Elemi statisztika fizikusoknak

Pollner Péter

Biológiai Fizika Tanszék

pollner@elte.hu

Mire jó a statisztika?

- **Mérési eredmények kiértékelésére**
- **Kísérletek megtervezésére**
- **Számítógéppel szimulált adatok feldolgozására**

Használják

- **a tudományokban (fizika, kémia, biológia, …**
- **a fejlesztésben (mérnökök, orvosok …**
- **a technológiában (minőségbiztosítás …**
- **a gazdaságban (marketing, vállalati statisztika …**
- **a kormányzásban (felmérések, népszámlálások …**

1. Előadás
Bevezetés a statisztikába

1-1. rész
Áttekintés

Áttekintés

A mérések, felmérések és más adatgyűjtési eszközök célja, hogy egy nagy csoport kis részéről gyűjtsünk be adatokat annak érdekében, hogy megtudjunk valamit a nagy csoportról. Ezen az előadáson arról lesz szó, hogy mire kell ügyelnünk eközben.

Példák:

•„Szokott ön időnként alkoholos italokat, mint sör, bor vagy égetett szeszes italok, fogyasztani vagy ön teljesen absztinens ?” A megkérdezettek válaszaiból (pl. 1000 ember) próbálunk a teljes népességre (pl. 10000000 ember) következtetni.

• Népszámlálás (Cenzus) Megpróbálunk mindenkit megkérdezni.

Definíciók

- **Adat**

  összegyűjtött megfigyelések (mérések, kérdőíves válaszok, felmérések)

- **Statisztika**

  adatokon alapuló kísérlettervezési, gyűjtési, rendezési, összesítési, ábrázolási, analizálási, értelmezési és következtetési módszerek összessége

Definíciók

❖Populáció (alapsokaság)

a tanulmányozandó elemek összessége, teljessége (pl. eredmények, mérések, stb.). A gyűjtemény **teljes** abban az értelemben, hogy tartalmaz minden tanulmányozandó tárgyat.

❖Cenzus

adatok gyűjteménye a populáció minden eleméről

❖Minta

a populációból kiválasztott elemek rész halmaza

1-2.
Az adatok típusai

Definíciók

❖ Paraméter

a populációt jellemző numerikus érték

populáció

paraméter

Definíciók

❖Statisztika

a mintát jellemző numerikus érték.

minta

statisztika

A véletlen szerepe

- **A mintát megfelelő módon kell gyűjteni, mint amilyen a véletlen kiválasztás.**

- **Ha az adatok nem így lettek gyűjtve, akkor általában statisztikai módszerekkel sem lehet ezt kijavítani, az adatokat nem lehet használni.**

Definíciók
számosság szerint

- **Kvantitatív adatok** méréseket vagy leszámlálásokat jellemző számok.

Pl.: az emberek súlya

- **Kvalitatív (kategória vagy tulajdonság) adatok**

kategóriákra bonthatók, melyek valamilyen nem-numerikus jellemzők alapján különböztethetők meg
Példa: profi atléták nemei (férfi/nő).

Kvantitatív adatok

- **Kvantitatív adatok két altípusa**

❖Diszkrét

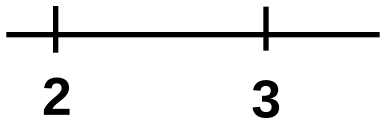
amikor a lehetséges adatok száma véges vagy legalábbis megszámlálható.

0, 1, 2, 3, . . .

Példa: Tyúkok által tojt tojások száma.

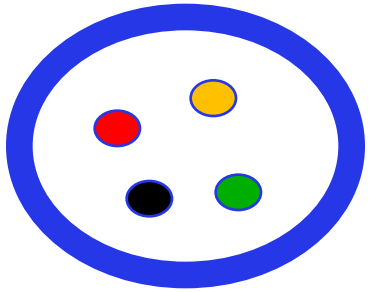
❖Folytonos

(numerikus) adat ami végtelen sok lehetséges értéket vehet fel valamilyen folytonos skálán, és nincsenek benne lyukak, szakadások.



Pl.: A tehén által naponta adott tej mennyisége (8.86965517 liter) .

Definíciók
szintek szerint



- **nominális szintű mérések**

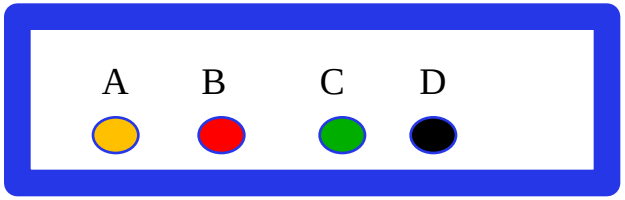
elnevezéseket, címkéket, vagy kategóriákat tartalmazó adatok, melyeket nem lehet valami szerint rendezni (pl. kicsitől nagyig)

