Elemi statisztika fizikusoknak

Vattay Gábor

Komplex Rendszerek Fizikája Tanszék

vattay@elte.hu

www.complex.elte.hu/vattay.html

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 …

Hogyan szerzünk jegyet ?

- Gyakorlati jegy:
  - Szerezz be egy e-mail címet (gmail nagyon jó) és írj egy levelet az elemistatisztika@gmail.com címre!
  - Szerezz be egy táblázatkezelőt és grafikon készítésre alkalmas programot (Excel, Open Office, google spreadsheet, GNUPLOT …)
  - Előadáson kiadott feladatokat a http://complex.elte.hu/elemistatisztika
  - Megoldásokat e-mailen küldd vissza következő vasárnap éjfélig
  - Beadott házi feladatok pontszáma alapján
    - Táblázatkezelővel megoldható numerikus feladat (20%)
    - Szöveges feladatok (80%)
    - 110 pont érhető el, 100 pontot elég elérni, 50 pont alatt elégtelen
  - Önálló munka, koppintott megoldások pontlevonással járnak
- Előadás jegy: írásbeli vizsga

Hat lépés távolság probléma

- six degrees of separation
- 1967 Stanley Milgram pszihológus
- Hány ismerőssel köthető össze két véletlenül választott amerikai lakos?
- levél 60 kísérleti személynek (Wichita, Kansas)
- továbbítsák ismerősökön keresztül egy nőnek Cambridge-be (Massachusetts)
- 50-en résztvettek a kísérletben
- 3 levél érkezett meg
- később többször megismételték, 35%-os sikerrel
- átlagos közbeeső ismerősök száma 6-nak adódott

- Six Degrees of Kevin Bacon játék
  www.cs.virginia.edu/oracle
- Erdős szám
- Kézfogás probléma (Usama bin Laden)
- A „kicsi világ probléma”
- Megbízhatók voltak-e Milgram adatai?
- Alátámasztják-e Milgram eredeti adatai a „hat lépés távolság” koncepcióját ?

Bevezetés a statisztikába

1-1 Áttekintés

1-2 Az adatok típusai

1-3 Kritikus szemlélet

1-4 Kísérlettervezés

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, használni vagy ön teljesen antialkoholista?” A megkérdezettek válaszaiból (pl. 1000 ember) próbálunk a teljes népességre (pl. 10000000 ember) következtetni.

• Népszámlálás (Cenzus) Megpróbálunk mindenkit megkérdezni.

Definíciók

- Adatok

összegyűjtött megfigyelések (mérések, kérdőíves válaszok, felmérések)

Definíciók

- 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.

Definíciók

- Cenzus

  adatok gyűjteménye a populáció minden eleméről

- Minta

  a populációból kiválasztott elemek rész halmaza

Kulcsfogalmak

- **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.**

1-2.
Az adatok típusai

Definíciók

❖ Paraméter

a populációt jellemző numerikus érték

populáció

paraméter

Definíciók

❖Statisztika

a mintát jellemző numerikus érték.

minta

statisztika

Definíciók

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

Pl.: az emberek súlya

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

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

Kvantitatív adatok

- A kvantitatív adatokat tovább bonthatjuk diszkrét és folytonos típusokra

Definíciók

- Diszkrét

  amikor a lehetséges adatok száma véges vagy legalábbis megszámlálható.

  0, 1, 2, 3, . . .

  Példa: Tyúkok által tojt tojások száma.

Definíciók

❖ 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) .

A mérések szintje

Egy másik lehetőség az adatok jellemzésére, hogy megadjuk a „szintjüket”. Négy példa jön.

Definíciók

- nominális szintű mérések

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

Példa: kérdőíves válasz igen, nem, nem

Definíciók

- ordinális szintű mérés

olyan adatok, melyeket lehet rendezni, de az adatok közti különbségeknek nincs értelmük, vagy nem lehet meghatározni

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

Definíciók

- intervallum szintű mérések

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

Példa: évek 1000, 2001, 1848, és 1526

Definíciók

- arány szintű mérés

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

Összefoglalás - A mérések szintjei

- Nominális – csak kategóriák
- Ordinális – kategóriák és rendezhetőség
- Intervallum – különbségeknek van értelme, de nincs természetes 0 pont
- Arány – a különbségeknek és arányoknak van értelme 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**

