap volt ...*" A patkányok számáról azonban nem tesznek említést. Mit gondolsz, hány patkányt használhattak, ha máshonnan tudjuk, hogy az ilyen korú patkányok tömegének szórása kb. 40 g?
Az átlag szórása egyenlő a változó szórása osztva a mintaelemszám gyökével. Rendezzük át ezt a képletet n-re:
$n = (SD/SE)^2 = (40\ g/20\ g)^2 = 2^2 =$ **4**

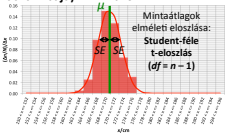
Megjegyzés: Ez egy eléggé alacsony elemszám egy vizsgálathoz, mely megmagyarázhatja, hogy a szerzők miért nem említették meg. Ez viszont az egész cikk hitelét megkérdőjelezi...

A „jó becslés” …

… hatásos

A becslés hatásos, ha szórása (értsd standard hibája) minimális.





Valószínűség becslésének standard hibája:

$$SE_{arány} = \sqrt{\frac{p(1-p)}{n}} \approx \sqrt{\frac{\frac{k}{n}\left(1-\frac{k}{n}\right)}{n}}$$

Várható érték becslésének standard hibája:

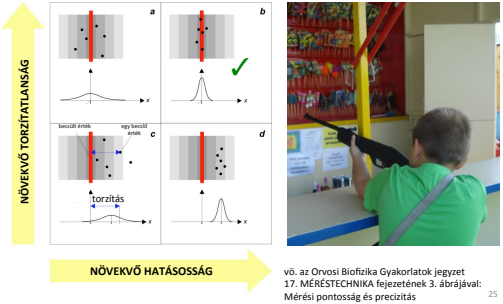
$$SE_{átlag} = s_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \approx \frac{s}{\sqrt{n}}$$

Általában kimondhatjuk, hogy a standard hiba négyzete egyenes arányban áll a statisztikai változó varianciájával, és fordított arányban a minta elemszámával.

Vagyis a hatásosság megduplázásához négyszeres mintaméretre van szükség!

A „jó becslés” …

… torzítatlan és hatásos

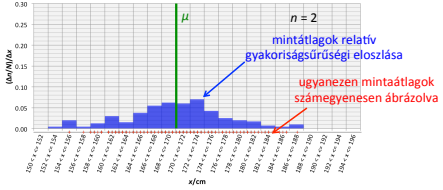


vö. az Orvosi Biofizika Gyakorlatok jegyzet 17. MÉRÉSTECHNIKA fejezetének 3. ábrájával: Mérési pontosság és precizitás

A „jó becslés” …

… konzisztens

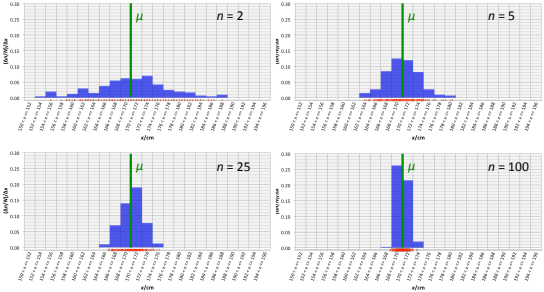
Képzeljük el egyre nagyobb méretű (*n*) mintákon elvégzett becslések sorozatát. Egy becslés konzisztens, ha a becslő érték a mintaméret növelésével egyre kevésbé tér el a becsült értéktől. Más szóval nagyobb elemszám esetén kisebb lesz a torzítás és a standard hiba is.



A testmagasság várható értékének becslésére 200 mintából számolt mintaátlagok eloszlása egyre növekvő mintaméretek esetén: *n* = 2, 5, 25, and 100

A „jó becslés” …

… konzisztens



A „jó becslés” …

… konzisztens

A konzisztenciát szigorú matematikai módszerekkel kell vizsgálni (mi nem fogjuk). Ehelyett csak szemléltessük a képleteiken, hogy a valószínűség és a várható érték becslésére tanult módszereink konzisztensek.

