Statisztika

Politológus képzés

Daróczi Gergely

Politológia Tanszék

2012. április 24.

Pázmány Péter
Katolikus Egyetem
Bölcsészettudományi Kar

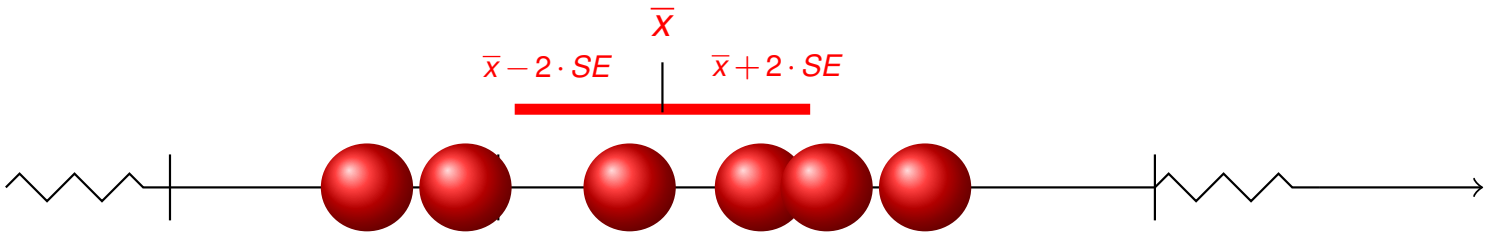
Outline

Szükséges képletek:

- **számtani átlag**: $\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
- **korrigált empirikus szórás**: $S^* = \sqrt{\sum_{i=1}^{n} \frac{(x_i - \overline{x})^2}{n}}$
- **standard/mintavételi hiba**: $SE = \frac{S^*}{\sqrt{n}} \cdot \sqrt{1 - \frac{n}{N}} \approx \frac{S^*}{\sqrt{n}}$
- **konfidencia-intervallum**: $\overline{x} \pm z \cdot SE$, ahol legtöbbször $z = 1,96$

Tehát:

- **konfidencia-intervallum**: $[\overline{x} - 2 \cdot SE; \overline{x} + 2 \cdot SE]$



„Az őszi kutatásban is megkérdezték az autósokat az üzemanyagárak lélektani határáról. A felmérés közben hétről hétre dőltek meg az üzemanyagár csúcsok, ezért a kutatás a 400 és a 450 forint közötti literenkénti ársávot vizsgálta. A gázolaj árának hatását most is rugalmasabban ítélték meg az autósok, még mindig sokan vannak, akik 450 forint feletti áron is ugyanannyit tankolnának, mint most. A benzinnél 420 forintos árnál a válaszadók többsége már nem tankolna annyit mint korábban, s jelentősen csökkentené az autó használatát."

Forensis Autóklub (2011.november)

A mintavételi hiba és konfidencia-intervallum

Gyakorlat

„Mi az az üzemanyag ár, ahol már hosszútávra leállítanád az autódat és nem tankolnál rendszeresen?”

410, 420, 420, 430, 500, 450, 400, 425, 460

„Mi az az üzemanyag ár, ahol már hosszútávra leállítanád az autódat és nem tankolnál rendszeresen?”

410, 420, 420, 430, 500, 450, 400, 425, 460

Leíró statisztikák:

- **számtani átlag**: $\overline{x} = \frac{410+420+420+430+500+450+400+425+460}{9} = 435$
- **medián**: 425
- **módusz**: 420
- **minimum érték**: 400
- **maximum érték**: 500
- **terjedelem**: 100

A mintavételi hiba és konfidencia-intervallum

Gyakorlat

„Mi az az üzemanyag ár, ahol már hosszútávra leállítanád az autódat és nem tankolnál rendszeresen?”

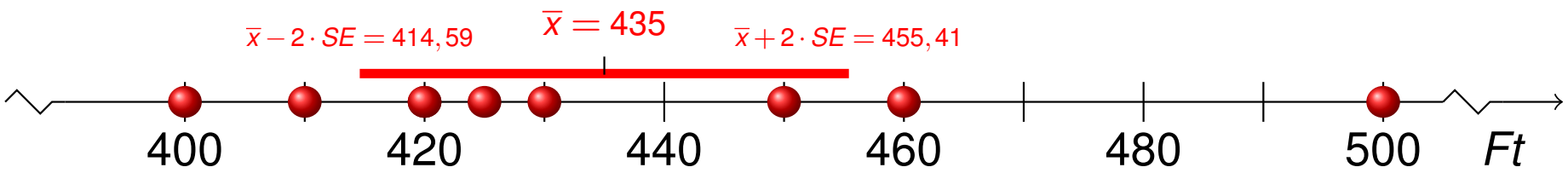
$$410, 420, 420, 430, 500, 450, 400, 425, 460$$

- **számtani átlag**: $\overline{x} = \frac{410+420+420+430+500+450+400+425+460}{9} = 435$
- **korrigált empirikus szórás**: $S^* = 30,619$
- **standard/mintavételi hiba**: $SE = \frac{30,619}{\sqrt{9}} = \frac{30,619}{3} = 10,206$
- **konfidencia-intervallum**: $435 \pm 2 \cdot 10,206 = [414,59; 455,41]$

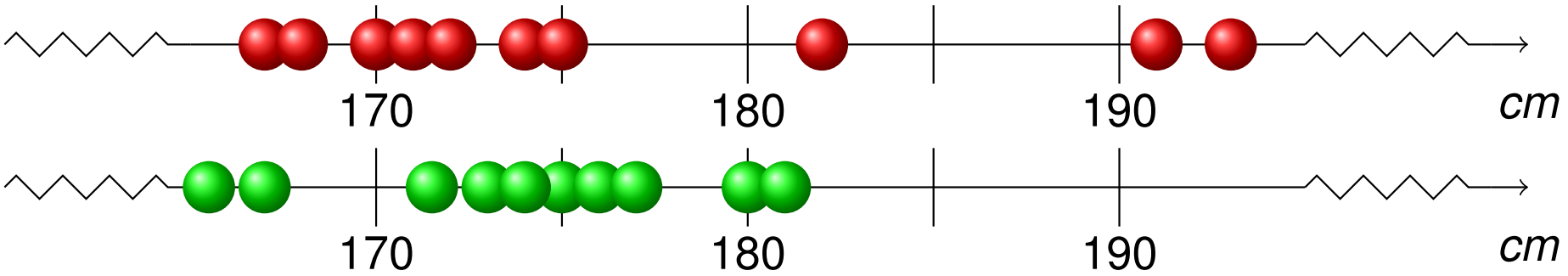
„Mi az az üzemanyag ár, ahol már hosszútávra leállítanád az autódat és nem tankolnál rendszeresen?”

$$410, 420, 420, 430, 500, 450, 400, 425, 460$$

- **számtani átlag**: $\overline{x} = \frac{410+420+420+430+500+450+400+425+460}{9} = 435$
- **korrigált empirikus szórás**: $S^* = 30,619$
- **standard/mintavételi hiba**: $SE = \frac{30,619}{\sqrt{9}} = \frac{30,619}{3} = 10,206$
- **konfidencia-intervallum**: $435 \pm 2 \cdot 10,206 = [414,59; 455,41]$



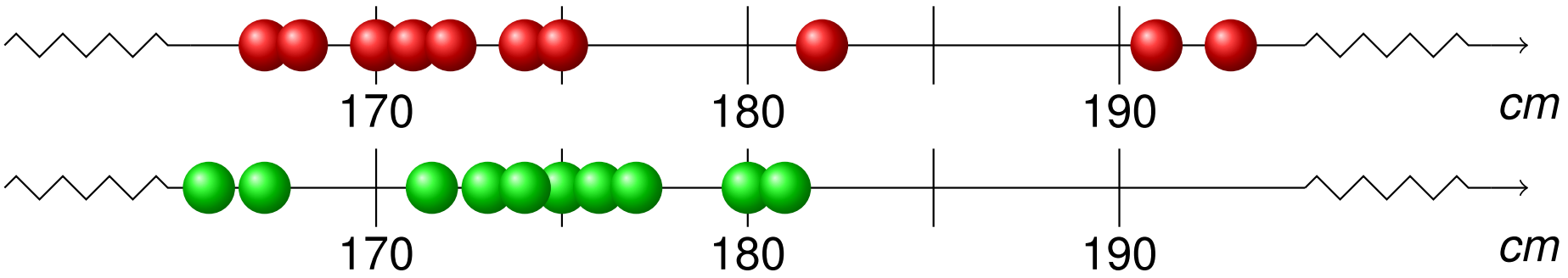
10-10 diák magasságát mértük 2 osztályteremben?

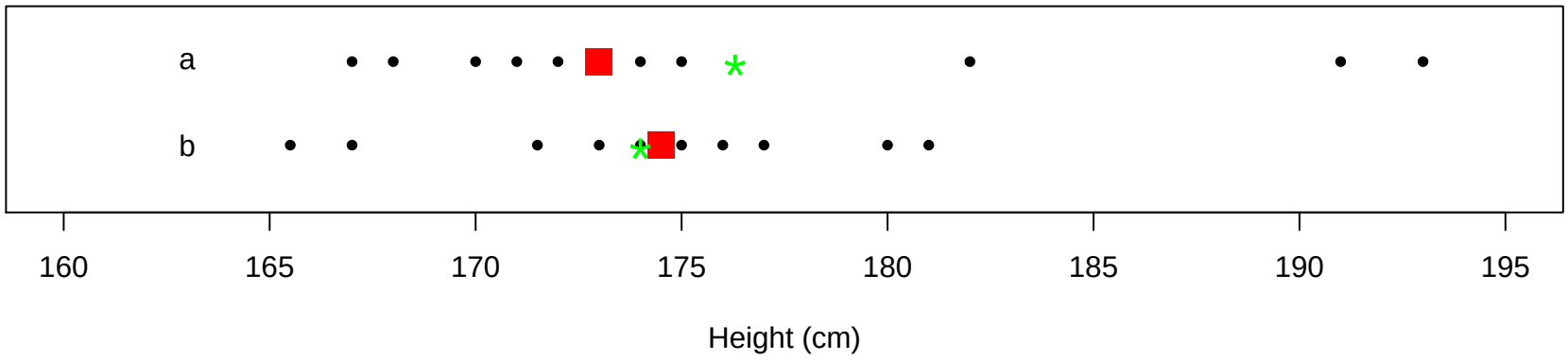


Melyik osztály diákjai a magassabbak a leíró statisztikák alapján adható becslések alapján?