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

Készakarva rosszul készített felmérések, laboratóriumi mérések

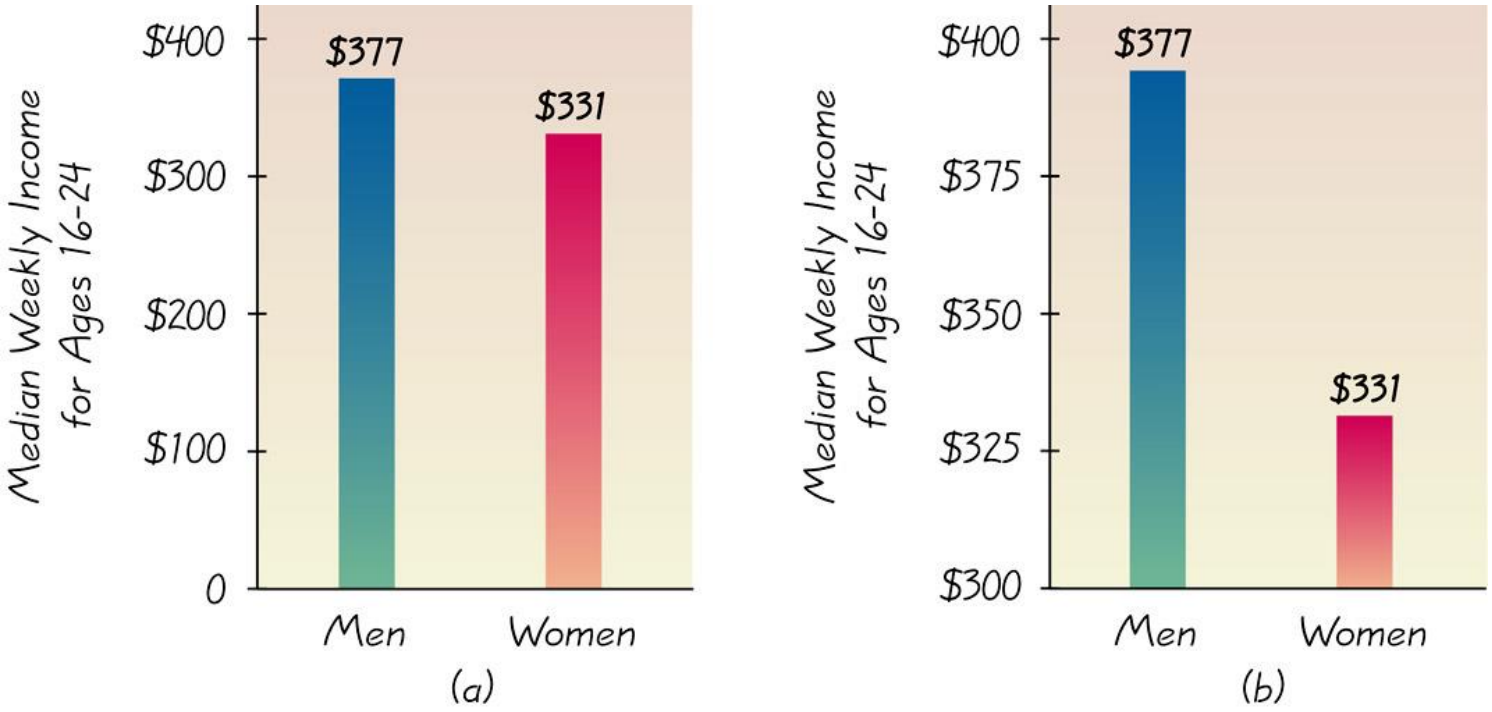
Buktatók

- **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

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

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

1-2 ábra

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.

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!

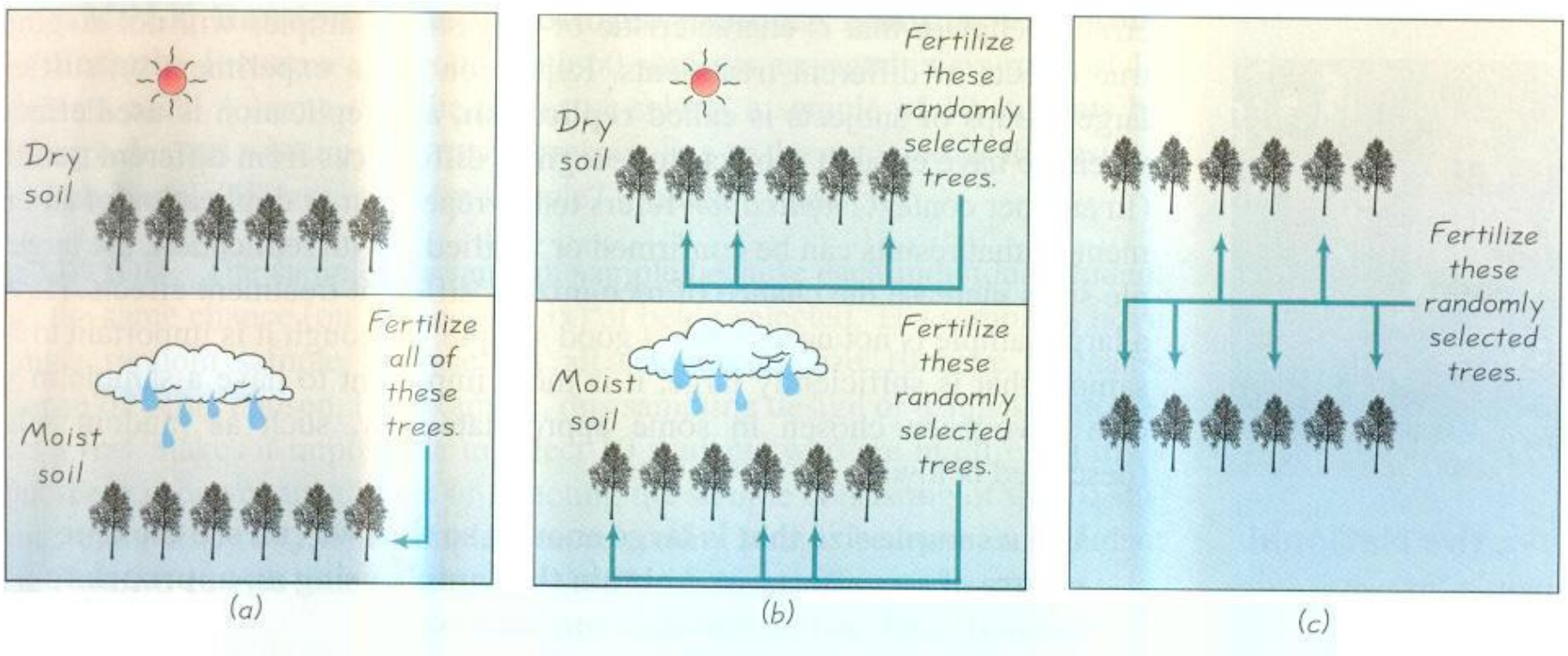
A változók hatásának kontroll alatt tartása