Példa: **kérdőíves válasz igen, nem, nem tudom**

SZÍN	GYAKORISÁG
Fekete	7
Piros	1

SZÍN	GYAKORISÁG
Piros	1
Fekete	7

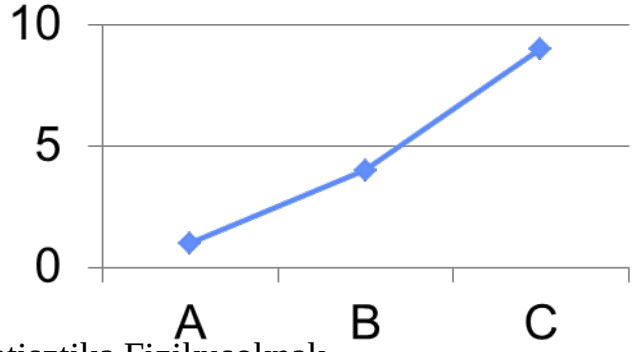
Definíciók
szintek szerint

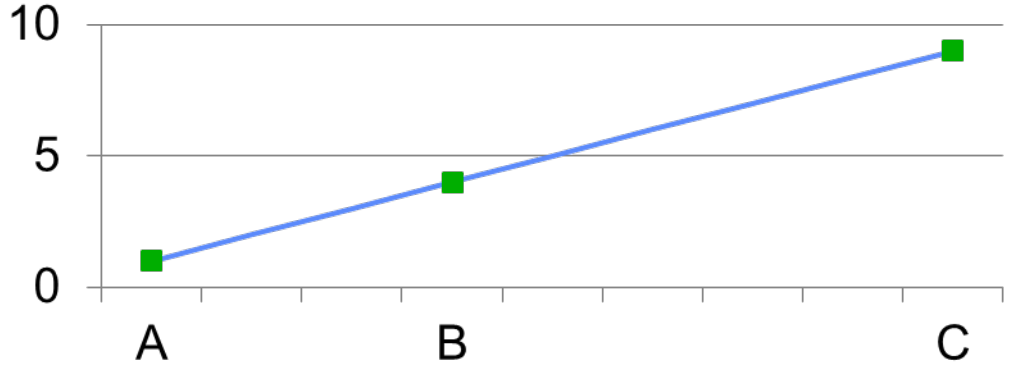


- **ordinális szintű mérés**

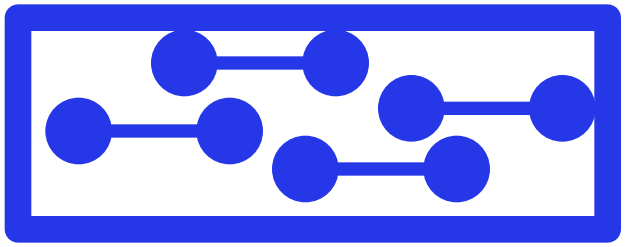
olyan adatok, melyeket lehet rendezni, de az adatok közti különbségeknek nincs értelmük, vagy nem lehet meghatározni, vagy értékpáronként eltérő a jelentése

Példa: Egyetemek sorrendje, érdemjegyek jeles, jó, közepes, elégséges vagy elégtelen, szavak ábc sorrendje





Definíciók
szintek szerint



- **intervallum szintű**

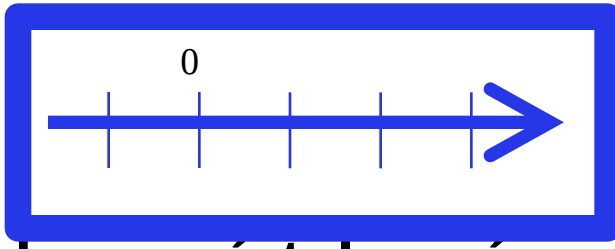
rendezhető adatok, melyeknél a különbségnek is van értelme, de nincs természetes 0 pont (olyan, ami valamilyen mennyiség jelen nem létét jelezné) és nincs értelme az arányoknak

Példa: **évek 1000, 2001, 1848, és 1526, hőmérséklet Celsius fokban**

Értelmetlen kijelentés: az évszám négyzetével nő a GDP

Definíciók

❖ arány vagy abszolút szint

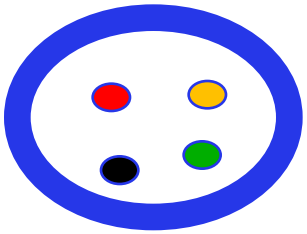


az adatok rendezhetők, a különbségnek van értelme és van természetes 0 pont, ami azt jelzi, hogy az adott mérendő mennyiség nincs jelen egyáltalán. Ebben az esetben az arányoknak is van értelme.