Ismétlés

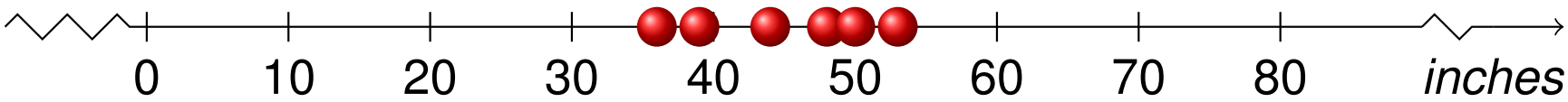
Középértékek



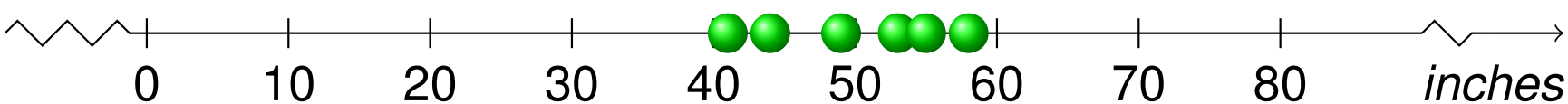


Hat gesztenyefa magasságát mértük meg a Margit-szigeten két időpontban. Számoljuk ki, határozzuk meg az átlagot, a szórás és a standard/mintavételi hibát!
Kimutatható szignifikáns eltérés a két minta átlaga között?

1. **2012. március 22**: 36 48 50 44 53 39



2. **2012. április 1**: 41 53 55 49 58 44



Hat gesztenyefa magasságát mértük meg a Margit-szigeten két időpontban.
Számoljuk ki, határozzuk meg az átlagot, a szórás és a standard/mintavételi hibát!
Kimutatható szignifikáns eltérés a két minta átlaga között?

1. **2012. március 22**: 36 48 50 44 53 39

$$X = \{36, 48, 50, 44, 53, 39\}$$

$$\overline{x} = \frac{36+48+50+44+53+39}{6} = \frac{270}{6} = 45$$

$$\sigma = \sqrt{\frac{(36-45)^2+(48-45)^2+(50-45)^2+(44-45)^2+(54-45)^2+(39-45)^2}{5}} =$$

$$= \sqrt{\frac{81+9+25+1+64+36}{5}} = \sqrt{\frac{216}{5}} = \sqrt{43.2} = 6.57$$

$$SE = \frac{6.57}{\sqrt{6}} = \frac{6.57}{2.44} = 2.68$$

A várható érték 40,5 és 49.5 között alakul (95% valószínűséggel).

Hat gesztenyefa magasságát mértük meg a Margit-szigeten két időpontban. Számoljuk ki, határozzuk meg az átlagot, a szórás és a standard/mintavételi hibát!
Kimutatható szignifikáns eltérés a két minta átlaga között?

1. **2012. április 1**: 41 53 55 49 58 44

$$X = \{41, 53, 55, 49, 58, 44\}$$

$$\overline{x} = \frac{41+53+55+49+58+44}{6} = \frac{300}{6} = 50$$

$$\sigma = \sqrt{\frac{(41-50)^2+(53-50)^2+(55-50)^2+(49-50)^2+(58-50)^2+(44-50)^2}{5}} =$$

$$= \sqrt{\frac{81+9+25+1+64+36}{5}} = \sqrt{\frac{216}{5}} = \sqrt{43.2} = 6.57$$

$$SE = \frac{6.57}{\sqrt{6}} = \frac{6.57}{2.44} = 2.68$$

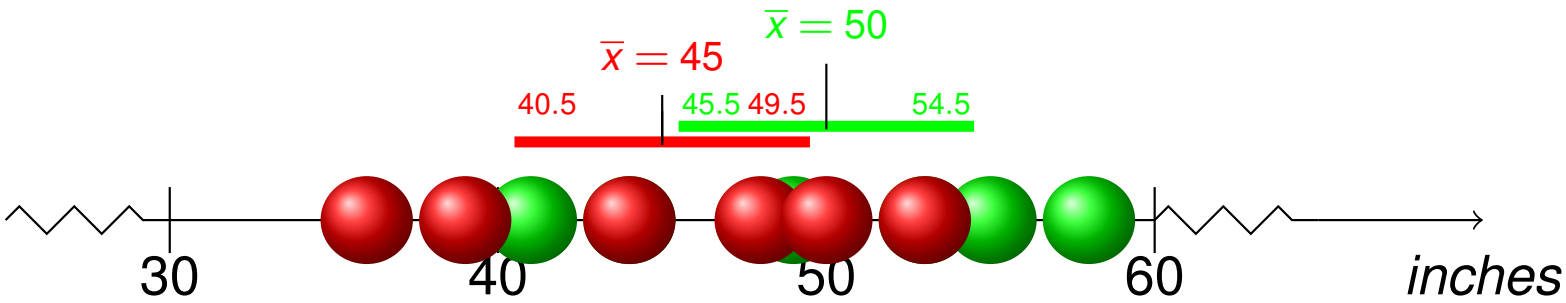
A várható érték 45,5 és 54.5 között alakul (95% valószínűséggel).

Minták összehasonlítása

Eredmények

Hat gesztenyefa magasságát mértük meg a Margit-szigeten két időpontban. Számoljuk ki, határozzuk meg az átlagot, a szórás és a standard/mintavételi hibát!
Kimutatható szignifikáns eltérés a két minta átlaga között?

1. **2012. március 22**: 36 48 50 44 53 39
2. **2012. április 1**: 41 53 55 49 58 44



Mintanagyság meghatározása általános esetben

Példa

Mekkora mintára van szükségem ahhoz, hogy 5 perc pontosság meg tudjam állapítani a napi tévénézésre fordított idő hosszát a felnőtt magyar lakosság körében?

Mintanagyság meghatározása általános esetben

Példa

Mekkora mintára van szükségem ahhoz, hogy 5 perc pontosság meg tudjam állapítani a napi tévénézésre fordított idő hosszát a felnőtt magyar lakosság körében?

- 5 perc pontosság 95 %-os döntési szinten: $SE = 2.5$,
- becsült szórás: 10
- $SE = \frac{s^*}{\sqrt{n}}$

$$\Downarrow$$

- $2,5 = \frac{10}{\sqrt{n}}$

$$\Downarrow$$

- $2,5 \cdot \sqrt{n} = 10$
- $\sqrt{n} = 4$
- $n = 16$

Mintanagyság meghatározása általános esetben

Példa

Mekkora mintára van szükségem ahhoz, hogy 1 perc pontosság meg tudjam állapítani a napi tévénézésre fordított idő hosszát a felnőtt magyar lakosság körében?

Mintanagyság meghatározása általános esetben

Példa

Mekkora mintára van szükségem ahhoz, hogy 1 perc pontosság meg tudjam állapítani a napi tévénézésre fordított idő hosszát a felnőtt magyar lakosság körében?

- 1 perc pontosság 95 %-os döntési szinten: $SE = 0.5$,
- becsült szórás: 10
- $SE = \frac{s^*}{\sqrt{n}}$

$$\Downarrow$$

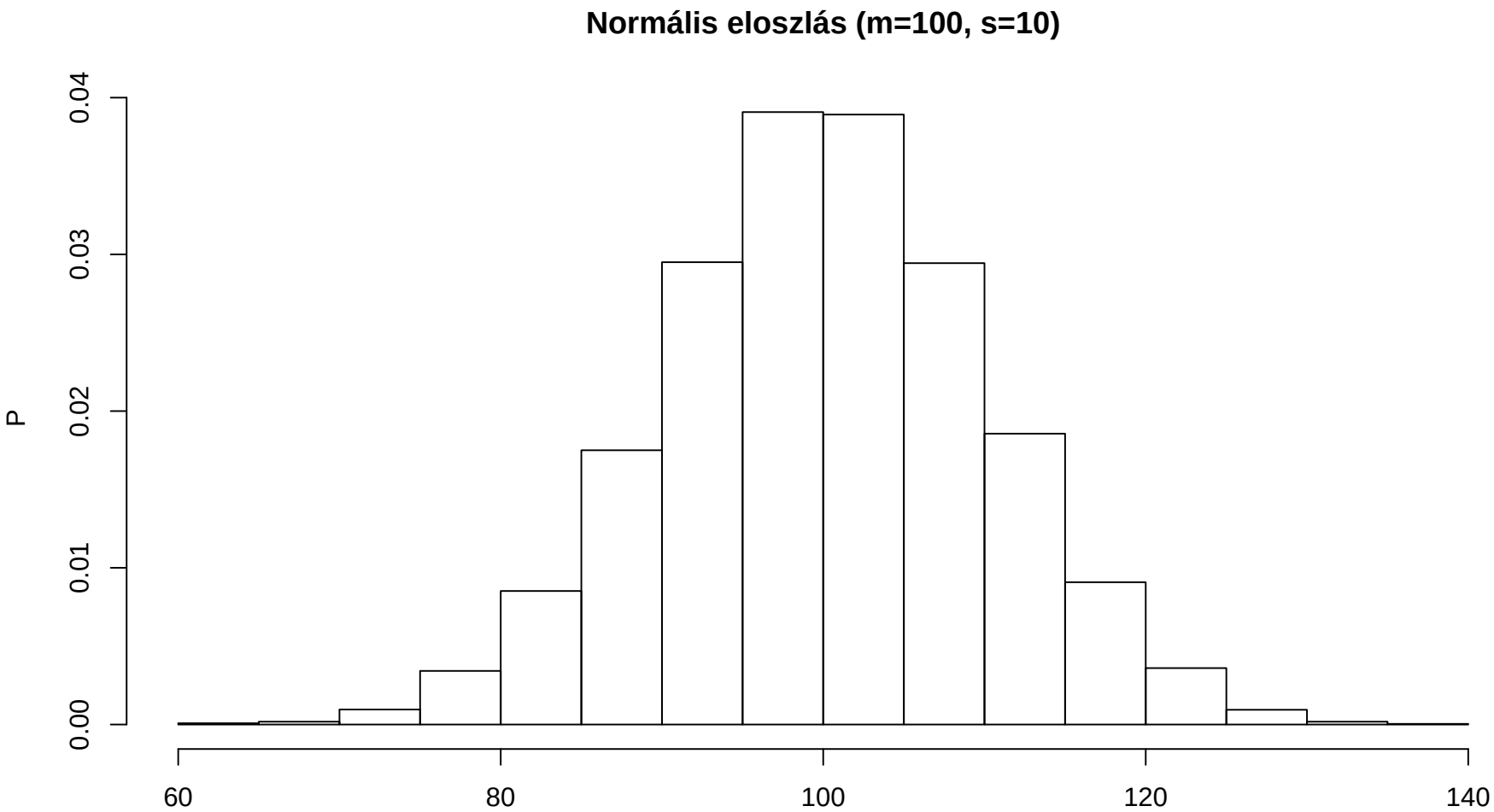
- $0,5 = \frac{10}{\sqrt{n}}$

$$\Downarrow$$

- $0,5 \cdot \sqrt{n} = 10$
- $\sqrt{n} = 20$
- $n = 400$

Mintanagyság meghatározása általános esetben

Példa



Mennyire homogén a vizsgált populáció?