- **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

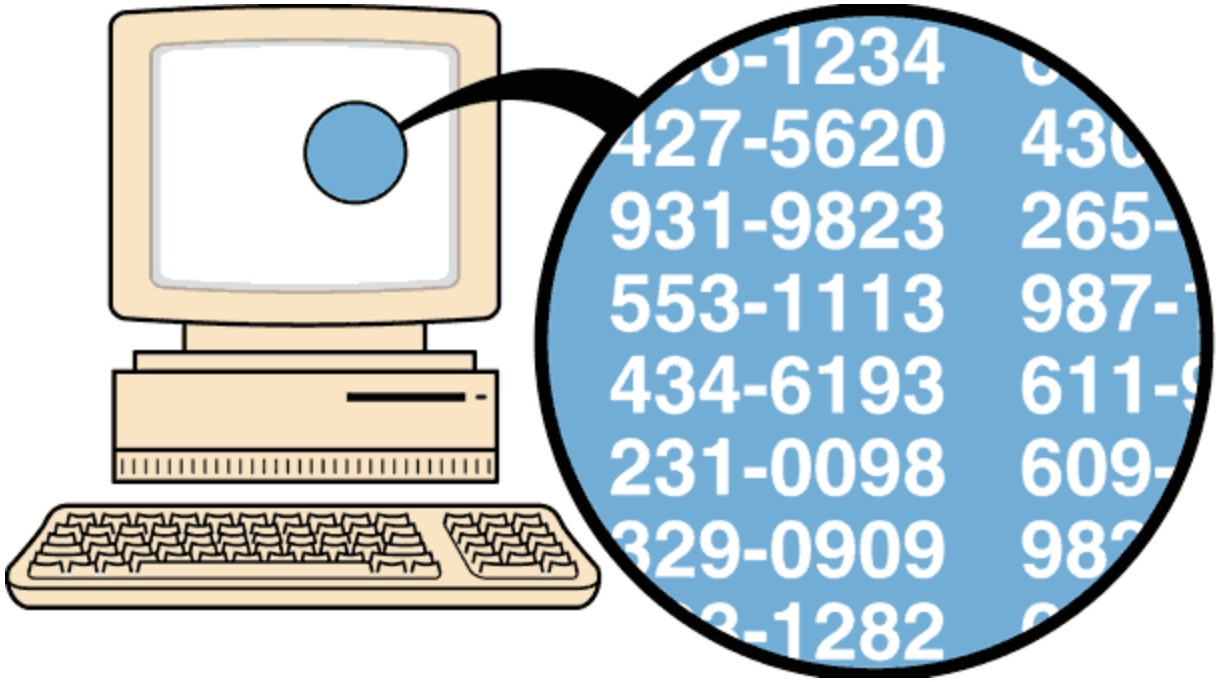
❖ Véletlen mintavétel

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



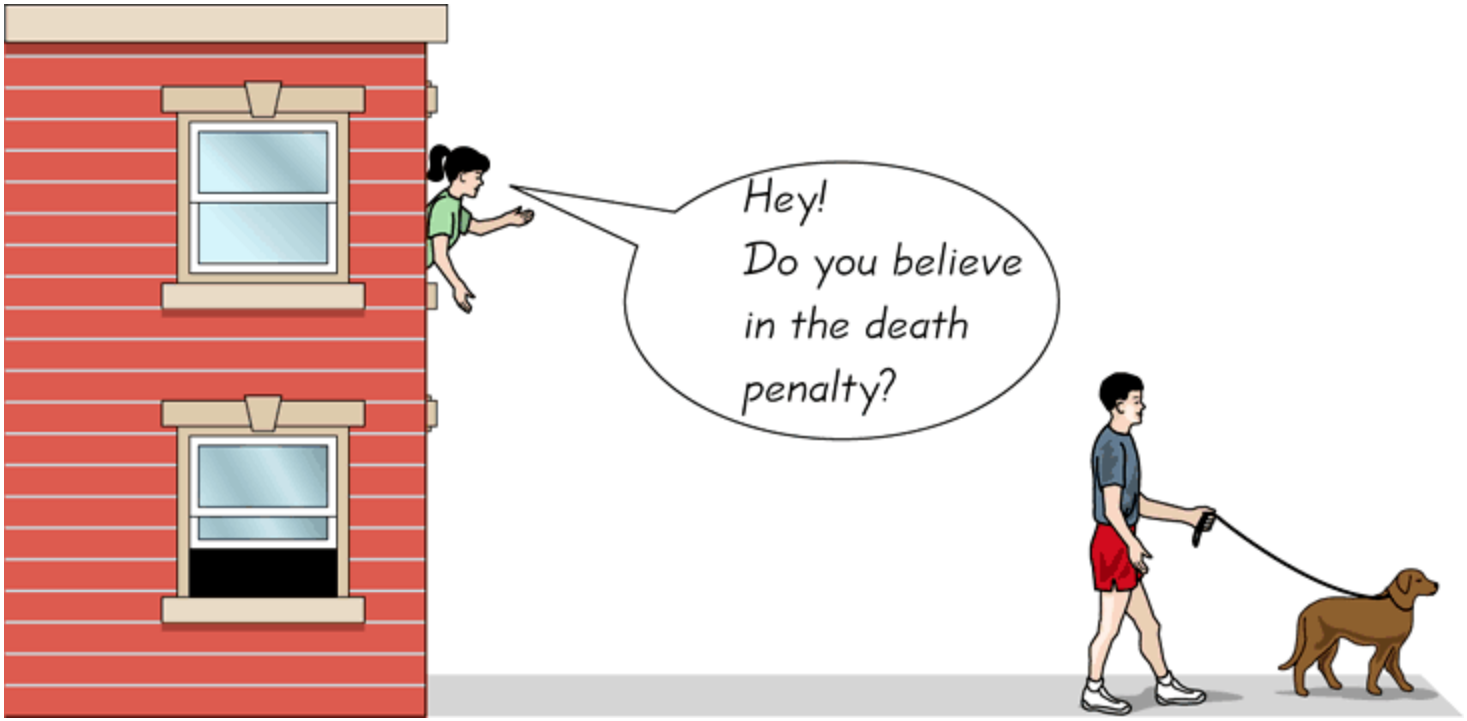
Szisztematikus mintavétel

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

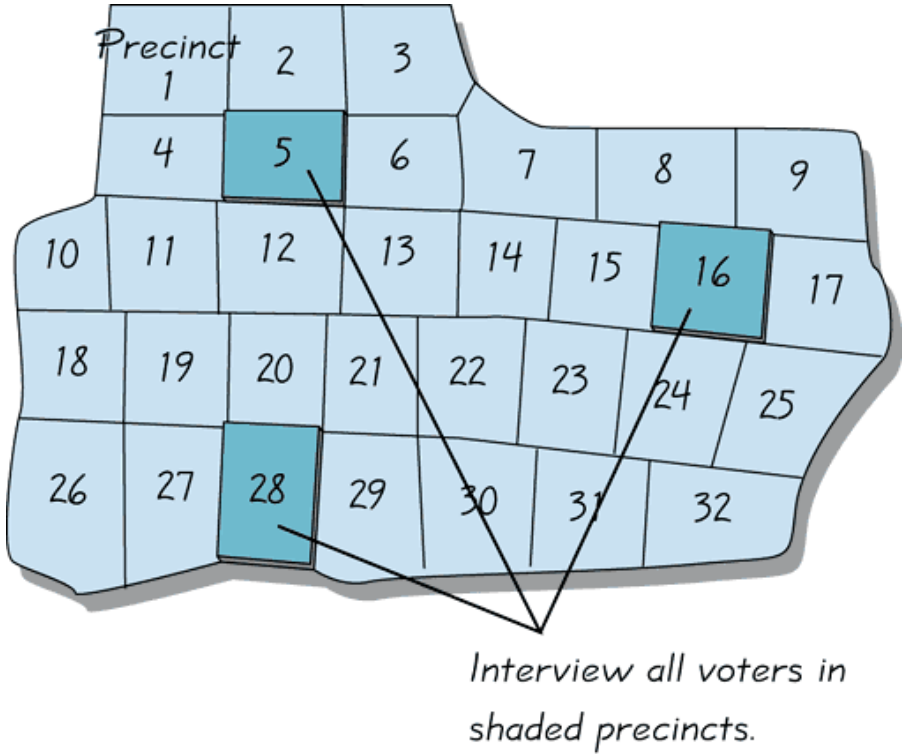


Rétegzett mintavétel

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.

Az adatok leírása, megismerése és összehasonlítása

2-1 Áttekintés

2-2 Gyakoriság eloszlások

2-3 Az adatok vizualizációja

2-4 A centrum mérőszámai

2-5 A szórás mérőszámai

2-6 A relatív elhelyezkedés mérőszámai

2-7 Exploratív adatelemzés

2-5. fejezet
A variabilitás mérőszámai

A variabilitás mérőszámai

A szórás a statisztika egyik legalapvetőbb fogalma, ezért fontos hogy megértsük a lényegét

Várakozási idő különböző bankokban
percekben

Bank of Nyúl	6.5	6.6	6.7	6.8	7.1	7.3	7.4	7.7	7.7	7.7
Csajágröcsögei Bank	4.2	5.4	5.8	6.2	6.7	7.7	7.7	8.5	9.3	10.0

	Bank of Nyúl	Csajágröcsögei Bank
Átlag	7.15	7.15
Medián	7.20	7.20
Módusz	7.7	7.7
Midrange	7.10	7.10

Definíció

Az adat halmaz terjedeleme (range) a legnagyobb és a legkisebb érték közti különbség

legnagyobb érték **–** **legkisebb érték**

Definíció

A minta halmaz szórása (standard eltérése, standard deviation) az adatok eltérését méri az átlag körül

A minta szórásának képlete

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}}$$

2-4. képlet

Példa: 1, 3, 14 (tábla)

A szórás kiszámításának procedúrája

- Számold ki az átlagot $\bar{x}$
- Vond le az átlagot minden egyes adatból $(x-\bar{x})$
- Minden így kapott eltérést emelj a négyzetre $(x-\bar{x})^2$
- Add össze ezeket az eltéréseket $\sum(x-\bar{x})^2$
- Az eredményt oszd el az adatok száma -1 $n-1$ -el.
- Vonjál belőle gyököt

Egyszerűsített képlet