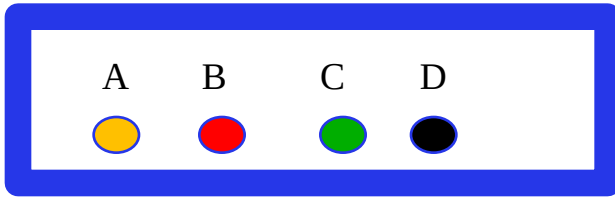
Példa: **A tankönyvek ára (0 Ft azt jelenti, hogy nem kerül semmibe), hőmérséklet Kelvin fokban**

Összefoglalás - A mérések szintjei

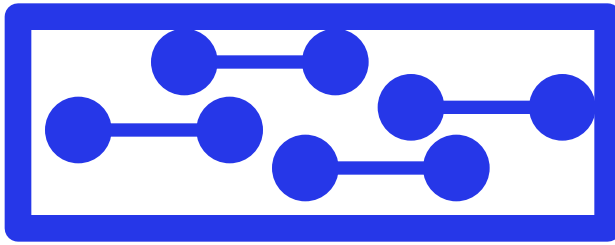
- **Nominális** – csak kategóriák



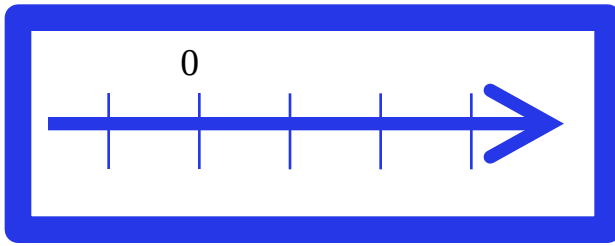
- **Ordinális** – kategóriák és rendezhetőség



- **Intervallum** – különbségek de nincs természetes 0 pont



- **Arány** – a különbségek és arányok értelmesek és létezik természetes kezdőpont



Ismétlés

Az 1-1 és 1-2 fejezetekben volt:

Néhány az adatokat jellemző kulcsfogalom

- **Paraméter vs. statisztika**
- **Az adatok fajtái (kvantitatív és kvalitatív)**
- **A mérések szintjei**

1-3 fejezet
Kritikus gondolkodás

A statisztikai módszerek sikere és buktatói

- Az elemi statisztikai módszerek használatakor a józan ész fontosabb mint a matematikai jártasság.

  Manapság a számítógépek és szoftver csomagok nagyban megkímélnek az elemi számítások elvégzésétől, de nekünk kell tudnunk, hogy mit miért csinálunk, és hogyan interpretáljuk az eredményeket.

- Most átnézzük, hogy általában mire kell ügyelni az adatok gyűjtésénél és interpretálásánál.

Buktatók

- Rossz minták: Készakarva rosszul készített felmérések, laboratóriumi mérések

  HF.: Dobj fel 500-szor egy pénzdarabot és írd le az eredményt! 6-szor egymásután fej?

  Benford törvény: az első digit 1 (30%), 2 (18%), 3 (12%), 4 (10%), 5 (8%), 6 (7%), 7 (6%), 8 (5%), 9 (5%)

  Publikációs elfogultság: csak vagy főleg olyan mintákat mutatok be, amik alátámasztják a megállapításaimat, az ellenpéldákat nem vagy nem a valóságos arányban mutatok be

