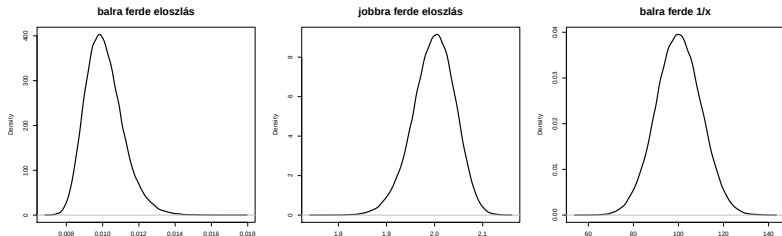


Valószínűség, pontbecslés, konfidenciaintervallum

Normális eloszlás tesztje

Kolmogorov-Szmirnov vagy Wilk-Shapiro próba.

R-funkció: `shapiro.test(vektor)`



Ha $p > 0,05$, elfogadjuk, hogy a minta normális eloszlású (ld. később).

Transzformációk

Unimodális, jobbra vagy balra ferde eloszlások gyakran átalakíthatóak normális eloszlásúvá.

Szokásos eljárások:

- ▶ $x = \log(x)$
- ▶ $x = 1/x$
- ▶ $x = \sqrt{x}$
- ▶ ...

Valószínűség a mindennapokban

Köznyelvi jelentés: tapasztalat alapú becslés (n megfigyelt esetből hányszor történt meg egy adott esemény). Pl.

„valószínűleg mindjárt elered az eső” (mert ha ilyen borús az ég, gyakran esik), „valószínűleg idén sem lesz fizetésemelés” (mert tíz éve nem volt).

A valószínűség soha nem jelent biztos tudást! Néha mégsem esik, ha borús az ég, és néha mégis van fizetésemelés.

Intuitív becslésnek kevés fokozata van: *nem túl valószínű, elég valószínű, nagyon valószínű, több mint valószínű.*

Valószínűség a szerencsejátékban

Fej vagy írás egy érme feldobásakor?

Megfigyelés: 10 dobás, 20, 30...

Valószínűség a szerencsejátékban

Fej vagy írás egy érme feldobásakor?

Megfigyelés: 10 dobás, 20, 30...

Fejek száma egyre jobban közelíti a 0,5-ös értéket.

Empirikus valószínűség P definíciója:

$P = \text{fej} / \text{összes dobás}$

ahol a dobások száma a végtelenhez közelít.

$\Rightarrow$ valószínűség értéke mindig 0 (egyáltalán nem valószínű) és 1 (biztos) között mozog.

Példák

1. Adott szám dobása kockával.
 2. Ász húzása egy 32 lapos kártyapakliból.
 3. Kétszer egymás után fej dobása.
-
5. Egy véletlenszerűen megkérdezett személy diplomás nő, ha a diplomások aránya 22,4%, és a nők aránya 50%.

Példák

1. Adott szám dobása kockával.
 2. Ász húzása egy 32 lapos kártyapakliból.
 3. Kétszer egymás után fej dobása.
-
5. Egy véletlenszerűen megkérdezett személy diplomás nő, ha a diplomások aránya 22,4%, és a nők aránya 50%.

Példák

1. Adott szám dobása kockával.
 2. Ász húzása egy 32 lapos kártyapakliból.
 3. Kétszer egymás után fej dobása.
-
5. Egy véletlenszerűen megkérdezett személy diplomás nő, ha a diplomások aránya 22,4%, és a nők aránya 50%.

Példák

1. Adott szám dobása kockával.
2. Ász húzása egy 32 lapos kártyapakliból.
3. Kétszer egymás után fej dobása.
4. Mekkora a valószínűsége annak, hogy egy véletlenszerűen kiválasztott magyar állampolgár katolikus, ha az összes megkérdezett közötti arány katolikus 51%, református 16%, evangélikus 3%, nem vallásos 14,5% stb.
5. Egy véletlenszerűen megkérdezett személy diplomás nő, ha a diplomások aránya 22,4%, és a nők aránya 50%.

Példák

1. Adott szám dobása kockával: $\text{adott szám} / \text{összes szám} = 1/6 = 0,1667$.

Példák

1. Adott szám dobása kockával: $\text{adott szám} / \text{összes szám} = 1/6 = 0,1667$.
2. Ász húzása egy 32 lapos kártyapakliból: $\text{ászok száma} / \text{összes kártya} = 4/32 = 0,125$.