$$s = \sqrt{\frac{n\left(\Sigma x^2\right) - \left(\Sigma x\right)^2}{n\left(n-1\right)}}$$

2-5. képlet

Levezetjük a táblánál!

Szórás - kulcspontok

- A szórás az átlag körüli variabilitás mértéke
- Az s szórás pozitív (vagy 0)
- A szórás s értéke dramatikusan megnő, ha egy vagy több outlier (a többitől messze eső) adat is van köztük
- Az s mértékegysége megegyezik az adatok mértékegységével

A populáció szórása

$$\sigma = \sqrt{\frac{\sum (x - \mu)^2}{N}}$$

Hasonló, mint a 2-4. képlet, azonban itt a populáció átlagát és a populáció nagyságát használjuk (és nem vonunk le 1-et).

Definíció

- A variancia (vagy szórásnégyzet) a szórás négyzete.
- Minta variancia: A minta szórásának négyzete.
- Populáció variancia: A populáció szórásának négyzete.

Variancia - Jelölések

négyzetre emelt szórás

Jelölés $\begin{cases} s^2 & \text{Minta variancia} \\ \sigma^2 & \text{Populáció variancia} \end{cases}$

Miért van n-1 a 2-4. képletben?

Szeretnénk, ha a mintából kiszámított s^2 szórásnégyzet a lehető legjobban megközelítené a populáció σ^2 varianciáját. Nagyon sokféle módon választhatunk ki n db mintaelemet az N elemű populációból, és így sok-sok különböző becslést kapunk a populáció szórására. Számításokkal alátámasztható, hogy a 2-4. képlet az n-1 osztóval átlagosan a helyes becslést adja a szórásra, amit torzítatlan becslésnek nevezünk.

Példa: 3 elemű populáció, véletlen (visszatevéses) mintavételezés

Példa: 3, 6, 9

N=3 $\mu = 6$ $\sigma^2 = ((3-6)^2 + (6-6)^2 + (9-6)^2)/3 = 6.0$

n=2

3,6 és 6,3 $\overline{x} = 4.5$ $s^2 = ((3-4.5)^2 + (6-4.5)^2)/1 = 4.5$

6,9 és 9,6 $\overline{x} = 7.5$ $s^2 = ((6-7.5)^2 + (9-7.5)^2)/1 = 4.5$

3,9 és 9,3 $\overline{x} = 6$ $s^2 = ((3-6)^2 + (9-6)^2)/1 = 18.0$

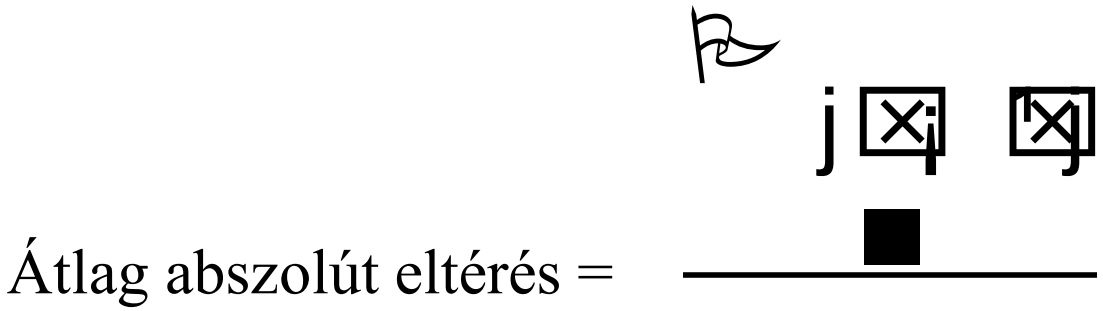
3,3 $\overline{x} = 3$ $s^2 = ((3-3)^2 + (3-3)^2)/1 = 0.0$