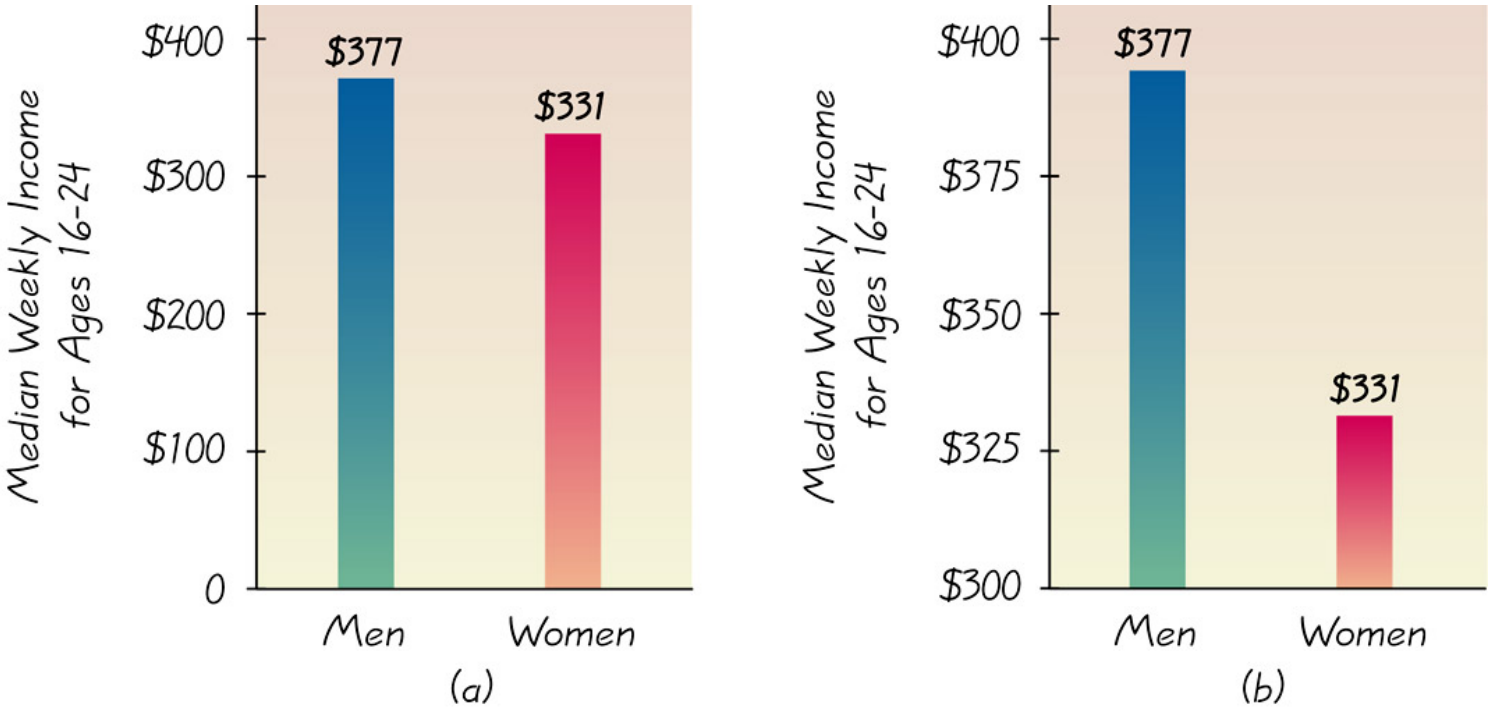
Buktatók

- **Túl kicsi minták**

Megkérdeztünk 1000 véletlenül kiválasztott magyar lakost a pártpreferenciájáról.

A 18-25 éves korosztály pártpreferencia megoszlása ilyen és ilyen volt ….

❖ Félrevezető grafika



1-1 ábra

Ha a kocka oldalhosszúságait megduplázzuk, a térfogata a nyolcszorosára nő!

1-2 ábra

Hogy korrektül interpretáljunk egy diagrammot vagy más grafikus megjelenítést, a benne szereplő számokat kell figyelembe vennünk, és nem szabad engednünk, hogy a kép formája félrevezessen!

Buktatók

- Önkéntes válaszadóktól gyűjtött adatok, önkényesen kiválasztott mérési adatok

ahol a válaszadók döntik el, hogy válaszolnak-e (pl. SMS szavazás),
ahol a lemért adatok közül valami önkényes módon szelektálunk (pl. a többitől nagyon eltérő, outlier adatokat eldobjuk)

Ilyen esetekben nem lehet valós következtetéseket levonni.

A mintának mindig jól kell reprezentálnia sokaságot. A sokaságból csak VÉLETLEN kiválasztás útján szerezhetünk torzítatlan képet.

Arányok viszonyítás nélkül

A mérés során 100g-ról 125g-ra növekedett a tömege:

„A mérés során 25%-kal növekedett a tömeg:”

(125-100)/100 = 25%

„A mérés végén 20%-kal lett nehezebb:”

(125-100)/125 = 20%

Becsapós kérdések

19% igen: Túl keveset költünk szociális kiadásokra.

63% igen: Túl keveset fordítunk a szegények megsegítésére.

A kérdések sorrendje

Ön mit mondana, a közlekedési vagy az ipari légszennyezés magasabb?

Ön mit mondana, az ipari vagy a közlekedési légszennyezés magasabb?

45% -27%
24%- 57%

Nem válaszolók

- **Michael Wheeler: Hazugságok, szemenszedett hazugságok, statisztika**

„Azok, akik nem válaszolnak a telefonos kérdésekre általában különböznek azoktól, akik válaszolnak.”

Precíz számok

- **Magyarország lakosságának száma**
  10 198 315 fő (2001-es népszámlálás)

Korreláció és kauzalitás

- **A korreláció nem jelenti azt, hogy valami valamit okoz is**
- **Pl.: az IQ és a vagyon korrelált, mégsem oka az egyik a másiknak**

Önérdekeltség

- „Egy országos, 250 „emberi erőforrás” szakember között végzett felmérés kimutatta, hogy a kopott cipő a vezető ok abban ha a férfi munkakeresők rossz első benyomást tesznek”
- A felmérést a „Kiwi Brands” támogatta
- A gyógyszercégek fizetnek azoknak a klínikai orvosoknak, akik valamely terméküket használják és erről fontos orvosi lapokban cikket jelentetnek meg.
- Általában nem szabad elhinnünk az olyan statisztikai vizsgálatok eredményét, ahol a támogató anyagilag érdekelt az eredményben.

Részleges ismeretek

- **„Az általunk az utolsó 10 évben az országban eladott autók 90%-a még mindig az utakat járja”**
- **A cég valójában csak három éve adta el az első autót az országban …**

Hiányzó adatok

- **Előfordul, hogy véletlen okokból hiányzik egy-egy adat.**
- **Ha valami speciális okból hiányzik, akkor az használhatatlanná teszi az adatsort.**

- **Pl.: Népszámlálási adatokból hiányoznak az otthontalanok. Jövedelmi adatok esetén az emberek nem mondanak igazat. A laboratóriumi mérések közül kihagyjuk azokat, amik túl nagyok …**

Készakart hamisítások

- http://en.wikipedia.org/wiki/Scientific_misconduct
- **Mendel, Millikan „megerősítési torzió”**