Valószínűség becslése

$\hat{p}_i = (k_i/n)$ ✓

A pontbecslések torzítatlanok

Várható érték becslése

$\bar{x} = \sum_{i=1}^{n} \frac{k_i}{n} \cdot x_i = \frac{1}{n} \sum_{i=1}^{n} k_i \cdot x_i$ ✓

$SE_{arány} = \sqrt{\frac{p(1-p)}{n}} \approx \sqrt{\frac{\frac{k}{n}\left(1-\frac{k}{n}\right)}{n}}$ ✓

$SE_{átlag} = s_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \approx \frac{s}{\sqrt{n}}$ ✓

A hiba nullához tart, ha növeljük az elemszámot (*n* a nevezőben szerepel)

A „jó becslés” …

… elégséges

Egy becslés akkor elégséges, ha a becsült érték **minden, a mintából kinyerhető információt** tartalmaz, amely a becsült érték szempontjából releváns.

Példa: Ha egy statisztikai változót legalább intervallumskálán mérünk, az átlag elégséges becslése a várható értéknek, mivel a minta minden megfigyelt elemének értékét felhasználja. Azaz a teljes minta ismerete nem ad több információt a várható értékre vonatkozóan annál, mintha csak az átlagot ismerjük.

Ellenpélda: Ugyanebben az esetben a medián nem elégséges, mivel csak a mintaelemek rangját használja.

Ellen-ellenpélda: Ha azonban a statisztikai változót csak ordinális szinten mérjük, a medián már elégséges becslő értéke a középértéknek (mivel az átlag nem használható).

A „jó becslés” …

… torzítatlan

Végtelensok mintavételből számolt becslő értékek átlaga egyenlő a becsült értékkel.

… hatásos

A becslés standard hibája (vagyis a becslő értékek szórása) kicsi.

… konzisztens

A mintaméret növelésével egyre kevéssé tér el a becsült értéktől a becslő érték.

… elégséges

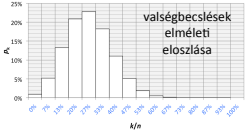
Ugyanannyi információt tartalmaz, mintha a teljes mintát adnánk meg.

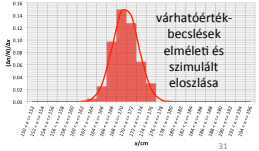
Intervallumbecslés

Az intervallum becslés során a becsült értékhez tartományt rendelünk (a becslő érték mint valségi változó skáláján egy alsó és egy felső határt), melynek neve **konfidenciaintervallum** (angolul *confidence interval*, jele: **CI**) és ehhez egy valséget, melynek neve **konfidenciaszint** (jele: 1–*α*), mely a becslési eljárás megbízhatóságát fejezi ki (konfidencia = megbízhatóság).

A CI létrehozásának tipikus lépései:

- mintavétel
- pontbecslés kiszámítása a mintaadatokból
- pontbecslés valószínűségi eloszlásának meghatározása
- az eloszlás alapján standard hiba számítása (opcionális)
- mindezek ismeretében egy tartomány megadása, ahol a becsült érték „lehet”





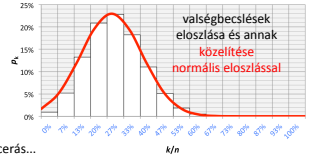
Intervallumbecslés ...

... valségre

Mi a valsége, hogy egy véletlenül kiválasztott gólyának kék szeme van?

A CI létrehozásának lépései:

- mintavétel: *n* elemű mintában *k* kék szemű
- pontbecslés: $p_k \approx k/n = \ldots$
- a becslő érték elméleti eloszlása: binomiális eloszlás
- *SE*: $SE_{arány} \approx \sqrt{\frac{\frac{k}{n}\left(1-\frac{k}{n}\right)}{n}} = \ldots$
- a megfigyelt *k* értékek binomiális eloszlást követnek, melyet át kell skálázunk *k*/*n*-re. Lehet ezzel az is számolni (ún. egzakt eljárás) de macerás...
- könnyebb ehelyett a normális eloszlással való közelítést használni: Wald-intervallum (egyszerű és elterjedt, de eléggé megbízhatatlan módszer)



Intervallumbecslés ...