Mintanagyság meghatározása általános esetben

Példa

Mekkora mintára van szükségem ahhoz, hogy 5 perc pontosság meg tudjam állapítani a napi tévénézésre fordított idő hosszát a felnőtt magyar lakosság körében?

Mintanagyság meghatározása általános esetben

Példa

Mekkora mintára van szükségem ahhoz, hogy 5 perc pontosság meg tudjam állapítani a napi tévénézésre fordított idő hosszát a felnőtt magyar lakosság körében?

- 5 perc pontosság 95 %-os döntési szinten: $SE = 2.5$,
- becsült szórás: 100
- $SE = \frac{S^*}{\sqrt{n}}$
- $2,5 = \frac{100}{\sqrt{n}}$
- $2,5 \cdot \sqrt{n} = 100$
- $\sqrt{n} = 40$
- $n = 1600$

Mintanagyság meghatározása általános esetben

Példa

Mekkora mintára van szükségem ahhoz, hogy 5 perc pontosság meg tudjam állapítani a napi tévénézésre fordított idő hosszát a felnőtt magyar lakosság körében?

- 5 perc pontosság 95 %-os döntési szinten: $SE = 2.5$,
- becsült szórás: 100
- $SE = \frac{S^*}{\sqrt{n}}$
- $2,5 = \frac{100}{\sqrt{n}}$
- $2,5 \cdot \sqrt{n} = 100$
- $\sqrt{n} = 40$
- $n = 1600$

Annál nagyobb minta kell, . . .

- minél nagyobb pontosságra törekszem,
- minél nagyobb a vizsgált változó szórása a populációban.

A mintaválasztás

A mintaválasztás célja

Miért vegyünk mintát?

- Nem áll rendelkezésre megfelelő információ, ismeret az érintett csoporton belül (pl. „egyetemisták zenehallgatási szokásai”).
- Teljes populáció megkérdezésének lehetetlensége, nehézsége (költséghatékonyság, korlátozott racionalitás).
- Az alapsokaság egyes tulajdonságainak, paramétereinek becslése annak egy kiválasztott része alapján.

Másképp: Mi célból vegyünk mintát?

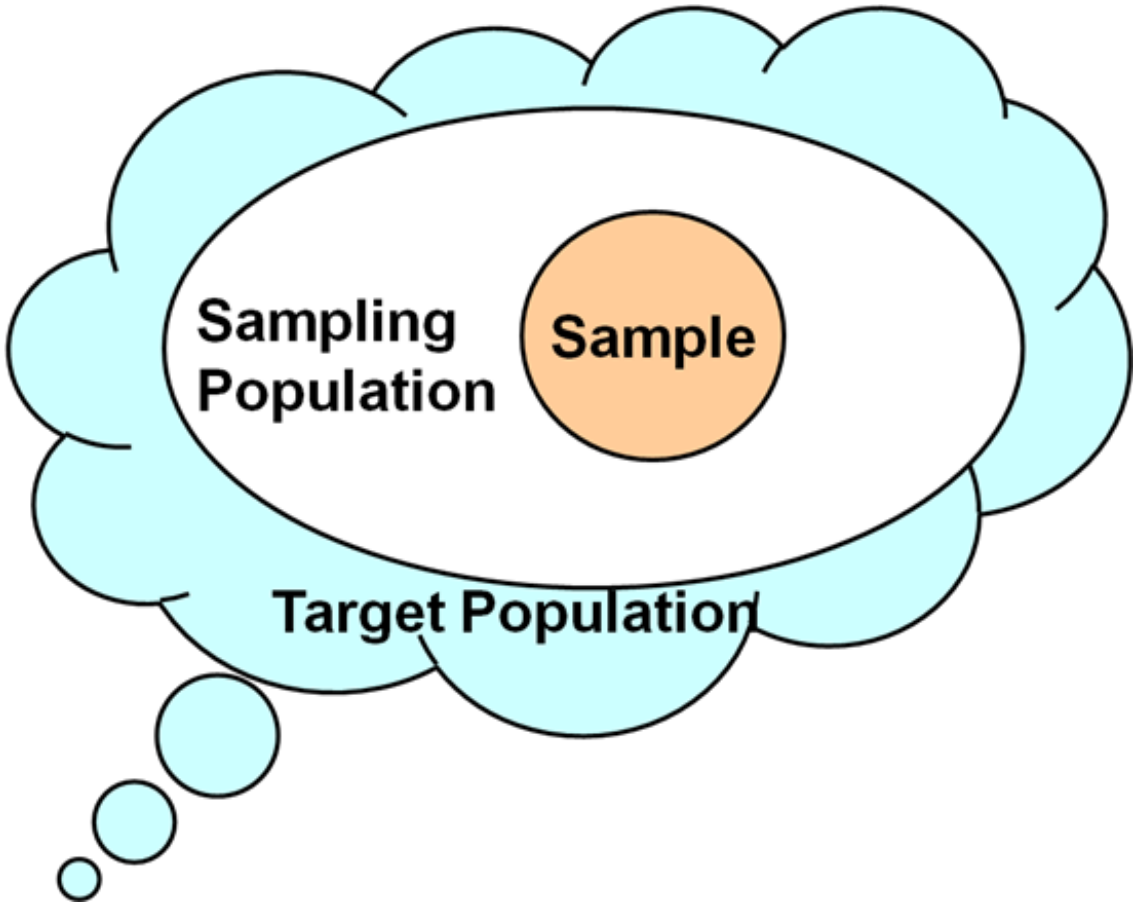
- **Felderítő kutatás**: „Vajon hallgatnak-e zenét?”
- **Leíró adatok gyűjtése**: „Milyen zenét hallgatnak?”
- **Hipotézisvizsgálat**: „Tényleg a trash-metal a legkedveltebb műfaj?”
- **Minőségbiztosítás:** „A kortárs zene jótékony hatásai?”
- **Döntéshozatal segítése**: „A büfében érdemes-e egy állandó DJ-t alkalmazni?”

- **Populáció**: azon elemek összessége, akikre általánosítani akarom, érvényesnek tekintem a mintából levont következtetéseket
- **Mintavételi keret**: a mintavételi egységekről készült lista, melyet a mintavétel céljára használunk
- **Minta**: a ténylegesen megfigyelt elemek összessége
- **Vizsgálati populáció**: a mintavételi keret által lefedett populáció
- **Elem**: akikről információt gyűjtünk, akikről hipotéziseink szólnak
- **Mintavételi egység**: az elemek vagy azok valamilyen csoportja

I. *Mintavételi egység, válaszadó, megfigyelési egység, eset*

A mintaválasztás

A mintaválasztás lépései



Forrás: http://www.femwiki.com/fem/w/wiki/concepts-in-sampling.aspx

Célcsoport meghatározása:

- a populáció a kutatás tárgyának függvénye.

Vizsgált csoport meghatározása:

- nem mindig tudatos döntés eredménye,
- szisztematikus (l. konceptualizálás).

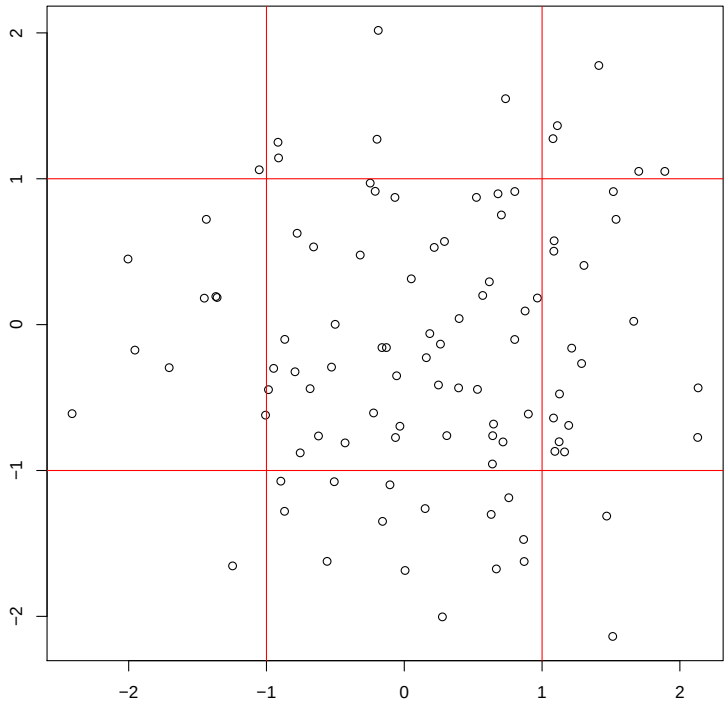
Mintavételi keret meghatározása:

- rendelkezésre álló erőforrások függvénye.

Elemek meghatározása:

- valószínűségi mintavétel,
- nem-valószínűségi mintavétel,.

- **Alapsokaság**: a vizsgálati populáció (elméletileg meghatározott) összes eleme
- **Mintavételi keret**: a kiválasztásnál figyelembe vett (elérhető) elemek összessége
- **Megfigyelési egység**: az alapsokaság elemeinek tekintett egységek
- **Mintavételi egység**: a kiválasztásnál figyelembe vett, legelemibb egységek



Az elemek kiválasztása alapvetően két, egymástól jól elkülöníthető módszer szerint történhet.

Valószínűségi mintavétel:

amikor a populáció (l. mintavételi keret) minden eleme (a mintavétel előtt már) ismert, nem nulla (és egyenlő) eséllyel kerül a mintába.

VAGY

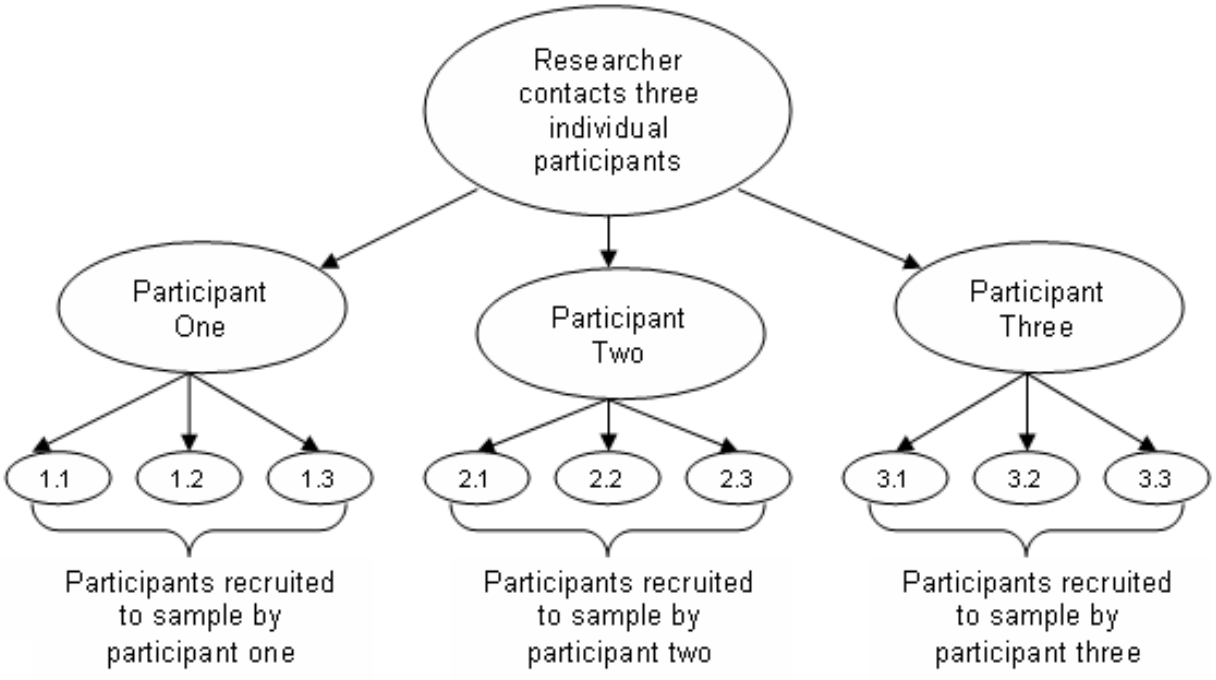
Nem-valószínűségi mintavétel:

minden egyéb kiválasztás.