Buktatók

- Hibás minták
- Kicsi minták
- Félrevezető ábrák
- Becsapós ábrák
- Játék a százalékokkal
- Beugrató kérdések
- A kérdések sorrendje
- Válasz megtagadás
- Korreláció és kauzalitás
- Önérdekeltség a vizsgálatban
- Precíz számok
- Részleges képek
- Készakart hamisítás

Összefoglalás

Ebben a fejezetben:

- **Áttekintettünk néhány buktatót.**
- **Bemutattuk miért fontos a józan ész mielőtt statisztikai vizsgálatokat végeznénk**

1-4 fejezet
A kísérletek megtervezése

Fő pontok

- **Ha a mintát nem megfelelő módon gyűjtjük, akkor az annyira használhatatlan lesz, hogy semmiféle statisztikai manipulációval sem tudjuk megmenteni.**

- **A véletlen tipikusan kritikus szerepet játszik abban, hogy mely adatokat gyűjtsük össze.**

Definíció

- **Megfigyeléses vizsgálat (Observational study)**

bizonyos jellemző tulajdonságok megfigyelése és mérése anélkül, hogy megváltoztatnánk a vizsgálat tárgyát/alanyát

pl.: közvéleménykutatás, csillagászati/asztrofizikai megfigyelések

Definíció

- **Kísérlet (Experiment)**

valamilyen kezelést végzünk és azután megfigyeljük a hatásait a kísérlet tárgyán/alanyán

Pl.: klínikai gyógyszervizsgálat, részecske ütközések a CERN gyorsítójában

Definíciók

- **Keresztmetszeti vizsgálat (Cross Sectional Study)**

  Az adatokat egy időpontban mérjük, figyeljük meg és gyűjtjük be.

- **Utólagos vizsgálat (Retrospective Study)**

  Múltbéli adatokat használunk. (pl.: az autóbalesetben meghaltak és nem abban meghaltak összehasonlítása)

- **Előre tervezett (Prospective Study)**

  Az adatokat a jövőben gyűjtjük, olyan csoportokból, melyek valamilyen közös faktorban megegyeznek. (pl.: a mobil telefont használó és nem használó vezetők csoportjainak összehasonlítása)

Definíció

Zavar (bezavarás)

akkor lép fel egy kísérletben, ha a kísérletet végző nem tudja megkülönböztetni az egyes faktorokat

Pl.: Mindenkitől levonunk 1 pontot, ha nem jelenik meg az előadáson, javul-e a részvételi arány? Tfh. hogy javul. De lehet, hogy idén jobb volt az időjárás. A két faktor nem különböztethető meg.

Úgy kell a kísérletet megtervezni, hogy ne lépjen fel zavar!

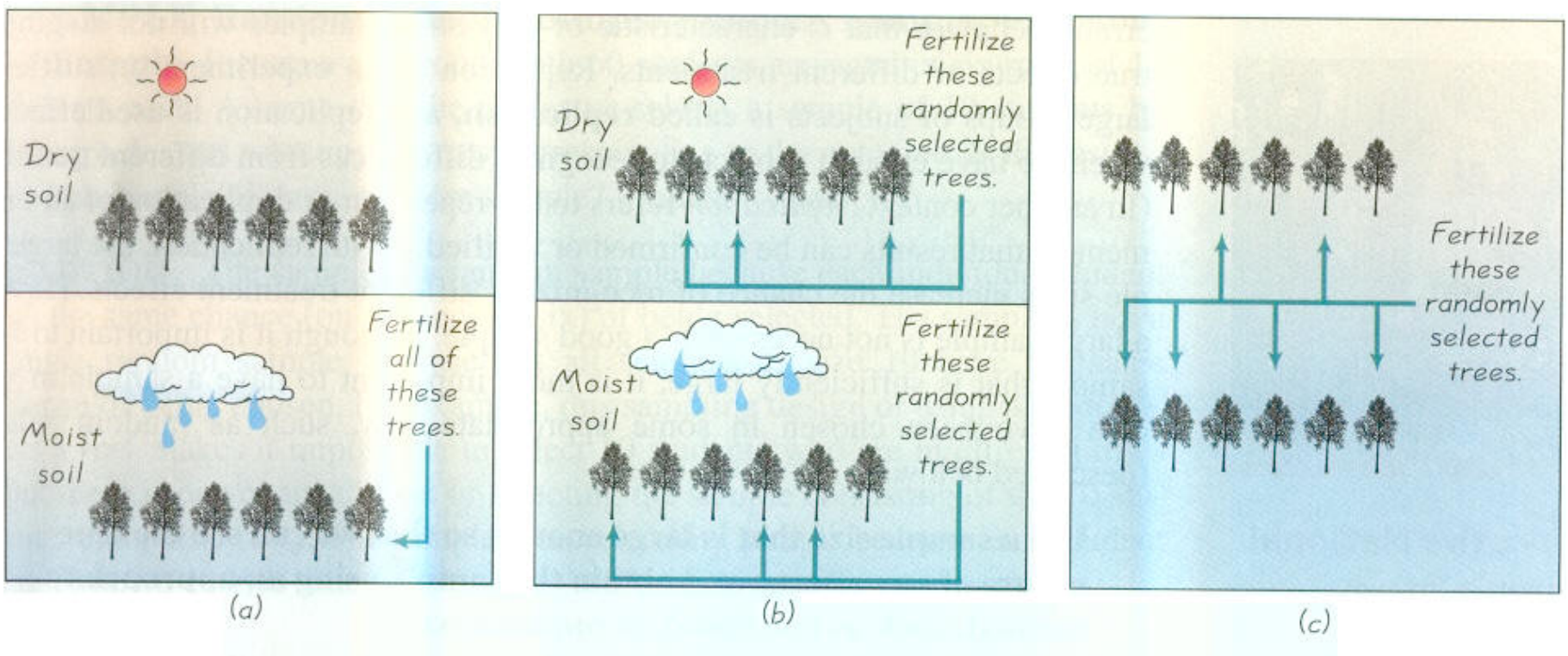
A változók hatásának kontroll alatt tartása