... valségre

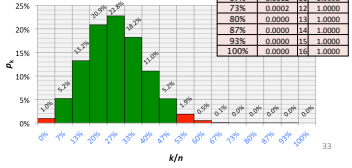
1. módszer: egzakt eljárás

- számold ki minden egyes kimenetel (*k*) valségét a becsült *p* felhasználásával, és rendezd őket csökkenő valség szerint
- kezdd el ezeket összeadogatni kezdve a legnagyobb valségűvel, majd a második legvalószínűbbel s.í.t.
- ha az összeg meghalad egy előre kiválasztott limitet, állj meg
- a kimenetelek tartománya, mely bekerült a végső összegbe, lesz az egzakt CI, maga a végső összeg pedig a tényleges konfidenciaszint (1–α)

k: *n*	p_k	#	SZUMMA
0%	0.0095	9	0.9933
7%	0.0520	6	0.9134
13%	0.1324	4	0.7510
20%	0.2087	2	0.4364
27%	0.2277	1	0.2277
33%	0.1821	3	0.6185
40%	0.1104	5	0.8614
47%	**0.0516**	**7**	**0.9650**
53%	0.0188	8	0.9838
60%	0.0053	10	0.9986
67%	0.0012	11	0.9998
73%	0.0002	12	1.0000
80%	0.0000	13	1.0000
87%	0.0000	14	1.0000
93%	0.0000	15	1.0000
100%	0.0000	16	1.0000

3. példa:

- használjuk az 1. példa adatait
- lásd a kiszámolt táblázatot és a grafikont, mely a **95%-os (valójában 96,5%-os) CI-t mutatja: 7% – 47%**

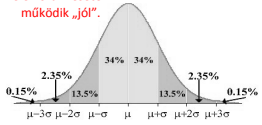


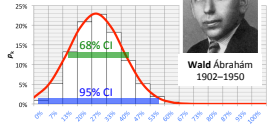
Intervallumbecslés ...

... valségre

A közelítés *n* > 30 elemszám esetén működik „jól”.

2. módszer: közelítés normális eloszlással (Wald Á.)





Wald Ábrahám
1902–1950

A $\mu \pm \sigma$ közötti tartományhoz kb. 68% valség tartozik.
Az így definiált becslési tartomány neve **68%-os konfidencia-intervallum (CI)**, a valségé pedig **konfidenciaszint (1–α)**.
A 3. példában: 68% CI = 15% – 38%. (lásd a jobb oldali ábrát is)

$$68\%\ CI \approx \frac{k}{n} \pm \sqrt{\frac{\frac{k}{n}\left(1-\frac{k}{n}\right)}{n}}$$

A $\mu \pm 2\sigma$ közötti tartományhoz kb. 95% valség tartozik.
Ez a **95%-os konfidenciaintervallumnak** felel meg, és igen gyakran használják az élettudományokban. Kis mintáknál ez a CI akár ki is lóghat a [0,1] tartományból! –> széleit le kell nyesni.
A 3. példában: 95% CI = 4% – 50%. (lásd a jobb oldali ábrát és vö. az egzakt intervallummal.)

$$95\%\ CI \approx \frac{k}{n} \pm 2 \cdot \sqrt{\frac{\frac{k}{n}\left(1-\frac{k}{n}\right)}{n}}$$

Intervallumbecslés ...

... valségre

6. feladat: Meg szeretnénk becsülni a Rhesus faktor *prevalenciáját* (alapsokaságon belüli arányát) a budapesti lakosság körében. Véletlenszerűen kiválasztottunk 42 embert, s meghatároztuk a vércsoportjukat: 35 volt Rh+.
a) Adj pontbecslést a Rhesus faktor prevalenciájára.
p(Rh+) ≈ $\hat{p}$(Rh+) = *k*(Rh+)/*n* = 35/42 = **0,833**
b) Add meg a 95%-os CI-t a Wald-intervallum eljárással.
Számoljuk ki a *SE*-t:

$$SE \approx \sqrt{\frac{\frac{k}{n}\left(1-\frac{k}{n}\right)}{n}} = \sqrt{\frac{\frac{35}{42}\left(1-\frac{35}{42}\right)}{42}} = 0.00331$$

A 95%-os CI kb. $\hat{p}$ (Rh+) ± 2*SE* között van, azaz **0,833 ± 0,006** vagy **0,826 – 0,839** .