A mintaválasztás

Nem-valószínűségi mintavételi eljárások

- A minta elemei nem valószínűségi alapon kerülnek kiválasztásra.
- A minta torzítása (alapsokaságtól való eltérése, ill. annak nagysága) matematikailag nem számítható.
- Bizonyos helyzetekben elkerülhetetlen (vagy nem célszerű) az alkalmazásuk.
- Gyors, költséghatékony módszerek.



Általánosan ismert típusai:

- hólabda módszer,
- egyszerűen elérhető alanyok megkeresése,
- szakértői mintavétel,
- kvótás mintavétel:
  - **1936**: Gallup vs. Literary Digest
  - Az alapsokaságot leíró mátrix alapján kerülnek az egyedek kiválasztásra.
  - A súlyozás következtében bizonyos jellemzők mentén reprezentatív a minta.
  - Hátrányok és buktatók:
    - l. 1948
    - Megfelelő kiindulási mátrix
    - Megfelelően kiválasztott tulajdonságok
    - „Marginális” elemek figyelmen kívül hagyása

Miért nem valószínőségiek ezek a technikák?

- Vannak olyan elemek, melyek sose kerülhetnek a mintába – például:
  - kvótás mintavételnél: megfigyelő számára antipatikus emberek
  - az egyszerűen nem elérhető alanyok
  - hólabdás mintavételnél: a mintába került alanyok által nem ismert emberek
- Nem ismert az a valószínőség, amivel a kiválasztott elemek a mintába kerültek
  - A mintába kerülés esélye szubjektív – ismertségen, szimpátián alapul
  - A szubjektív esélyt nem tudjuk számszerűsíteni

- Minden egyed azonos (pontosabban: meghatározott) valószínűséggel kerül kiválasztásra.
- A valószínűségek alapján lehetőség nyílik a mintavétel során elkövetett hibát (*mintavételi hiba*), a minta torzítását számolni, azaz meghatározni azt, hogy minta által felvázolt jellemzők milyen jól jellemzik az alapsokaságot.

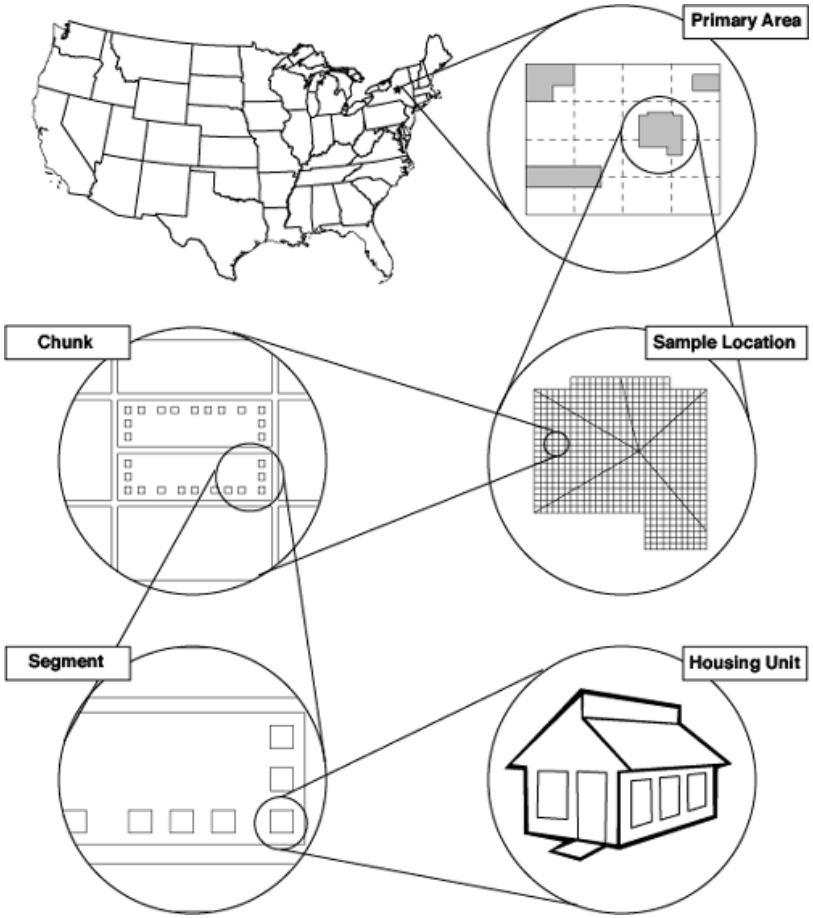
Reprezentativitás

Az alapsokaság minden eleme meghatározott, nem nulla (egyenlő) valószínűséggel kerülhet kiválasztásra.

- **Megbízhatósági szint:** a minta alapján számolt becslések milyen valószínűséggel lesznek igazak az alapsokaság tagjaira.
- **Mintavételi hiba:** a minta alapján becsült paraméter milyen mértékben ingadozik a valós érték körül (konfidencia intervallum).

Típusai:

- Egyszerű véletlen mintavétel (SRS)
- Szisztematikus véletlen mintavétel (systematic sampling)
- Rétegzett mintavétel (stratified sampling)
- Rétegzett-szisztematikus mintavétel
- Csoportos mintavétel (cluster sampling)
- Többlépcsős technikák (multi-stage sampling)

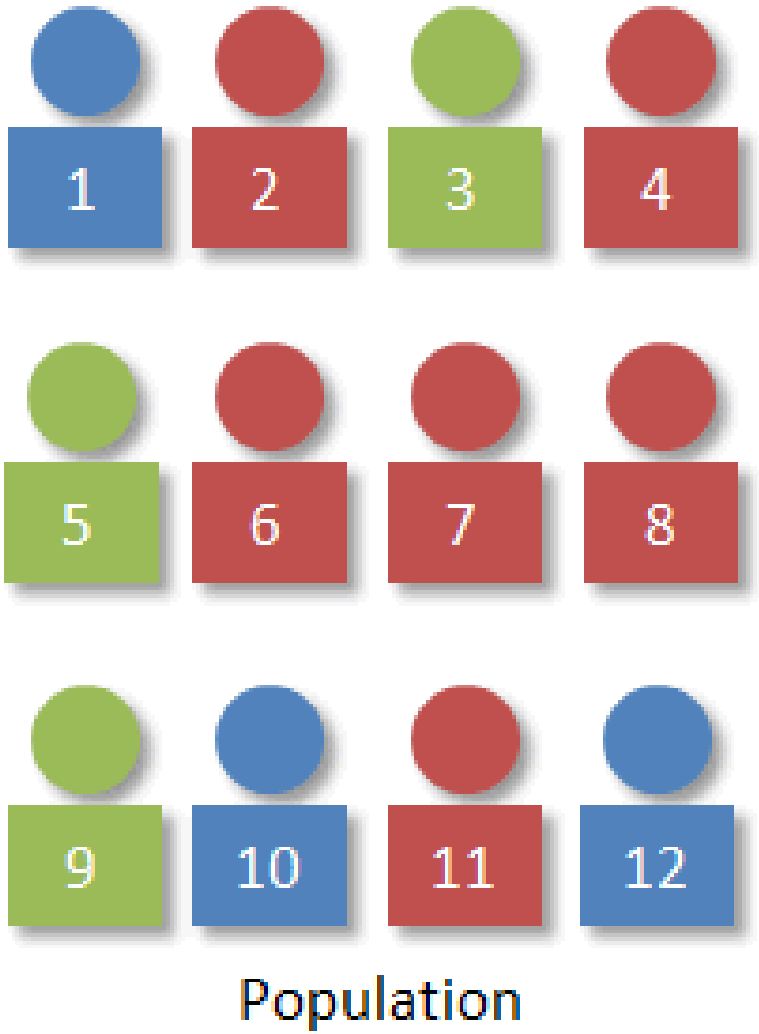


Véletlen mintaválasztás történik-e, amikor:

- Budapest egy véletlen módon kiválasztott buszmegállójában megkérdezek minden harmadik embert?
- a reprezentativitás elérése érdekében a mintatagokat az alapsokaság arányában választjuk ki: a kérdezőbiztosok 40 férfit és 60 nőt kérdeznek le a Campus véletlen bejárása során.
- felhívunk 500 számítógép által generált telefonszámot (l. 2+7 véletlen számjegy)?
- a kurzus látogatóinak neveivel ellátott papírfecniket tartalmazó kalapból csupa női nevet húzunk ki?
- matematika órán a tanár a napló „felcsapásával” választja ki a felelőt?