Példák

1. Adott szám dobása kockával: $\text{adott szám} / \text{összes szám} = 1/6 = 0,1667$.
2. Ász húzása egy 32 lapos kártyapakliból: $\text{ászok száma} / \text{összes kártya} = 4/32 = 0,125$.
3. kétszer egymás után fej dobása:
 $(\text{fej}+\text{fej})+(\text{fej}+\text{írás})+(\text{írás}+\text{fej})+(\text{írás}+\text{írás}) = 1/4 = 0,25$.

Példák

1. Adott szám dobása kockával: $\text{adott szám} / \text{összes szám} = 1/6 = 0,1667$.
2. Ász húzása egy 32 lapos kártyapakliból: $\text{ászok száma} / \text{összes kártya} = 4/32 = 0,125$.
3. kétszer egymás után fej dobása:
 $(\text{fej} + \text{fej}) + (\text{fej} + \text{írás}) + (\text{írás} + \text{fej}) + (\text{írás} + \text{írás}) = 1/4 = 0,25$.
4. Mekkora a valószínűsége annak, hogy egy véletlenszerűen kiválasztott magyar állampolgár katolikus, ha az összes megkérdezett közötti arány katolikus 51%, református 16%, evangélikus 3%, nem vallásos 14,5% stb.: $51\% = 0,51$.

Példák

1. Adott szám dobása kockával: adott szám/összes szám = $1/6 = 0,1667$.
2. Ász húzása egy 32 lapos kártyapakliból: ászok száma/összes kártya = $4/32 = 0,125$.
3. kétszer egymás után fej dobása:
(fej+fej)+(fej+írás)+(írás+fej)+(írás+írás) = $1/4 = 0,25$.
4. Mekkora a valószínűsége annak, hogy egy véletlenszerűen kiválasztott magyar állampolgár katolikus, ha az összes megkérdezett közötti arány katolikus 51%, református 16%, evangélikus 3%, nem vallásos 14,5% stb.: $51\% = 0,51$.
5. Egy véletlenszerűen megkérdezett személy diplomás nő, ha a diplomások aránya 22,4%, és a nők aránya 50%: $0,224 \cdot 0,5 = 0,112$.

Inferenciális statisztika: oksági vagy relációs alapú statisztika. A minta értékei alapján következtet a populációra.

DE: a minta alapján a populációra csak **becsléseket** tehetünk.

1. probléma: különböző minták különböző átlagokat eredményeznek, még véletlenszerű kiválasztás esetén is.
2. probléma: véges az időnk, csak véges mintával tudunk dolgozni.

A normális eloszlás jelentősége

Feltételezés: n minta $\bar{x}$ átlagai normális eloszlást mutatnak a populáció μ átlaga körül, ha a populáció szórása σ .

A mintaátlagok μ körüli szórása egyenlő a standard hibával, azaz $se = \frac{s}{\sqrt{n}}$ -vel.

A **szórás** egyes adatpontok **mintaátlagtól** való távolságát fejezi ki.

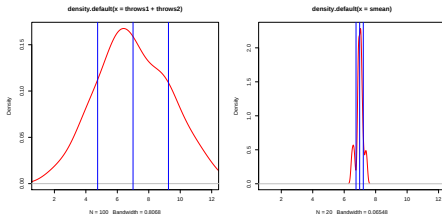
A **standard hiba** a mintaátlagok **populációátlagtól** való távolságát fejezi ki.

Előny: egyetlen minta átlaga és szórása alapján következtethetünk a populáció ismeretlen értékeire.

A minta és a populáció szórása

Két kockával dobunk 100-szor.

Bal: egyetlen, 100 dobásból álló minta összegei. Jobb: 20 minta átlagai, amelyek egyenként 100 dobásból állnak.



Minta átlaga = 7, átlagok átlaga = 6,98.

Minta szórása = 2,25, átlagok szórása = 0,23.

Standard hiba EGYETLEN mintából számolva:

$$2,25/\sqrt{100} = 0,225$$

→ a 20 mintából számolt szórás jó közelítése 20 minta híján is.

Pontbecslés

Véletlen minta átlaga függ a véletlentől, azaz egy **becsült pont**.

Megmérjük egy véletlenszerűen kiválasztott, 300 fős, férfi egyetemistából álló csoport testmagasságát.

$$s = 6,3 \text{ cm}$$

A minta részmintáiból számolt átlagok szórása függ az elemszámtól: a tíz fős minták szórása a minta átlaga körül $se = 6,3/\sqrt{10} = 1,99$, ötven fős mintáké $se = 6,3/\sqrt{50} = 0,89$, stb.