7. feladat: add meg a 95%-os konfidenciaintervallumot a kék szemszín prevalenciájára a gólyák közt, ha egy 10 hallgatóból álló mintában kettőnek volt kék a szeme.
Az ismert képlettel számolva: $\hat{p}$ (Rh+) = 0,200 és *SE* = 0,126. Ebből a következő 95%-os CI adódna: –0,052 – 0,452 . Azonban mivel valséget becslünk, a CI nem lóghat ki a [0,1] intervallumból; szélek nyesése után a CI: **0 – 0,452**. Itt a konfidenciaszint már biztos nem 95%-os. Ez csak egy példa, hogy mennyire megbízhatatlan a Wald-intervallum – de ennek ellenére szoktuk használni nem túl kis minták esetén, ha gyors becslésre van szükség.

Intervallumbecslés ...

... várható értékre

Student-féle t-eloszlást használva (W. S. Gosset)

Az arányokhoz hasonló intervallumokat lehet definiálni, ez esetben azonban az $n-1$ szabadsági fokú (df) Student-féle t-eloszlást fogjuk használni. Így nem használhatjuk a korábbi közelítő intervallumokat sem ($\mu \pm \sigma$ és $\mu \pm 2\sigma$). Helyette függvénytáblában vagy Excel paranccsal határozzuk meg a kétszélű p-t: =T.INVERZ.2SZ(valószínűség,szabadságfok)

4. példa:

- használjuk a 2. feladat adatait
- add meg a 95%-os CI t-értékét:
  - Excel-ben: a =T.INVERZ.2SZ(5%, 15-1) függvénnyel: 2,1445
  - (a Képlettárban lévő függvénytáblából: 2.14)

ez itt α, vagyis 1 – konfidencia-szint

T-ELOSZLÁS

szabadságfok	p (valószínűség, kétszélű)							
	0,5	0,2	0,1	0,05	0,02	0,01	0,002	0,001
1	1,00	3,08	6,31	12,7	31,8	63,7	318,3	636,6
2	0,82	1,89	2,92	4,30	6,96	9,92	22,3	31,6
3	0,76	1,64	2,35	3,18	4,54	5,84	10,2	12,9
4	0,74	1,53	2,13	2,78	3,75	4,60	7,17	8,61
5	0,73	1,48	2,02	2,57	3,37	4,03	5,89	6,87
6	0,72	1,44	1,94	2,45	3,14	3,71	5,21	5,96
7	0,71	1,41	1,89	2,36	3,00	3,50	4,79	5,41
8	0,71	1,40	1,86	2,31	2,90	3,36	4,50	5,04
9	0,70	1,38	1,83	2,26	2,82	3,25	4,30	4,78
10	0,70	1,37	1,81	2,23	2,76	3,17	4,14	4,59
11	0,70	1,36	1,80	2,20	2,72	3,11	4,02	4,44
12	0,70	1,36	1,78	2,18	2,68	3,05	3,93	4,32
13	0,69	1,35	1,77	2,16	2,65	3,01	3,85	4,22
14	0,69	1,35	1,76	2,14	2,62	2,98	3,79	4,14
15	0,69	1,34	1,75	2,13	2,60	2,95	3,73	4,07
16	0,69	1,34	1,75	2,12	2,58	2,92	3,69	4,01

Ábra: A t-érték meghatározása a függvénytáblából, ha α = 5% és df = 14.

Intervallumbecslés ...

... várható értékre

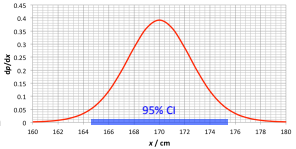
Student-féle t-eloszlást használva (W. S. Gosset)

A statisztika gyakrolaton Excel-t használunk a t kiszámolására. Azonban t eloszlása standard (μ = 0 és σ = 1), így az $x = \bar{x} + t \cdot SE$ képletet kell használnunk a megfelelő x-értékek kiszámolására. (x a statisztikai változónk, jelen esetben a magasság). –t-ből tudjuk a CI alsó határát (x_A) számolni, míg +t-ből a felsőt (x_F):

$x_A = \bar{x} - t \cdot SE$ = 170 cm – 2,1445 * 2,5 cm = 164,6 cm

$x_F = \bar{x} + t \cdot SE$ = 170 cm + 2,1445 * 2,5 cm = 175,4 cm

A 95%-os CI tehát 164,6 cm – 175,4 cm.