6,6 $\overline{x} = 6$ $s^2 = ((6-6)^2 + (6-6)^2)/1 = 0.0$

9,9 $\overline{x} = 9$ $s^2 = ((9-9)^2 + (9-9)^2)/1 = 0.0$

(4.5+4.5+4.5+4.5+18.0+18.0+0.0+0.0+0.0)/9=54.0/9=6.0

Miért nem használjuk az abszolút eltérést?



nem additív és nem torzítatlan becslése a populáció átlagtól való abszolút eltérésének

Definíció

A variációs együttható (CV) megadja a szórást az átlag százalékában kifejezve

Minta

$$CV = \frac{s}{\bar{x}} \bullet 100\%$$

Populáció

$$CV = \frac{\sigma}{\mu} \bullet 100\%$$

Arra jó, hogy különböző skálákon mért variabilitásokat össze tudjunk hasonlítani.

Példa:

- Megvizsgáltuk 100 férfi magasságát és súlyát
- Magasság: $\bar{x}$
- Magasság átlaga = 173.58 cm
- Magasság szórása S= 7.67 cm
- Súly:
- Súly átlaga = 78.26 kg
- Súly szórása S= 11.94 kg
- $CV_{magasság}$=7.67cm/173.58cm=4.42%
- $CV_{súly}$=11.94kg/78.26kg=15.26%

A szórás kiszámítása gyakoriság eloszlásból

2-6. képlet

$$s = \sqrt{\frac{n\,[\Sigma(f \cdot x^2)] - [\Sigma(f \cdot x)]^2}{n\,(n-1)}}$$

Használjuk x értékeknek az osztályfelező pontokat

Definíció

Csebisev tétel

Az adatok legalább 1-1/K^2 –ad része mindig közelebb van az átlaghoz mint K szórás, ahol K egy 1-nél nagyobb pozitív szám.

- **K = 2 esetén, legalább ¾-e (vagy 75%-a) az adatoknak nem tér el jobban az átlagtól mint 2 szórás**
- **K = 3 esetén, legalább 8/9-ada (vagy 89%-a) az adatoknak nem tér el jobban az átlagtól mint 3 szórás**

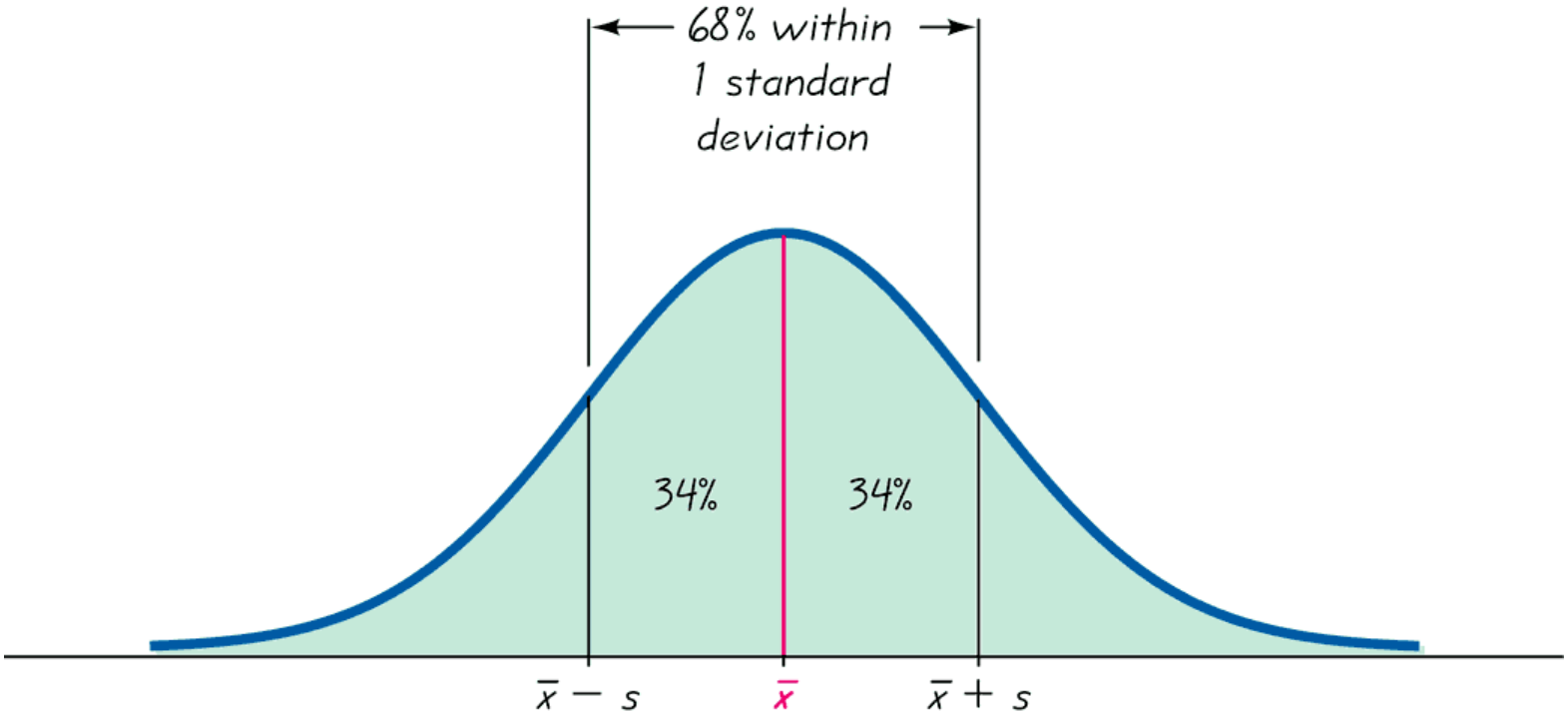
Definíció

Empirikus (68-95-99.7) szabály

Közelítőleg haranggörbe alakú eloszlás esetén a következő tulajdonságok igazak:

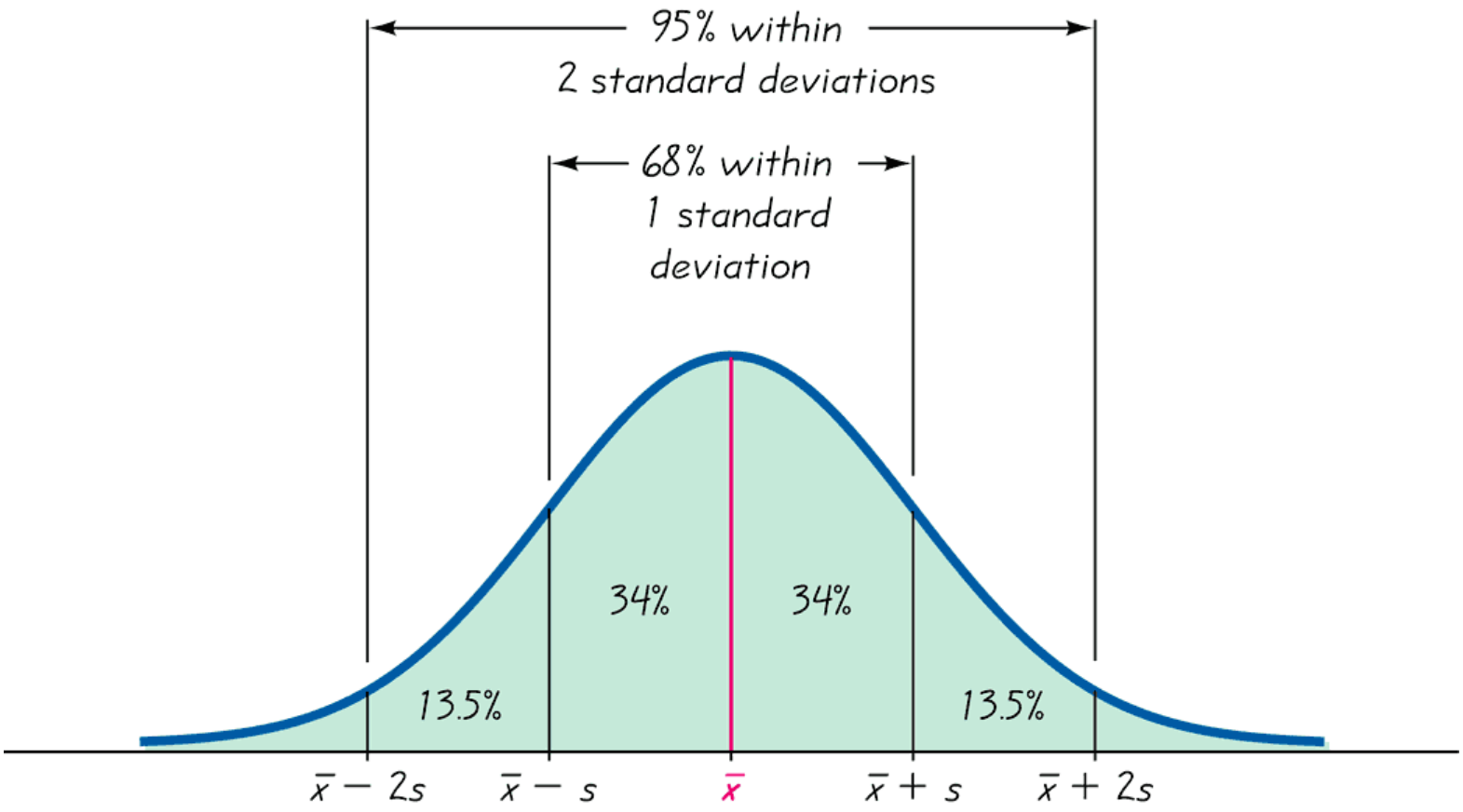
- **Mintegy 68%-a az értékeknek az átlag 1 szórásnyi környezetébe esnek**
- **Mintegy 95%-a az értékeknek az átlag 2 szórásnyi környezetébe esnek**
- **Mintegy 99.7%-a az értékeknek az átlag 3 szórásnyi környezetébe esnek**