$\Rightarrow$ minél nagyobb az elemszám, annál kisebb a szórás, azaz az egyes mintaátlagok annál jobban közelítik a populáció átlagát.

Feladat

Fájl letöltése:

<http://clara.nytud.hu/~mady/courses/statistics/materials/testmagassag.txt>

300 férfi egyetemi hallgató testmagassága.

Testmagasság adatai testmagassag.txt nevű fájlban.

átlag: `mean()`

szórás: `sd()`

gyök: `sqrt()`

Hogyan számoljuk ki az első tíz fő testmagasságának átlagát?

Feladat

Fájl letöltése:

<http://clara.nytud.hu/~mady/courses/statistics/materials/testmagassag.txt>

300 férfi egyetemi hallgató testmagassága.

Testmagasság adatai testmagassag.txt nevű fájlban.

átlag: `mean()`

szórás: `sd()`

gyök: `sqrt()`

Hogyan számoljuk ki az első tíz fő testmagasságának átlagát?

```
mean(testmagassag$height[1:10])
```

Teljes minta standard hibája?

Feladat

Fájl letöltése:

<http://clara.nytud.hu/~mady/courses/statistics/materials/testmagassag.txt>

300 férfi egyetemi hallgató testmagassága.

Testmagasság adatai testmagassag.txt nevű fájlban.

átlag: `mean()`

szórás: `sd()`

gyök: `sqrt()`

Hogyan számoljuk ki az első tíz fő testmagasságának átlagát?

```
mean(testmagassag$height[1:10])
```

Teljes minta standard hibája?

```
sd(testmagassag)/sqrt(300)
```

0,36

Konfidenciaintervallum

Kérdés: igaz-e, hogy a véletlen minta átlaga beleesik az ismeretlen populációátlag körül szóródó mintaátlagokba?

Nehézség: μ -t nem ismerjük, csak $\bar{x}$ -et.

$\Rightarrow$ döntés nem lehetséges, csak egy adott valószínűségi határon, azaz **konfidenciaintervallumon** belüli valószínűség megállapítása.

Kérdés: igaz-e, hogy $\bar{x}$ 95%-os valószínűséggel beleesik a μ körül standard hibával szóródó mintaátlagok tartományába?

Konfidenciaszint ebben az esetben: $p = 0,95$.

Kiindulás

- ▶ Véletlenszerű minták átlagai normális eloszlásúak.
- ▶ Átlagok 95%-a $\pm 1,96 \cdot \text{szórás } (s)$, itt $s/\sqrt{n}$, azaz $1,96 \cdot \text{standard hiba } (se)$.
- ▶ Keresett μ a populáció eloszlásának középpontja (szimmetria feltételezése miatt).

tehát:

$$p(-1,96 \cdot se + \mu < \bar{x} < \mu + 1,96 \cdot se) = 0,95$$

Cél: a 95%-os konfidenciaintervallumon belüli határértékek meghatározása negatív és pozitív irányban.

Konfidenciaintervallum $\bar{x}$ alapján

$$p(-1,96 * se + \mu < \bar{x} < \mu + 1,96 * se) = 0,95$$

$$-\mu$$

$$p(-1,96 * se < \bar{x} - \mu < 1,96 * se) = 0,95$$

$$* - 1$$

$$p(1,96 * se > \mu - \bar{x} > -1,96 * se) = 0,95$$

$$+\bar{x}$$

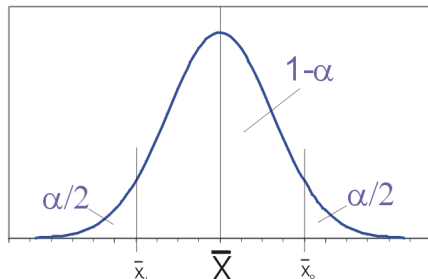
$$p(1,96 * se + \bar{x} > \mu > \bar{x} - 1,96 * se) = 0,95$$

$$p(-1,96 * se + \bar{x} < \mu < \bar{x} + 1,96 * se) = 0,95$$

Konfidenciaszint

Konfidenciaintervallum: értéktartomány, amely a becslendő paramétert előre rögzített valószínűséggel tartalmazza.

Konfidenciaintervallumon kívüli tartomány: $\alpha = 1 - p$.