- **Vak vizsgálat (Blinding), duplán vak vizsgálat**

a vizsgálat alanya nem tudja, hogy kezelést kap-e vagy placebót, duplán vak, ha a kísérletező sem tudja (pl.: a gyermekbénulás Salk vakcina kipróbálása az USA-ban 1954-ben)

❖ Blokkosítás – felosztjuk a populációt

olyan alcsoportokra amelyekben a kísérlet szempontjából fontos tulajdonságai megegyeznek . Mindegyik blokkban véletlenszerűen választjuk ki a kezelteket



❖ **Teljesen randomizált (véletlenszerűsített) kísérleti elrendezés**

véletlen kiválasztással választjuk ki azokat, akik kezelést kapnak

pl.:

❖ **Szigorúan kontrollált elrendezés**

nagyon körültekintően kiválasztott egyedek

pl,: ha pl. vérnyomáscsökkentőt tesztelünk, akkor ha az egyik blokkban van egy 30 éves túlsúlyos cigarettázó férfi, aki szereti a sós és zsíros ételeket, akkor a másik blokkba is teszünk ilyet

Ismétlés és a minta mérete

❖ Ismétlés

a kísérlet megismétlése, amikor van elegendő alany ahhoz, hogy észrevehessük a különböző kezelések közti eltéréseket

❖ Minta mérete

akkora mintát kell használni, ami elég nagy ahhoz, hogy kimutathassuk benne az effektust

Definíciók

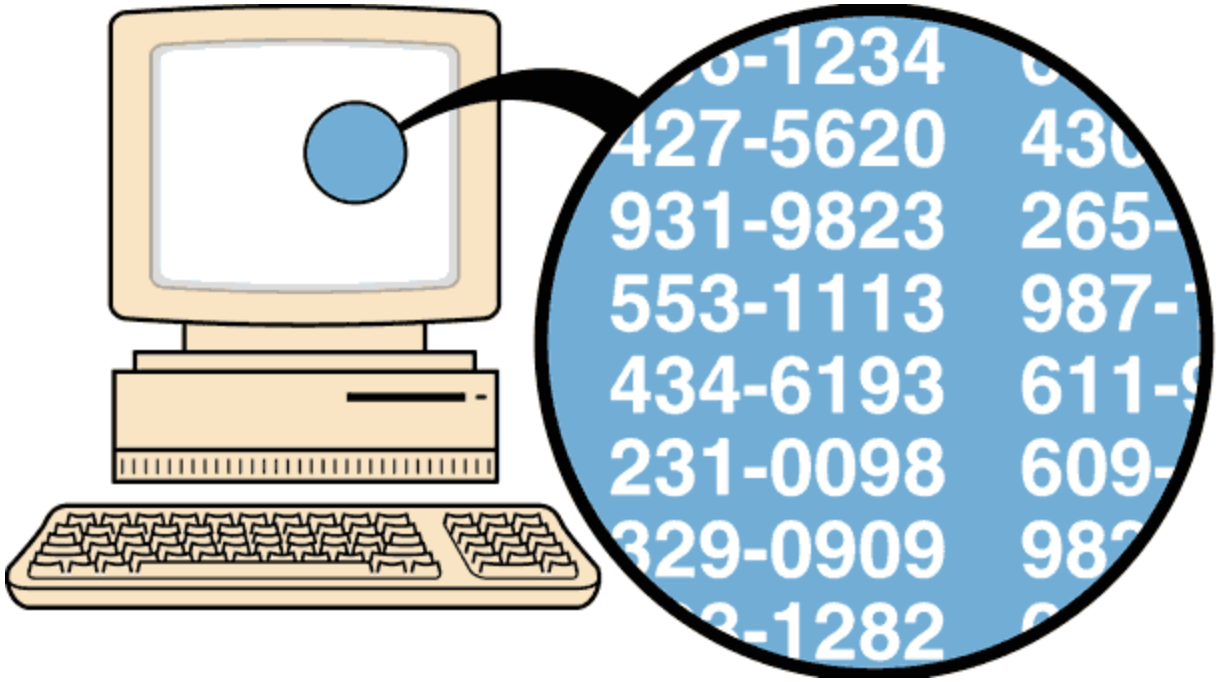
❖ Véletlen mintavétel

a populáció minden tagjának ugyanakkora esélye van arra, hogy a mintába bekerüljön