Egyszerű véletlen mintavétel

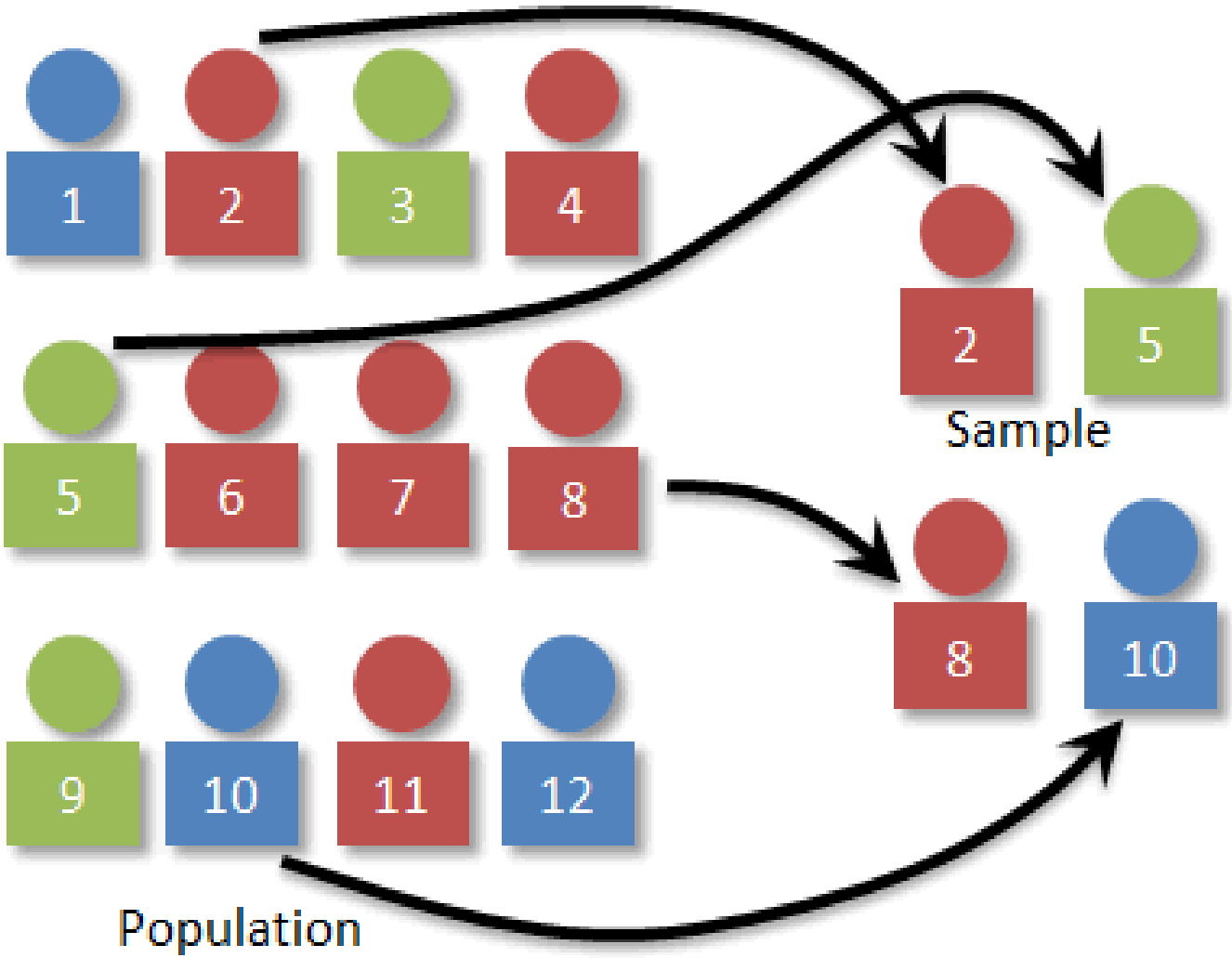
A kiválasztás menete



Forrás: Dan Kerlner, Elgin Community College

Egyszerű véletlen mintavétel

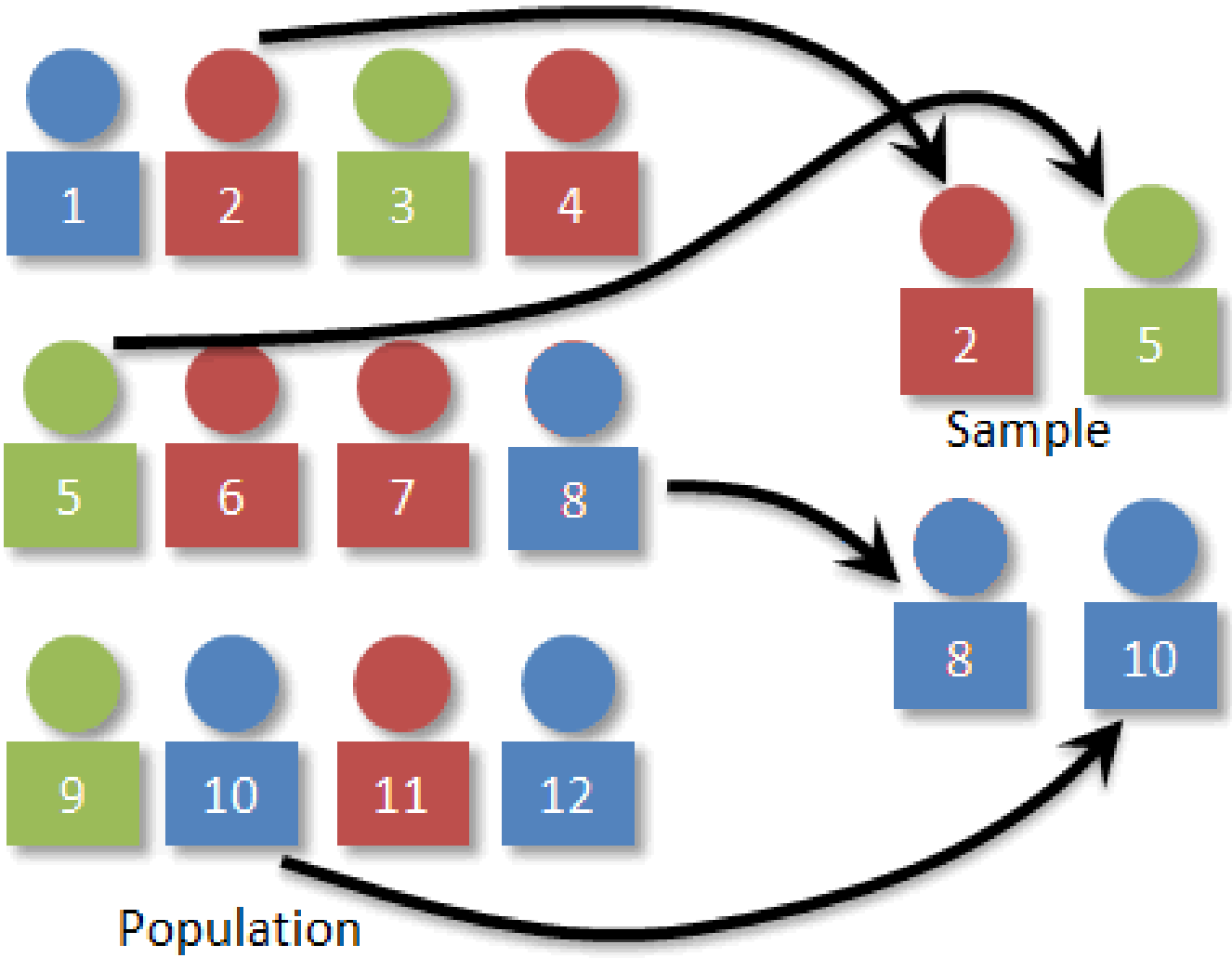
A kiválasztás menete



Forrás: Dan Kerlner, Elgin Community College

Egyszerű véletlen mintavétel

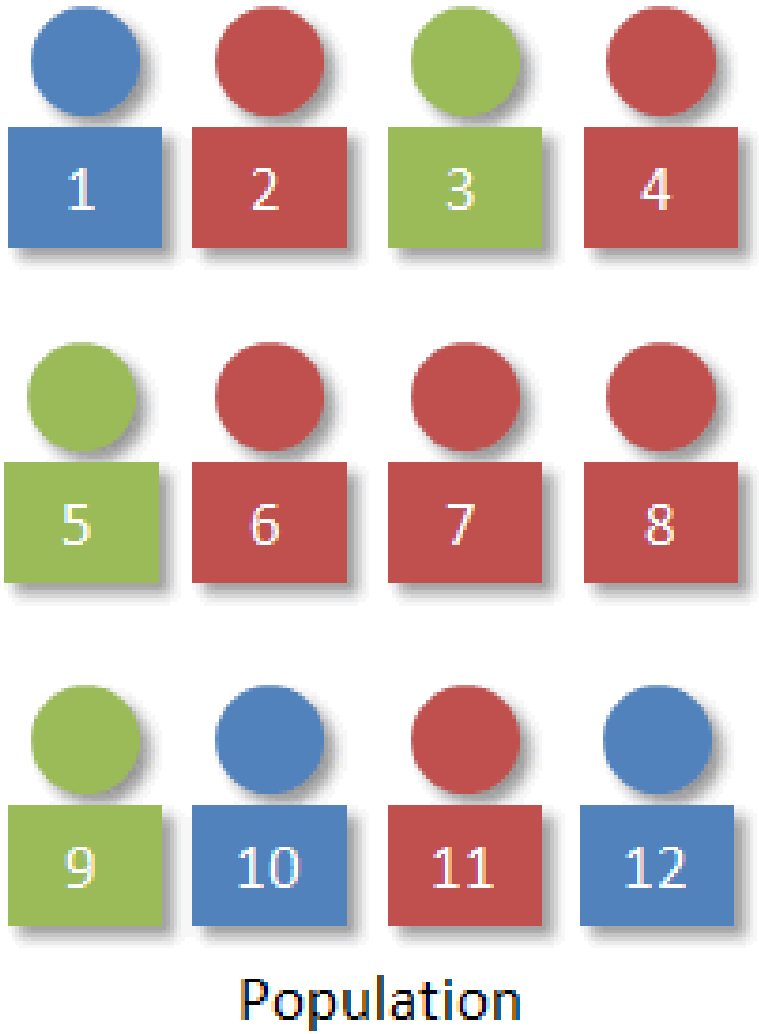
A kiválasztás menete



Forrás: Dan Kerlner, Elgin Community College

Rétegzett mintavétel

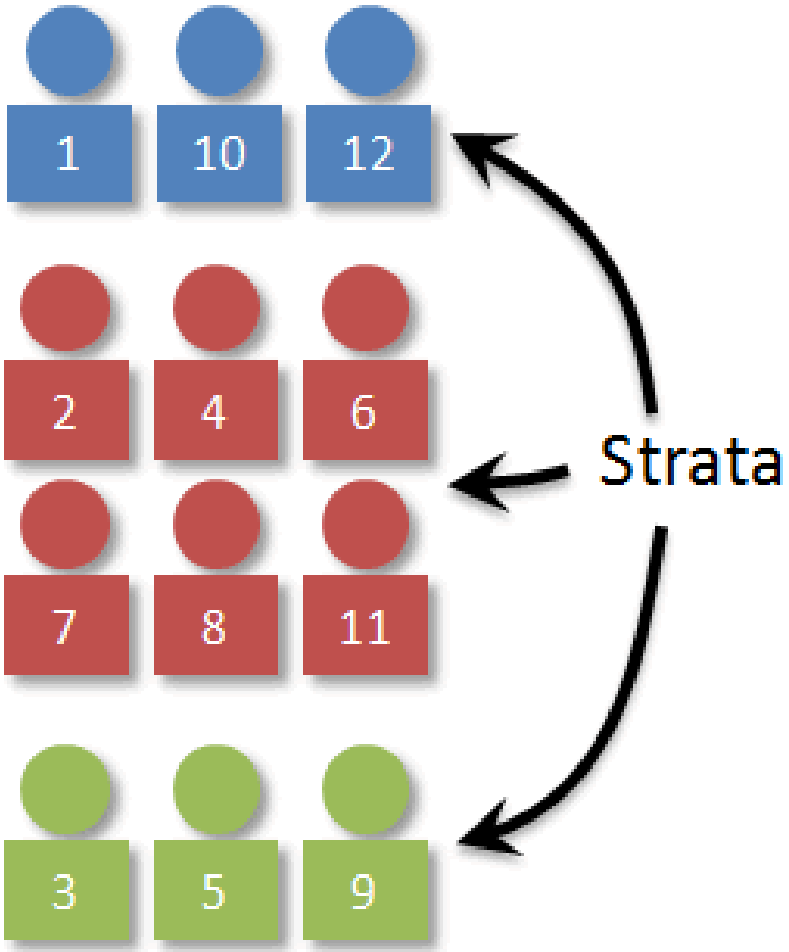
A kiválasztás menete



Forrás: Dan Kerlner, Elgin Community College

Rétegzett mintavétel

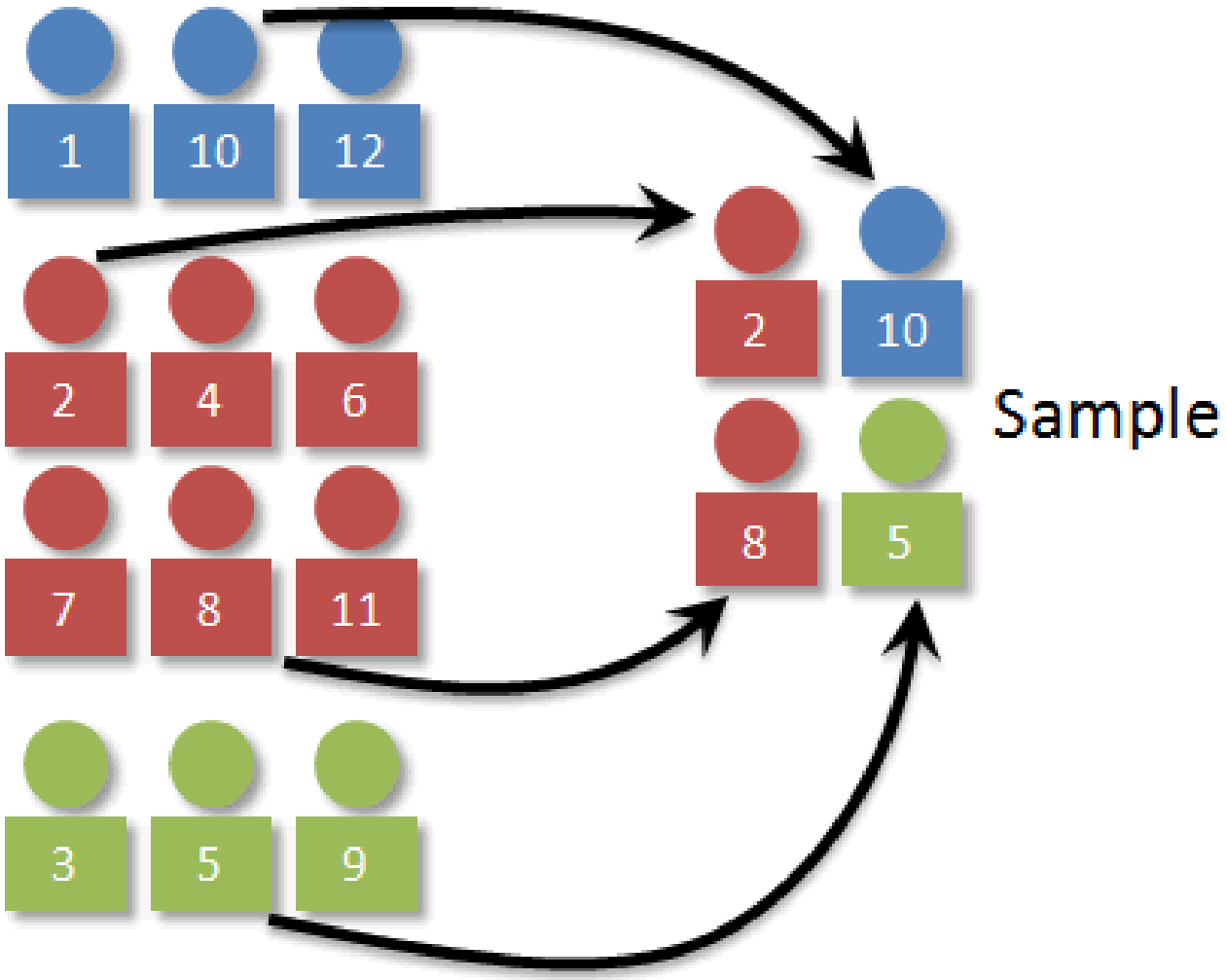
A kiválasztás menete



Forrás: Dan Kerlner, Elgin Community College

Rétegzett mintavétel

A kiválasztás menete



Forrás: Dan Kerlner, Elgin Community College

Rétegzett mintavétel

A mintavételi hiba

Négy hallgatót kérdeztünk meg arról, hogy hány macskát tart otthon:

	Budapest	vidék
Lányok	9	7
Fiúk	3	1

Mit gondolunk, hogyan alakulna a mintavételi hiba ha egyszerű véletlen mintát, és hogyan, ha rétegzett mintát vennénk?

Négy hallgatót kérdeztünk meg arról, hogy hány macskát tart otthon:

	Budapest	vidék
Lányok	9	7
Fiúk	3	1

Mit gondolunk, hogyan alakulna a mintavételi hiba ha egyszerű véletlen mintát, és hogyan, ha rétegzett mintát vennénk?

Két fős mintákat választva:

1. SRS: 6 lehetséges minta: (1,7) (1,9) (3,7) (3,9) (1,3) (7,9)

$$\overline{x} = \frac{4+5+5+6+2+8}{6} = 5, S^* = \frac{1+0+0+1+9+9}{6} = 3.33$$

2. Rétegzett: 4 lehetséges minta: (1,7) (1,9) (3,7) (3,9)

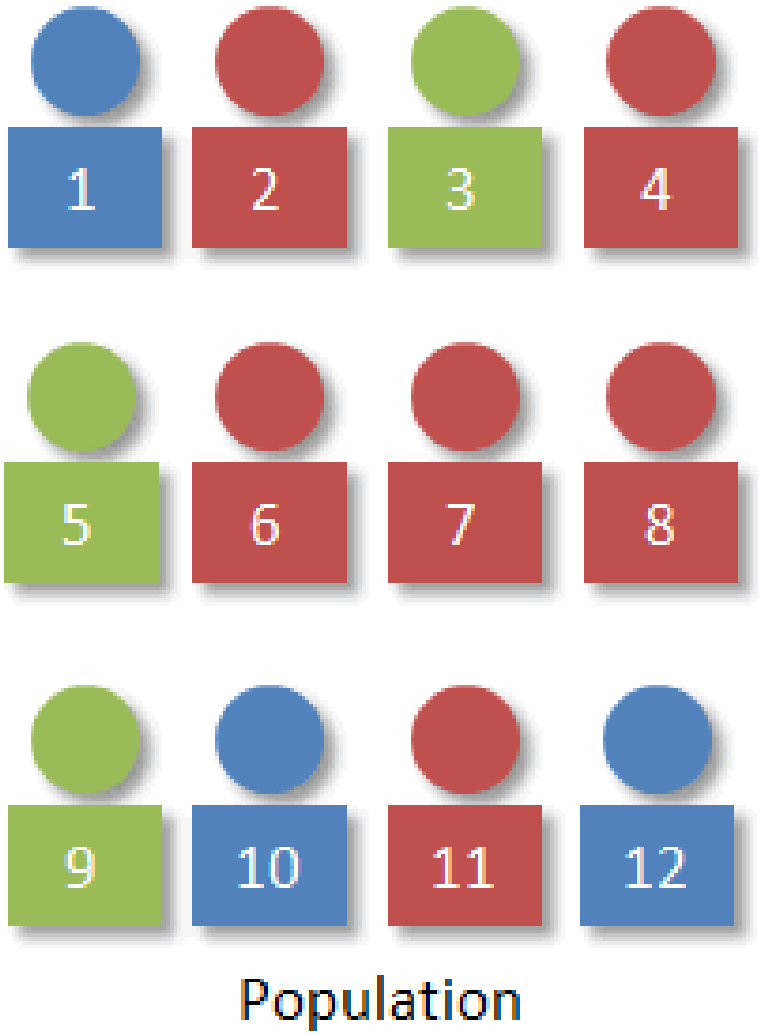
$$\overline{x} = \frac{4+5+5+6}{4} = 5, S^* = \frac{1+0+0+1}{4} = 0.5$$

3. Rétegzett: 4 lehetséges minta: (1,3) (1,9) (3,1) (3,7)

$$\overline{x} = \frac{2+5+2+5}{4} = 3.5, S^* = \frac{1.5+1.5+1.5+1.5}{4} = 1.5$$