Az empirikus szabály



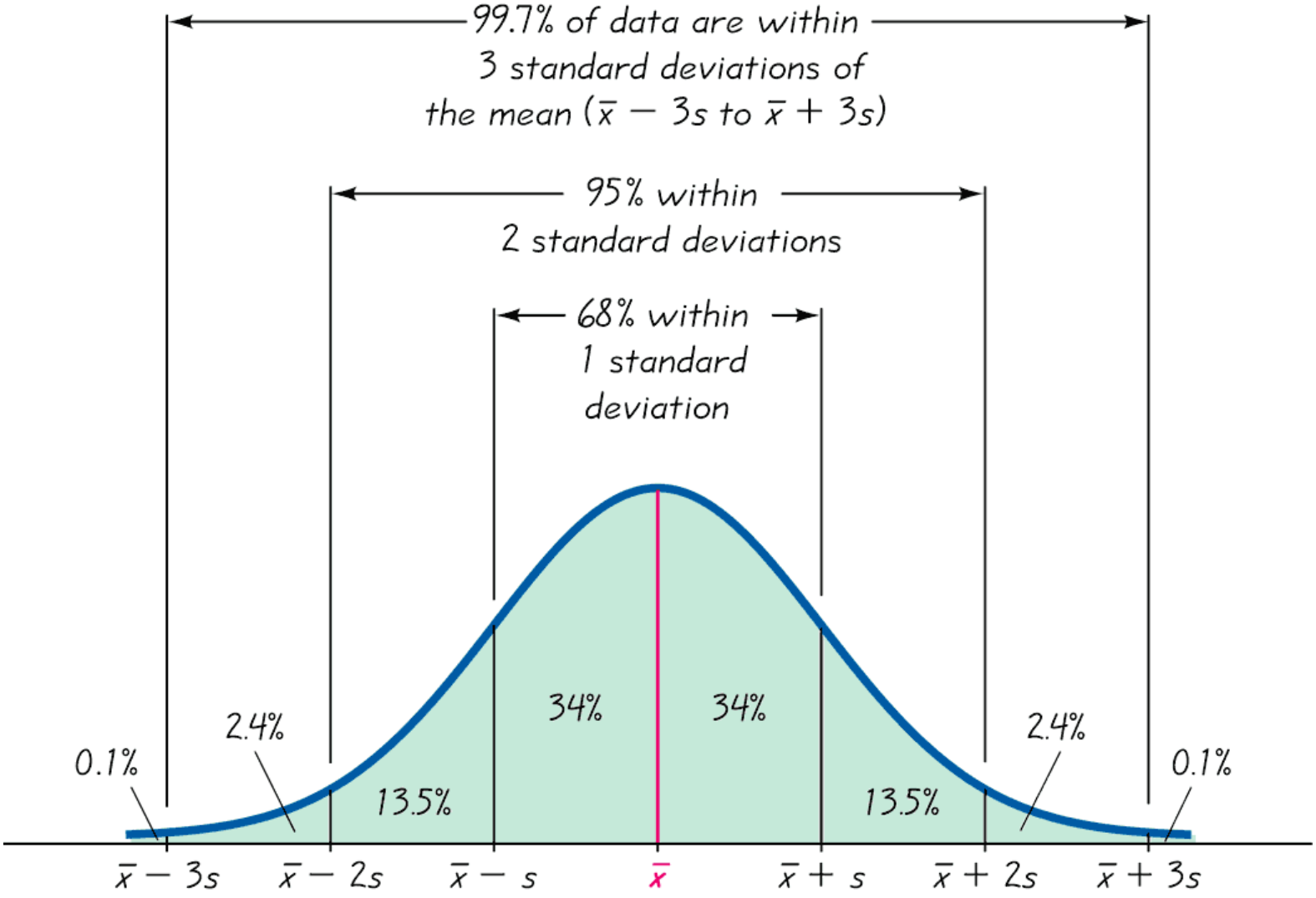
2-13. ábra

Az empirikus szabály



2-13. ábra

Az empirikus szabály



2-13. ábra

Összefoglalás

Ebben a fejezetben foglalkoztunk a:

- **Az adatok terjedelmével**
- **A populáció és a minta szórásával (SD)**
- **A populáció és a minta varianciájával (VAR)**
- **A variációs együtthatóval (CV)**
- **A szórás kiszámításával a gyakoriság eloszlásból**
- **Empirikus szabály**
- **Csebisev tételével**

2-6. fejezet
A relatív helyzet mérőszámai

Definíció

- z eltérés (vagy standard eltérés)

x pozitív vagy negatív eltérése az átlagtól szórás egységekben mérve.

Az eltérés mérése
z érték

Minta

$$z = \frac{x - \bar{x}}{s}$$

Populáció

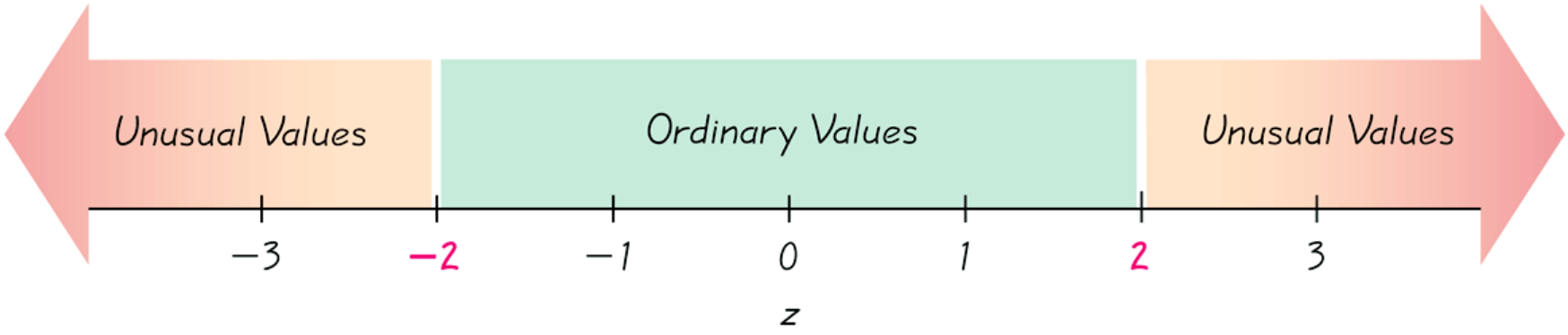
$$z = \frac{x - \mu}{\sigma}$$

Példa:

- Lyndon Johnson volt a legmagasabb amerikai elnök, 190.5 cm.
- Shaquille O'Neal a Miami Heat legmagasabb kosárlabda játékosa, 216 cm.
- Johnson volt-e sokkal magasabb mint az összes elnök, vagy O'Neal a csapattársainál a Miami Heat-ben?
- Elnökök átlaga 181.6 cm, szórása 5.3 cm.
- Miami Heat átlaga 203.2 cm, szórása 8.4 cm.
- $z=(190.5-181.6)/5.3=1.67$

A z eltérés interpretációja

2-14. ábra



Ha egy érték kisebb mint az átlag, akkor a z érték negatív.

Megszokott értékek: *z* értéke –2 és 2 között

Szokatlan értékek: *z* érték < -2 vagy *z* érték > 2

Einstein IQ-ja

- Az IQ eloszlása jó közelítéssel haranggörbe alakú
- Az emberek IQ átlaga 100, szórása 16.
- Einstein IQ-ja 160-volt.
- $z=(160-100)/16=3.75$

Definíció

- Q_1 (Alsó/első kvartilis) nagyság szerint rendezett adatok alsó 25%-át választja el a felső 75%-tól.
- Q_2 (Második kvartilis) ugyanaz mint a median; elválasztja az adatok alsó és felső 50%-át egymástól.

- Q_3 (Felső/harmadik kvartilis) az alsó 75%-ot a felső 25%-tól választja el.

Percentilisek

Ugyanúgy, ahogy a kvartilisek négy részre osztják az adatokat, a 99 percentilis (kvantilis)

P_1, P_2, . . . P_{99}, az adatokat 100 csoportra osztja.