Ábra: Az átlagok elméleti eloszlásának grafikonja (egy df = n–1 szabadsági fokú nem-standard t-eloszlás, melynek paraméterei: $\mu = \bar{x}$ és $\sigma = SE$) és az imént számolt 95%-os CI.

Intervallumbecslés ...

... várható értékre

8. feladat: Becslést szeretnénk adni a vérkoleszterinszint várható értékére és meg szeretnénk adni a becslés 95%-os CI-t . Nyolcelemű mintát vettünk, a megfigyelt értékek a táblázatban vannak.

Excel-ben fogjuk elvégezni a számításokat.

A várható érték pontbecslése a mintaátlag:

$\mu = \bar{x}$ =ÁTLAG(*adat*) = **152,9 mg/dL**

Számoljuk ki a SE-t a már tanult módon:

n =DARAB(*adat*) = 8

SD =SZÓR.M(*adat*) = 18,47 mg/dL

SE =SD/GYÖK(n) = 6,53 mg/dL

A mintaátlagok t-eloszlást követnek, a 95%-os CI határai:

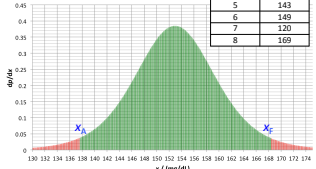
$df = n - 1 = 7$

t =T.INVERZ.2SZ(5%,df) = 2,365

x_A =$\bar{x}$-t*SE = **137,4 mg/dL**

x_F =$\bar{x}$+t*SE = **168,3 mg/dL**

megfigyelés sorszáma	koleszterin-szint (mg/dL)
1	174
2	161
3	139
4	168
5	143
6	149
7	120
8	169



Ábra: 95%-os konfidenciaintervallum (zölddel).

Intervallumbecslés ...

... várható értékre

9. feladat: Meg szeretnénk határozni a gólyák átlagmagasságát 95%-os CI-mal, de azt szeretnénk, hogy a CI ne legyen 1 cm-nél szélesebb. Mekkora legyen a minta mérete? Az irodalomból tudjuk, hogy az emberi testmagasság szórása általában 5 cm, és ezt a gólyákra is érvényesnek tekintjük.

Itt most „visszafelé" kell gondolkodnunk:

– Vann egy CI-unk (most még) ismeretlen x_A és x_F között.

– A CI szélessége a határoló értékek különbsége: *szélesség* = $x_A - x_F$

– Helyettesítsük be a képleteiket: *szélesség*= $(\bar{x}+t^*SE) - (\bar{x}-t^*SE) = 2^*t^*SE$

– Most van egy kis problémánk: a t-érték maga is függ a mintamérettől (pontosabban a szabadsági fokok számától)! Egy **feltevéssel** kell éljünk: vegyük úgy, hogy a **minta elég nagy**, és egyszerűen használjunk a $t = 2$ értéket; ekkor a képlet: *szélesség* = $4SE$.

– Tudjuk, hogy $SE = SD/n^{0,5}$, helyettesítsük be a fenti képletbe: *szélesség*= $4SD/n^{0,5}$.

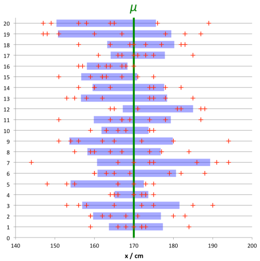
– Rendezzük át a képletet a mintaméretre: $n = (4SD/\textit{szélesség})^2$

– Helyettesítsük be az adatainkat: $n = (4^*(5\text{ cm})/(1\text{ cm}))^2$ = **400**

Intervallumbecslés

Konfidenciaszint ($1 - \alpha$): annak a valószínűsége, hogy egy adott módszerrel meghatározott CI-ok tartalmazzák-e a becsült értéket.
Ebből nem tudjuk meg, hogy egy *konkrét* CI *ténylegesen* tartalmazza-e a becsült értéket.
Magát a becslési eljárást jellemzi általában, nem annak egy konrét eredményét. Azt nem tudjuk megmondani, hogy a becsült érték benne van-e az éppen kiszámolt CI-ban, csak akkor, ha valahonnan ismerjük a becsült értéket – mely esetben az egész becslésre nincs szükség.