Szisztematikus mintavétel

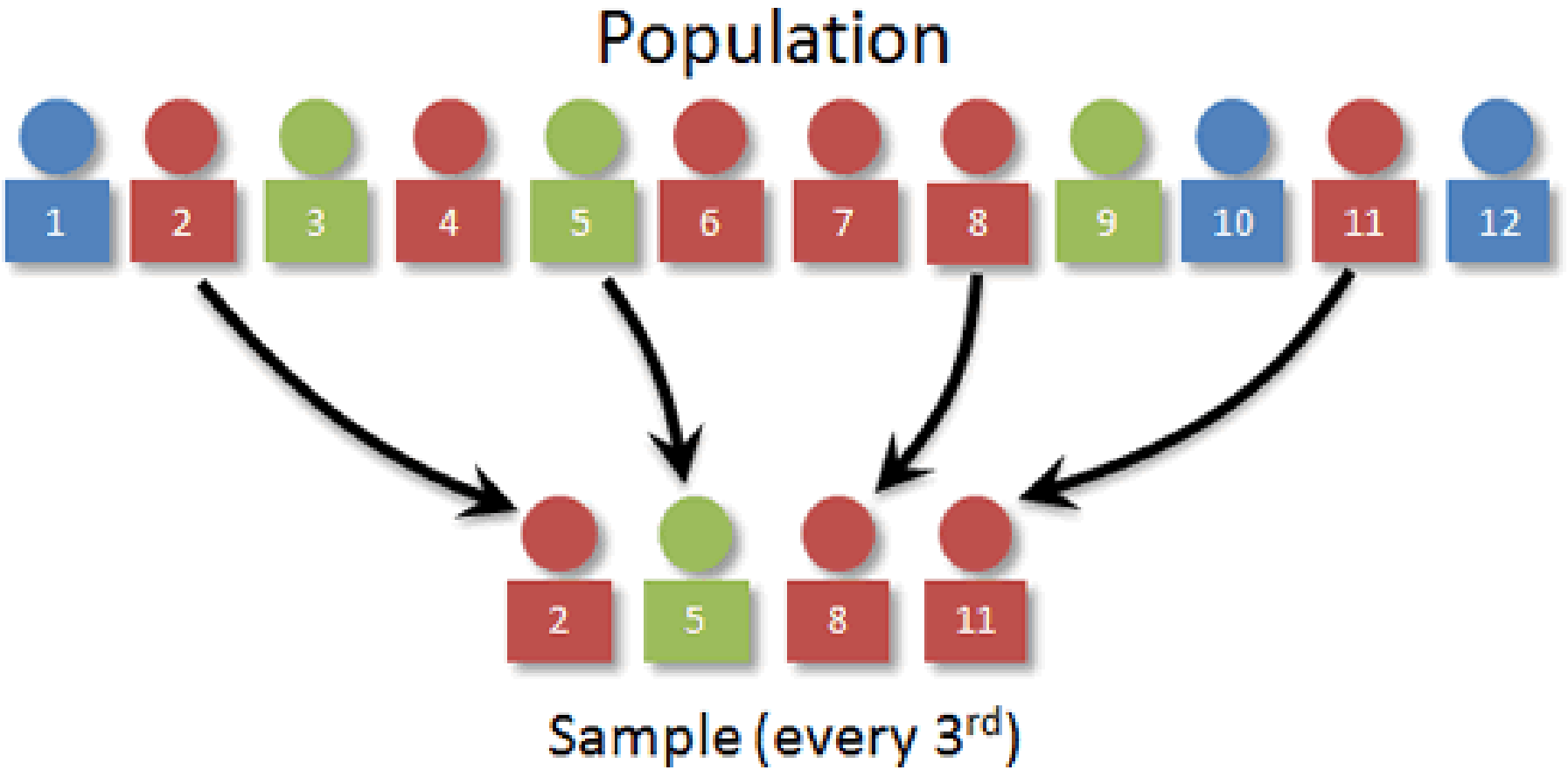
A kiválasztás menete



Forrás: Dan Kerlner, Elgin Community College

Szisztematikus mintavétel

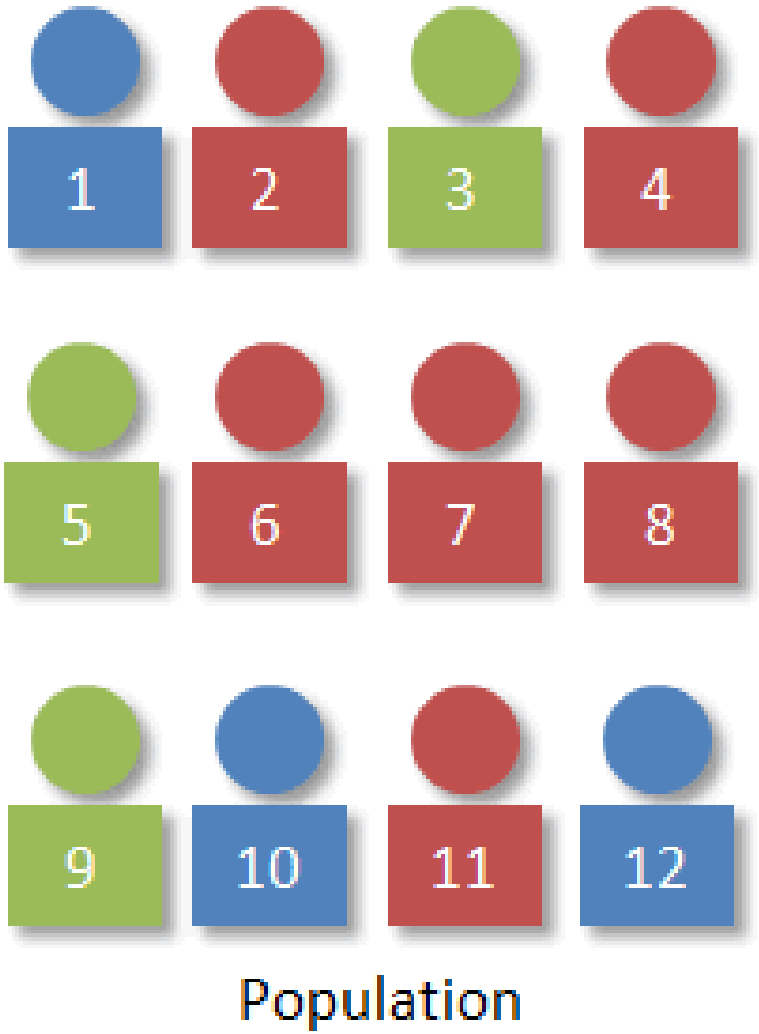
A kiválasztás menete



Forrás: Dan Kerlner, Elgin Community College

Szisztematikus-rétegzett mintavétel

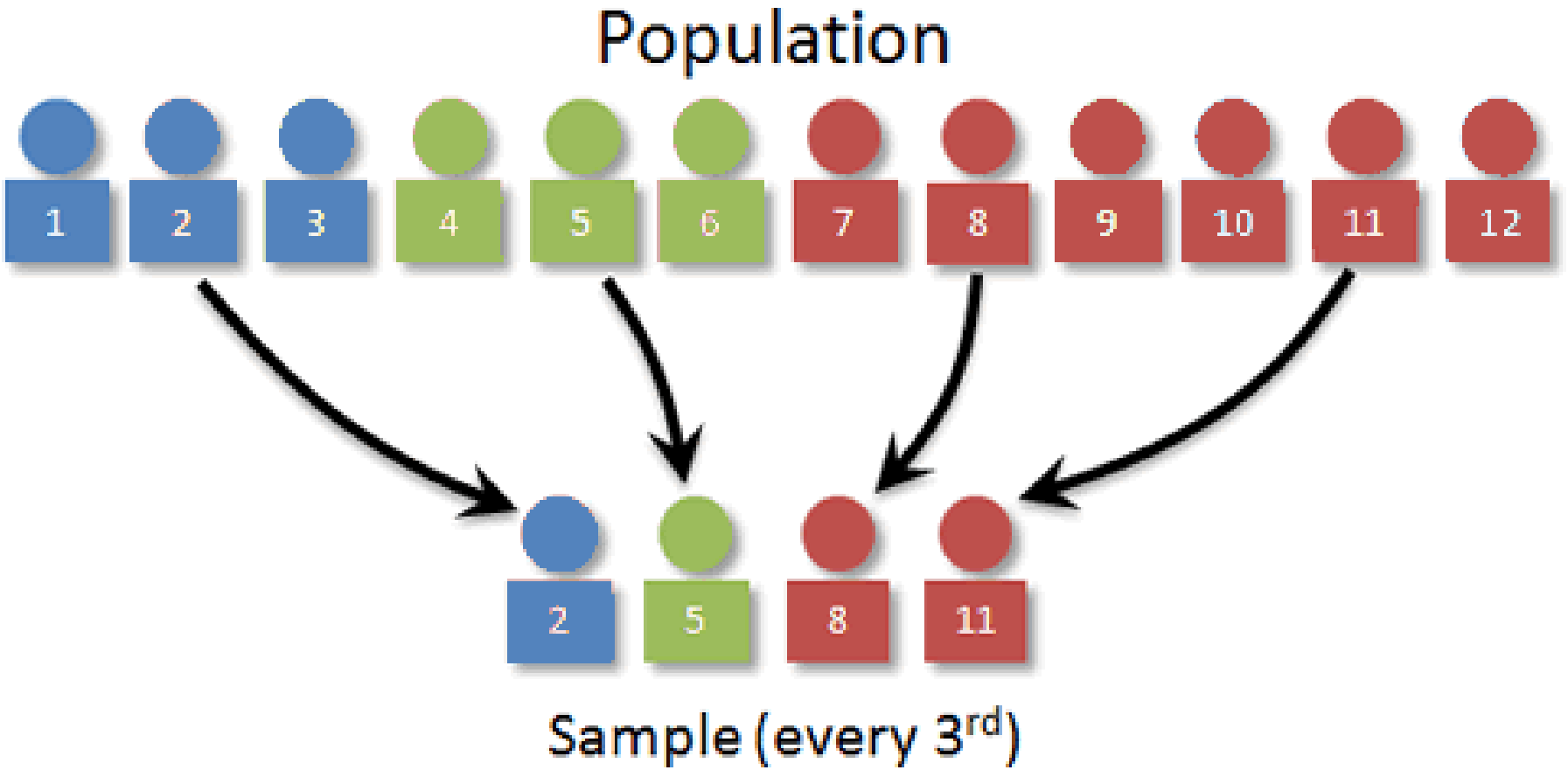
A kiválasztás menete



Forrás: Dan Kerlner, Elgin Community College

Szisztematikus-rétegzett mintavétel

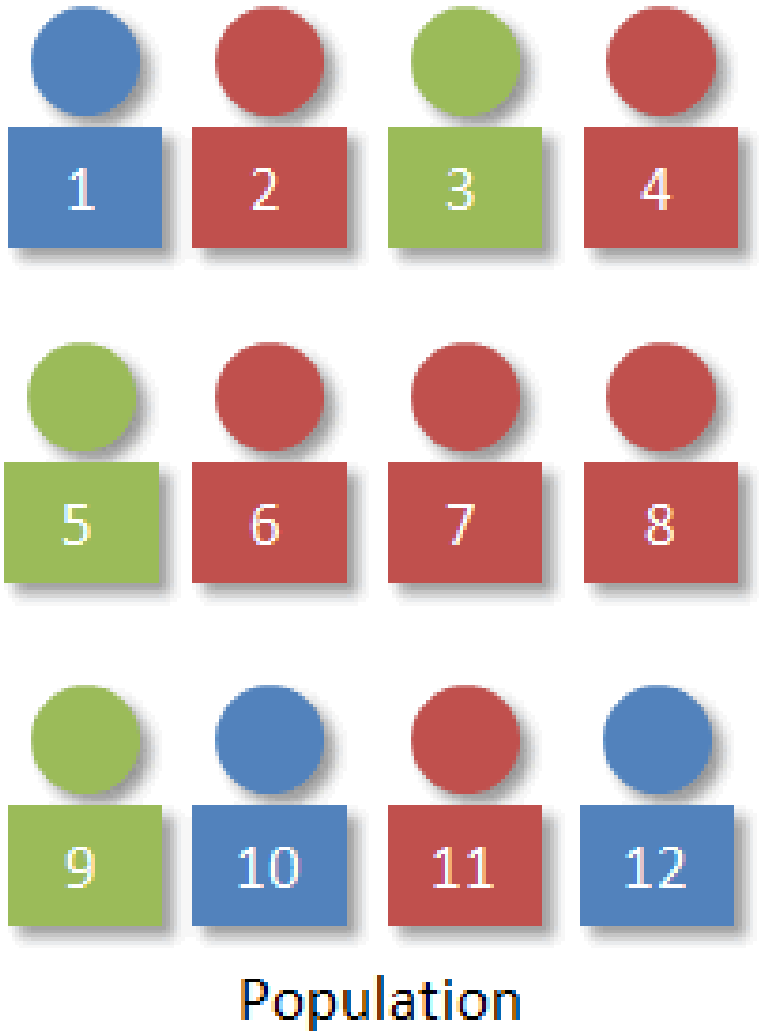
A kiválasztás menete



Forrás: Dan Kerlner, Elgin Community College

Csoportos mintavétel

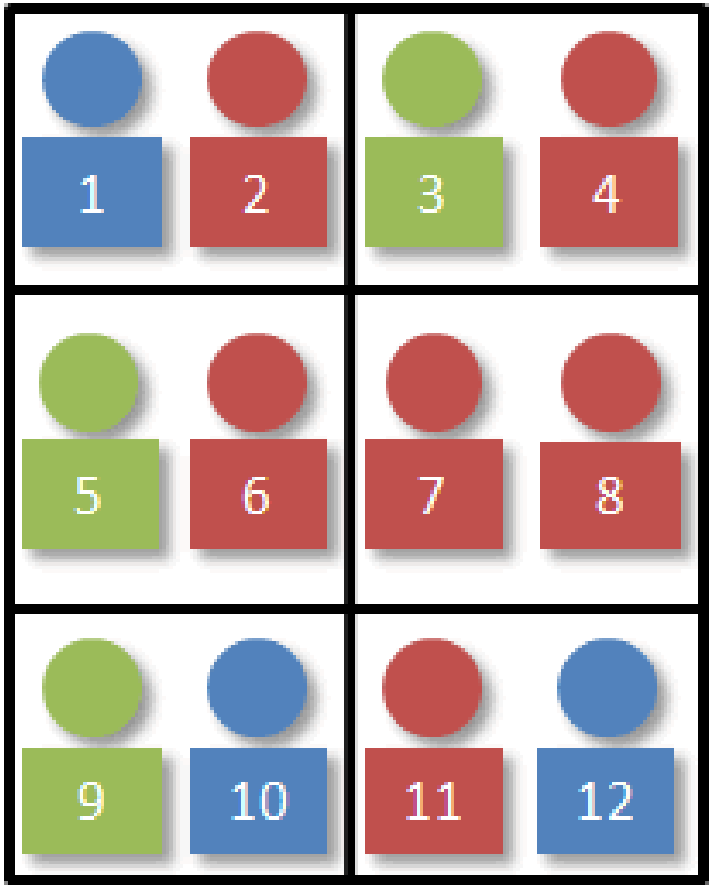