❖Egyszerű véletlen mintavétel (*n hosszúságú*)

a minta tagjait úgy választjuk ki, hogy bármelyik n hosszúságú mintának ugyanakkora a kiválasztási esélye

Véletlen számok generálása



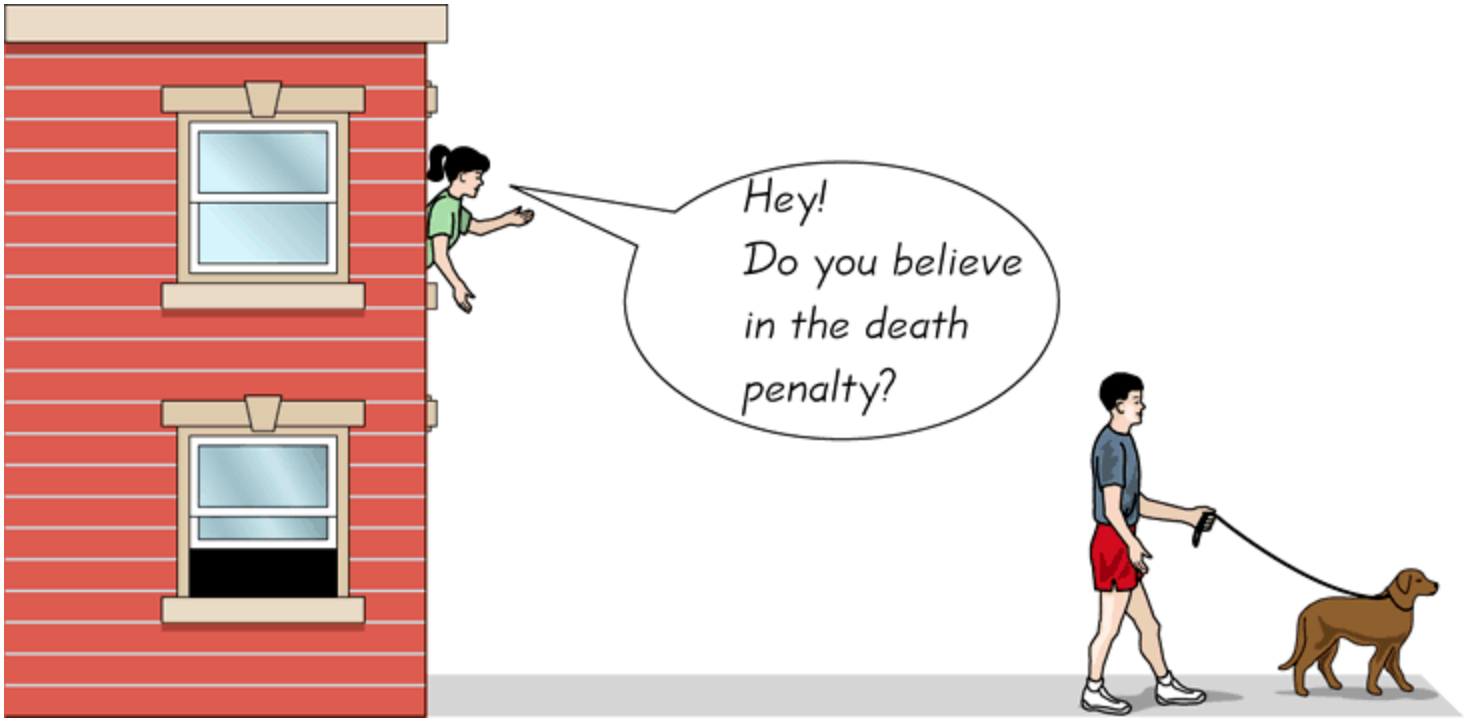
Szisztematikus mintavétel

Valamilyen kezdőponttól indulva kiválasztjuk minden K adik elemet a populációból

problémás lehet, ha a populáció is szisztematikusan van rendezve

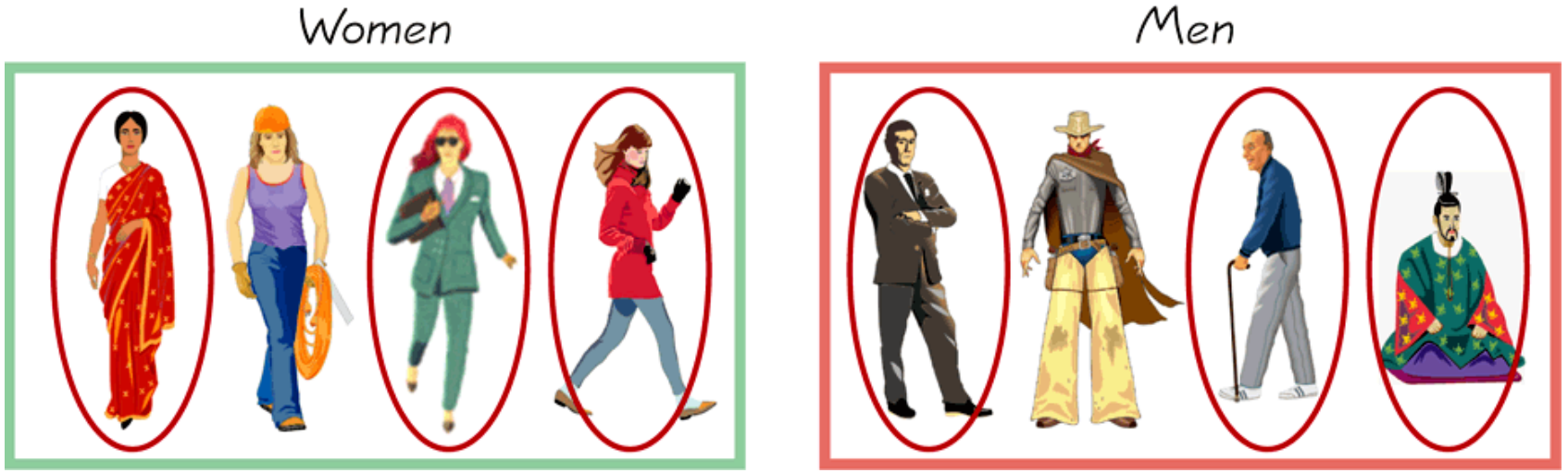
Kényelmes mintavétel

használjuk azt a mintát, amit a legkönnyebb beszerezni



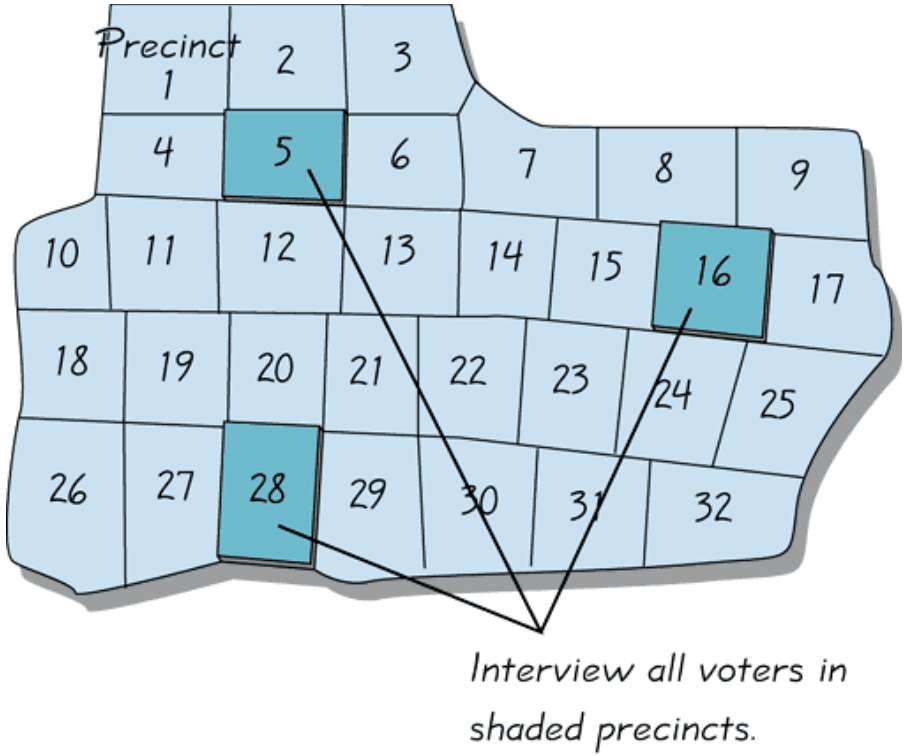
Rétegzett mintavétel

oszd fel a populációt kettő vagy több csoportra (rétegre), melyeken belül bizonyos (a kísérlet szempontjából fontos) tulajdonságok azonosak vagy hasonlóak,
majd vegyünk mintát mindegyik rétegből



Klaszter mintavétel

oszd a populációt valamilyen természetes módon klaszterekre; véletlenül válassz közülük, használd az összes tagot



A mintavételezés módszerei

- Véletlen
- Szisztematikus
- Kényelmi
- Rétegzett
- Klaszter

Definíciók

- Mintavételi hiba (Sampling error)

  a minta és a populáció eredménye közti eltérés, ami a minták fluktuációjából származik

- Nem mintavételi hiba (Non-sampling error)

  olyan eltérés, ami az inkorrekt adatgyűjtésből, adat felvitelből vagy analízisből ered

Összefoglalás

Ebben a fejezetben:

- **A vizsgálatok és mérések típusait**
- **A változók hatásának kontrollálását**
- **Randomizációt**
- **A mintavételezés típusait**
- **A minta hibáit**

tekintettük át.