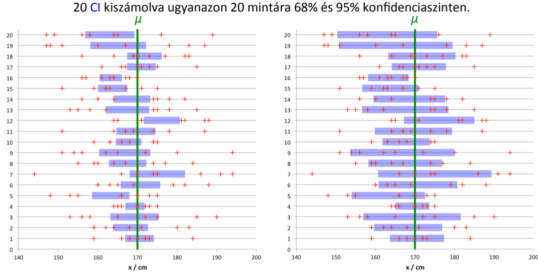
Szignifikanciaszint (α): a komplementer valség, pl. ha a konfidenciaszint 95%, akkor a szignifikanciaszint 5%. Hogy melyiket használjuk, az a kontextustól függ.



Ábra: A testmagasság várható értékre (μ) ugyanazon módszerel végzett 20 becslés szimulációja: 8-elemű minta vételezése (piros +), átlag és standard hibájának számítása, majd a CI megadása: *átlag* ± 2,36 × *SE* (kék sáv). A becsült értékek: μ = 170 cm, σ = 10 cm.

Intervallumbecslés

20 CI kiszámolva ugyanazon 20 mintára 68% és 95% konfidenciaszinten.



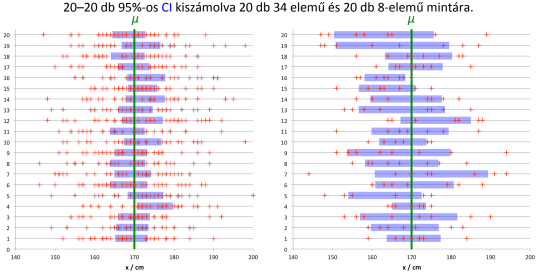
n = 8 (df = 7), μ = 170 cm, σ = 10 cm
$1 - \alpha$ = 68%

n = 8 (df = 7), μ = 170 cm, σ = 10 cm
$1 - \alpha$ = 95%

Magasabb konfidenciaszint =
kisebb valószínűséggel esik a CI-on kívül a becsült érték, de kevesebb az információtartalma.

Intervallumbecslés

20–20 db 95%-os CI kiszámolva 20 db 34 elemű és 20 db 8-elemű mintára.



n = 32 (df = 31), μ = 170 cm, σ = 10 cm
$1 - \alpha$ = 95%

n = 8 (df = 7), μ = 170 cm, σ = 10 cm
$1 - \alpha$ = 95%

Nagyobb mintaelemszám ugyanazon konfidenciaszinten = szűkebb CI.
(4-szeres elemszám átlagosan fele szélességű CI-kat eredményez.)

Függelék: Normáltartomány

Normáltartomány, referenciatartomány vagy referenciaintervallum: a statisztikai változóra vonatkozó tartomány, mely a **változó egy elemét 95%-os valószínűséggel tartalmazza**.

A normáltartomány is egyfajta intervallumbecslés, de itt a statisztikai változó eloszlását, nem pedig eloszlásának egy paraméterét jellemezzük. Más szóval: a normáltartomány a magukra az adatokra vonatkozó 95%-os CI. Ezt a tartományt szokták feltüntetni az orvosi laborleleteken is.

TESTS	RESULT	FLAG	UNITS	REFERENCE INTERVAL	LAB
CMP12+LP+6AC+CBC/D/Plt+UA					
Chemistries					01
Glucose, Serum	80		mg/dL	65-99	01
Uric Acid, Serum	7.2		mg/dL	2.4-8.2	01
BUN	21		mg/dL	5-26	01
Creatinine, Serum	0.94		mg/dL	0.76-1.27	01
Glom Filt Rate, Est	>59		mL/min/1.73	>59	01
If African-American	>59		mL/min/1.73	>59	01
Note: Persistent reduction for 3 months or more in an eGFR <60 mL/min/1.73 m2 defines CKD. Patients with eGFR values >/=60 mL/min/1.73 m2 may also have CKD if evidence of persistent proteinuria is present. Additional information may be found at www.kdoqi.org.					
BUN/Creatinine Ratio	22			8-27	01
Sodium, Serum	142		mmol/L	135-145	01
Potassium, Serum	4.4		mmol/L	3.5-5.2	01
Chloride, Serum	100		mmol/L	97-108	01
Calcium, Serum	10.1		mg/dL	8.5-10.6	01
Phosphorus, Serum	3.9		mg/dL	2.5-4.5	01
Protein, Total, Serum	8.2		g/dL	6.0-8.5	01
Albumin, Serum	5.1		g/dL	3.5-5.5	01
Globulin, Total	3.1		g/dL	1.5-4.5	01
A/G Ratio	1.6			1.1-2.5	01
Bilirubin, Total	1.2		mg/dL	0.1-1.2	01
Bilirubin, Direct	0.20		mg/dL	0.00-0.40	01
Bilirubin, Indirect	**1.00**	**High**	mg/dL	0.10-0.80	01
Alkaline Phosphatase, S	70		IU/L	25-150	01
LDH	142		IU/L	100-250	01
AST (SGOT)	35		IU/L	0-40	01
GGT	**70**	**High**	IU/L	0-65	01
Iron, Serum	115		ug/dL	40-155	01