A kiválasztás menete



Forrás: Dan Kerlner, Elgin Community College

Csoportos mintavétel

A kiválasztás menete

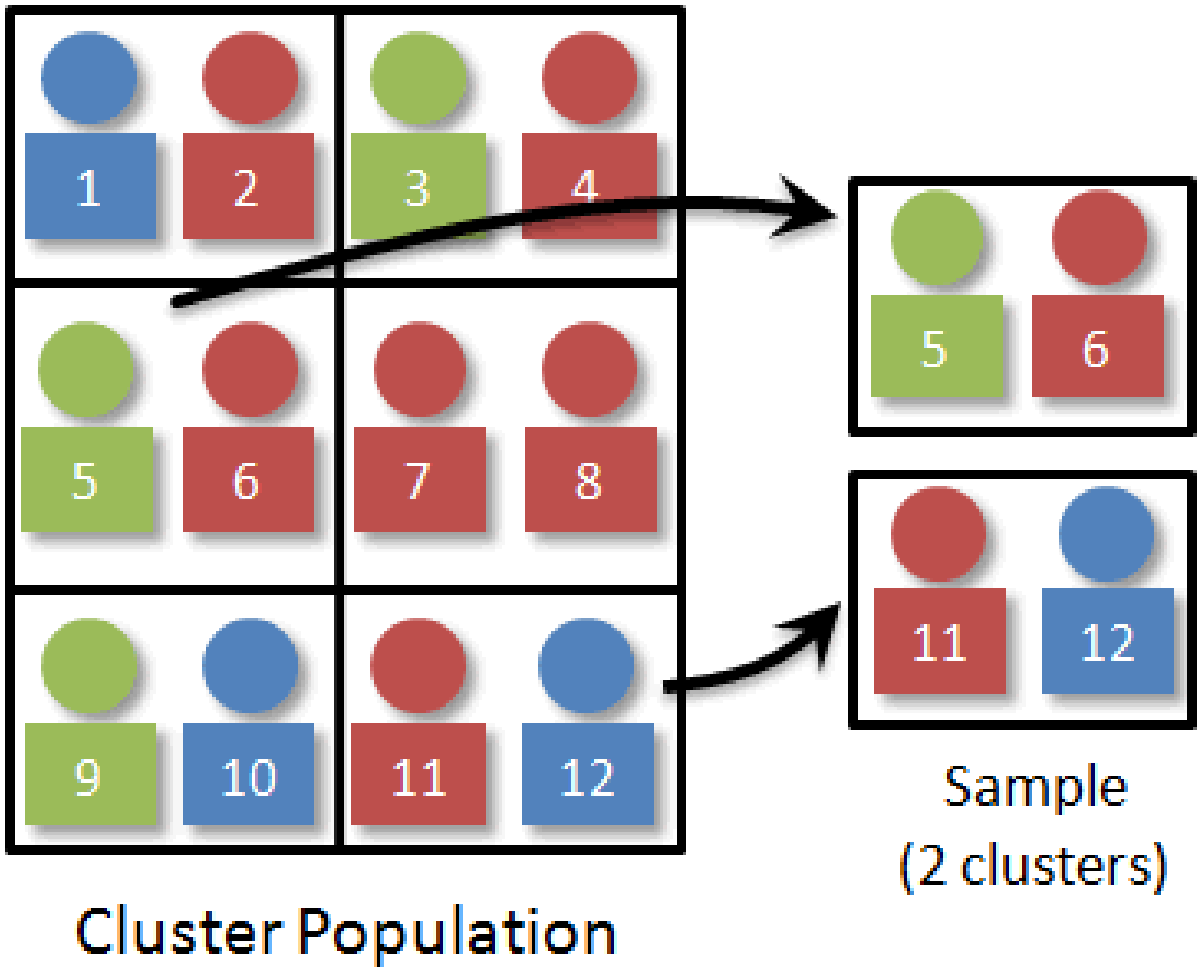


Cluster Population

Forrás: Dan Kerlner, Elgin Community College

Csoportos mintavétel

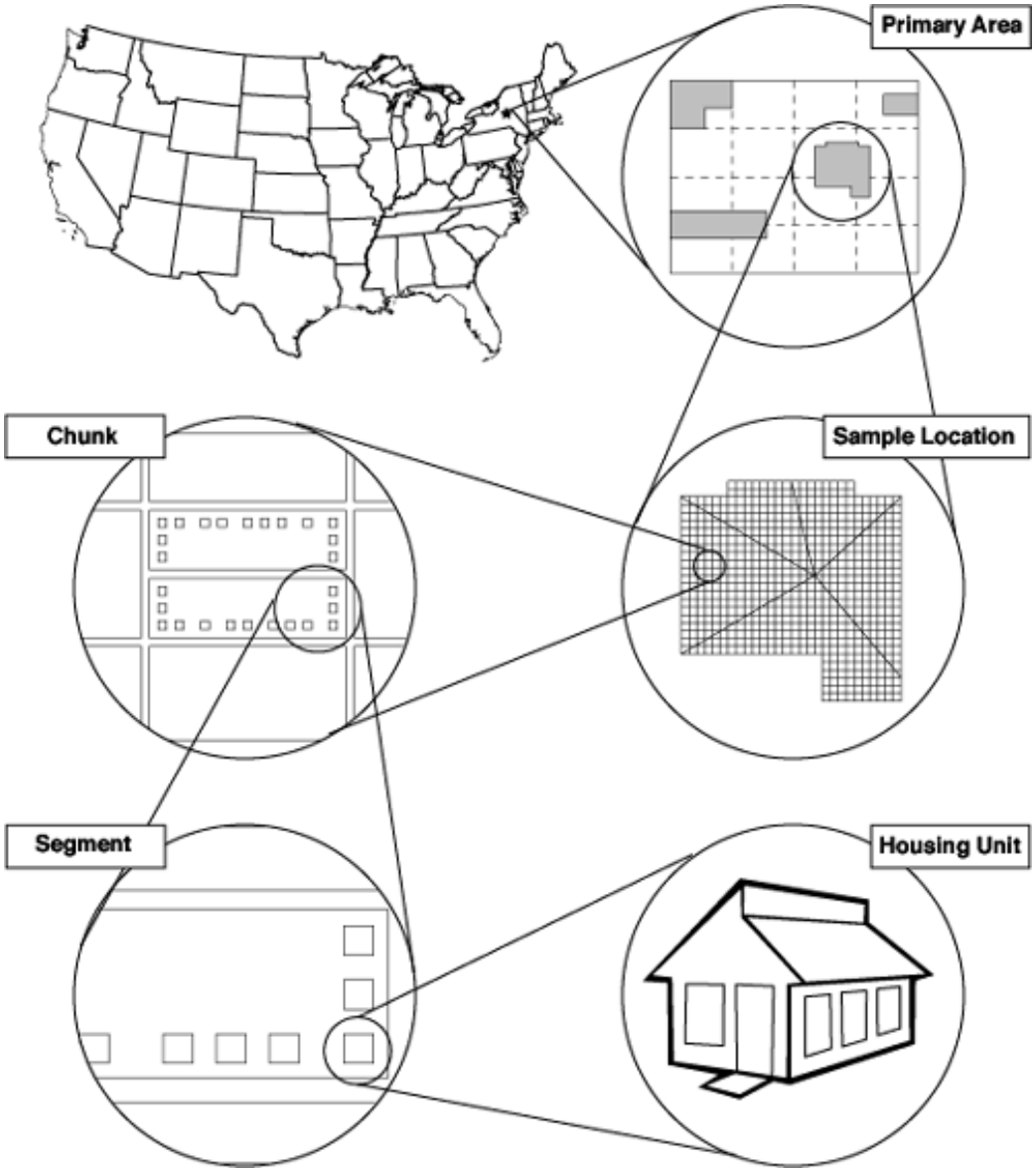
A kiválasztás menete



Forrás: Dan Kerlner, Elgin Community College

Csoportos mintavétel

A kiválasztás menete



Rétegzett mintavétel:

Kvótás mintavétel:

	Nő	Férfi	$\sum$
Elméleti matematika	10	10	20
Környezettudomány	40	10	50
Rendezvényszervező	10	20	30
$\sum$	60	40	100

Köszönöm a figyelmet!

Daróczi Gergely
daroczi.gergely@btk.ppke.hu