Ha $\bar{x}$ nem esik a 95%-os konfidenciaintervallumba, akkor is tartozhat az adott populációhoz! Tévedés valószínűsége 5%, ez az ún. alfa-hiba.

Kiindulási hipotézis tesztelése

Hipotézis állítása falszifikáción keresztül, azaz az állításunk **ellenhipotézisét** teszteljük.

Az empirikus vizsgálatokban általában abban vagyunk érdekeltek, hogy vizsgált érték $1 - p$, azaz α tartományba essen.

$\Rightarrow$ szignifikanciaszintet α értékével szokás megadni, azaz 0,05 vagy 5%.

Ha azt akarjuk bizonyítani, hogy egy adott minta nem tartozik az adott p konfidenciaintervallumba, akkor a mintának negatív és pozitív irányban az $\alpha/2$ tartományba kell tartoznia. **Tehát egy szimmetrikus, azaz kétoldalas tesztnél az azonosság elutasítása 2,5%-ra teljesül.**

A mintaméret jelentősége a hipotézistesztesztelés szempontjából

A $p = 0,95$ -es konfidenciahatár kritikus értékei:

alsó kritikus érték: $-1.96 * se + \bar{x}$

felső kritikus érték: $+1.96 * se + \bar{x}$

A standard hiba kiszámítása:

$$se = \frac{s}{\sqrt{n}}$$

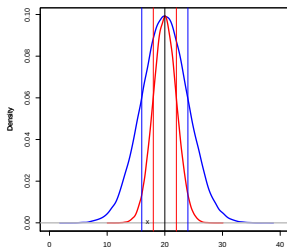
Ha n alacsony, a standard hiba nagyobb, mert a nevezőben $\sqrt{n}$ szerepel $\Rightarrow$ nagyobb standard hiba esetén a kritikus érték nagyobb lesz, ezért H_0 elutasításának lehetősége kisebb.

Kisebb és nagyobb standard hiba és hipotézistesztesztelés

Két populáció, amelyek átlaga $\mu = 20$. 1. populáció szórása $se = 1$, 2. populációé $se = 2$.

1. populáció alsó kritikus értéke : $20 - 1 \cdot 1.96 = 18.04$, 2. populációé: $20 - 2 \cdot 1.96 = 16.08$.

Egy olyan minta, amely átlaga $\bar{x} = 17$, ehhez a populációhoz tartozik?



1. populáció: $17 < 18.04 \Rightarrow H_0$ -t elutasíthatjuk.

2. populáció: $17 > 16.08 \Rightarrow H_0$ -t nem utasíthatjuk el.

Feladat

Számoljuk ki a `testmagassag` R-objektum első tíz elemének átlagát. Beleesik a teljes, 300 elemű minta 95%-os konfidenciaintervallumába?

Első tíz elem átlagának kiszámítása:

Feladat

Számoljuk ki a `testmagassag` R-objektum első tíz elemének átlagát. Beleesik a teljes, 300 elemű minta 95%-os konfidenciaintervallumába?

Első tíz elem átlagának kiszámítása:

```
mean(testmagassag$height[1:10])  
175.9037
```

95%-os konfidenciaintervallum határai?

Feladat

Számoljuk ki a testmagassag R-objektum első tíz elemének átlagát. Beleesik a teljes, 300 elemű minta 95%-os konfidenciaintervallumába?

Első tíz elem átlagának kiszámítása:

```
mean(testmagassag$height[1:10])  
175.9037
```

95%-os konfidenciaintervallum határai?

A régi szép időkben megnéztük az adott $\alpha/2$ tartományra megadott (standardizált!) z-értéket a függvénytáblázatban.

Feladat

Számoljuk ki a `testmagassag` R-objektum első tíz elemének átlagát. Beleesik a teljes, 300 elemű minta 95%-os konfidenciaintervallumába?

Első tíz elem átlagának kiszámítása:

```
mean(testmagassag$height[1:10])  
175.9037
```

95%-os konfidenciaintervallum határai?

A régi szép időkben megnéztük az adott $\alpha/2$ tartományra megadott (standardizált!) z-értéket a függvénytáblázatban.

Manapság letöltjük a `gmodels` nevű R-csomagot, és lekérdezzük a határokat a

`ci` függvényvel.

```
ci(testmagassag$height,0.95) Estimate CI lower CI  
upper Std. Error 178.0349657 177.3166967 178.7532346  
0.3649871
```