Normális eloszlású változó esetén a normáltartomány kb.: $\bar{x} \pm 2SD$.
(pontosabban: $\bar{x} \pm 1{,}96SD$)

10. feladat: Számold ki a szérum glükózszint μ, σ és σ^2 paramétereit a laborlelet adatainak segítségével.

$x_L = \mu - 2\sigma = 65$ mg/dL
$x_U = \mu + 2\sigma = 99$ mg/dL

$x_U + x_L = (\mu+2\sigma) + (\mu-2\sigma) = 2\mu$
$\mu = (x_U + x_L)/2 =$ **82 mg/dL**

$x_U - x_L = (\mu+2\sigma) - (\mu-2\sigma) = 4\sigma$
$\sigma = (x_U - x_L)/4 =$ **8,5 mg/dL**

$\sigma^2 =$ **72,25 (mg/dL)²**

Ellenőrző kérdések

– Mi a populáció?
– Mi a minta?
– Hogyan szerezhetünk információt egy statisztikai változóról?
– Milyen becsléstípusokat ismersz?
– Mik a becslés lépései?
– Mi a „plug-in" becslés?
– Mi a pontbecslés?
– Mi a pontbecslés hátránya?
– Mi a becslő értéke a valséghek, várható értéknek, elméleti varianciának és elméleti szórásnak?
– Mik a „jó becslés" tulajdonságai?
– Mi a torzítatlanság? Szemléltesd példával.
– Mi a hatásosság? Hogyan lehet matematikailag kifejezni?
– Milyen eloszlást követnek az arányok?
– Milyen eloszlást követ a várható érték becslő értéke?
– Mi a standard hiba?
– Hogyan számolható egy arány standard hibája?
– Hogyan számolható egy átlag standard hibája?
– Hogyan függ a standard hiba a változó szórásától?
– Hogyan függ a standard hiba a minta méretétől (elemszámától)?
– Meg szeretnénk háromszorozni egy becslés hatásosságát. Hogyan változtassunk a minta méretén?
– Mekkora egy 25 elemű mintából számolt arány legnagyobb lehetséges hibája?

Ellenőrző kérdések

– Mik egy konzisztens becslés ismérvei?
– Mit jelent, hogy egy becslés elégséges?
– Mi a Bessel-korrekció? Hol használjuk és mi a célja?
– Mi a CI jelentése?
– Mi a konfidenciaszint jele és jelentése?
– Hogyan adhatunk meg egzakt CI-ot egy arányhoz?
– Mi a Wald-intervallum alapja, mi az előnye és a hátránya?
– Hogyan változik a CI, ha emelem a konfidenciaszintet?
– Hogyan változik a CI, ha növelem a mintaméretet?
– Hogyan változik a CI, ha a vizsgált változó szórása kisebb?
– Hogyan adunk meg CI-t várható érték becslése esetén?
– Mi a referenciatartomány?
– Hogyan adható meg egy normális eloszlású valségi változó referenciatartománya?
– Egy laborleletünkben különféle laborértékek mellett referenciatartományokat látunk. Hogyan lehet ezekből a kórház által használt várható értékeket és szórásokat kiszámolni?
– Mi a kapcsolat a konfidenciaszint és a szignifikanciaszint között?
– Miért nem lehetséges megmondani, hogy egy adott CI ténylegesen tartalmazza-e a becsült értéket?