Hogyan találhatjuk meg, hogy egy érték melyik percentilis esik?

$$x \text{ percentilis értéke} = \frac{x\text{-nél kisebb értékek száma}}{\text{az összes értékek száma}} \cdot 100$$

Konverzió a *k*-adik percentilis és a megfelelő adat értékek között

Jelölés

$$L = \frac{k}{100} \cdot n$$

n az adatok száma

k a kvantilis száma

L lokátor, ami meghatározza a keresett adat sorszámát

P_k k-adik kvantilis

Keressük meg 0.8152 kvantilis értékét

2-16 Sorted Weights (in pounds) of Regular Coke in 36 Cans

0.7901	0.8044	0.8062	0.8073	0.8079	0.8110
0.8126	0.8128	0.8143	0.8150	0.8150	0.8152
0.8152	0.8161	0.8161	0.8163	0.8165	0.8170
0.8172	0.8176	0.8181	0.8189	0.8192	0.8192
0.8194	0.8194	0.8207	0.8211	0.8229	0.8244
0.8244	0.8247	0.8251	0.8264	0.8284	0.8295

11/36•100 =30.55556

Kerekítve 31

0.8152 a 31. kvantilisbe esik

2-16 Sorted Weights (in pounds) of Regular Coke in 36 Cans

0.7901	0.8044	0.8062	0.8073	0.8079	0.8110
0.8126	0.8128	0.8143	0.8150	0.8150	0.8152
0.8152	0.8161	0.8161	0.8163	0.8165	0.8170
0.8172	0.8176	0.8181	0.8189	0.8192	0.8192
0.8194	0.8194	0.8207	0.8211	0.8229	0.8244
0.8244	0.8247	0.8251	0.8264	0.8284	0.8295

Keressük meg P_{31} értékét (a 31. kvantilist).

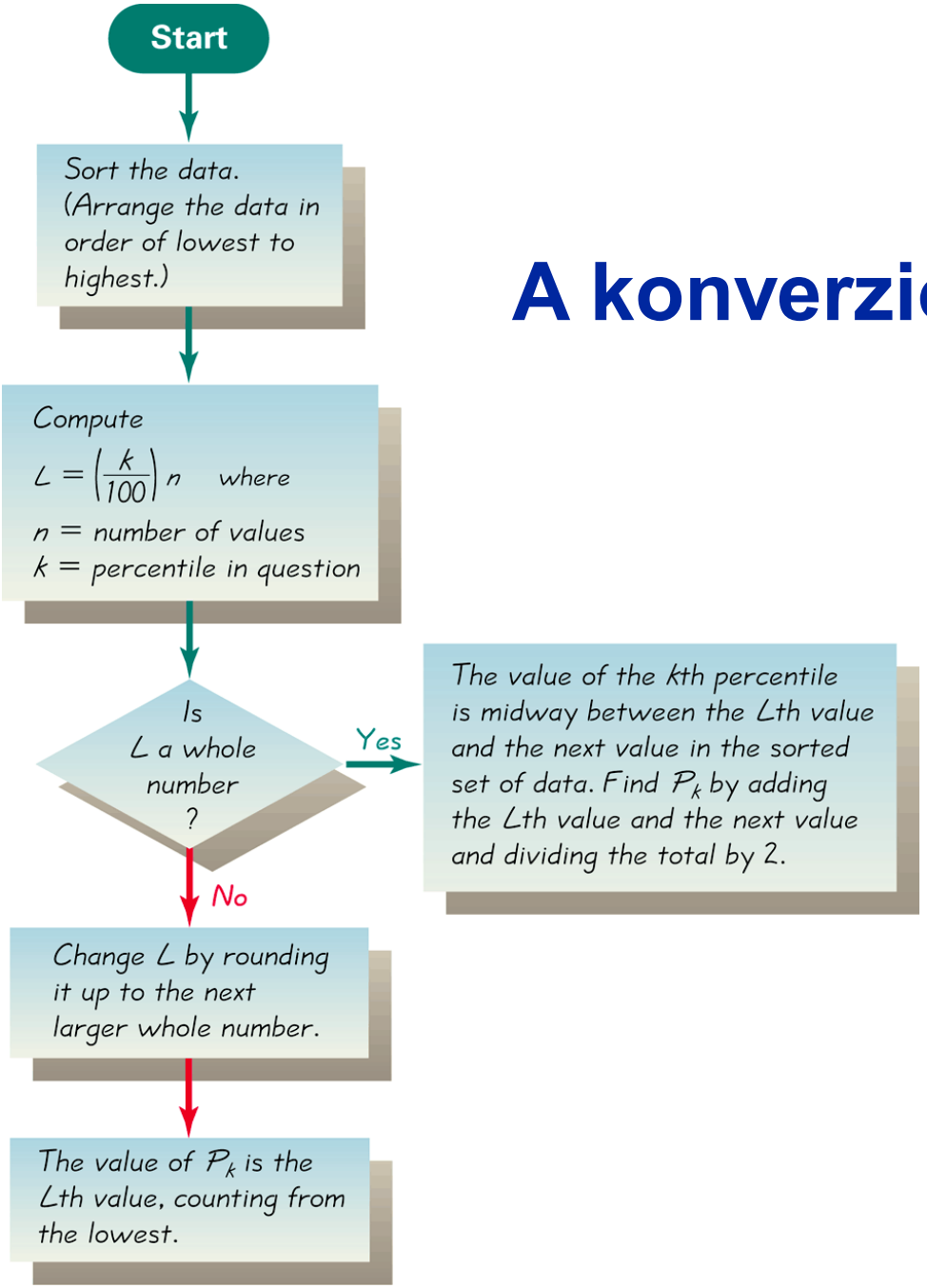
$$L = \frac{31}{100} \bullet 36 = 11.16$$

Kerekítsük fel: 12.

Kezdve a legkisebb értékkel, számoljunk el a 12.-ig a rendezett listában.

$P_{31} = 0.8152.$

A konverzió sémája



2-15. ábra

Néhány fontos jellemző

- **Interkvartilis terjedelem (IQR):** $Q_3 - Q_1$
- **Fél-interkvartilis terjedelem:** $\frac{Q_3 - Q_1}{2}$
- **Kvartilis felező:** $\frac{Q_3 + Q_1}{2}$
- **10 - 90 kvantilis terjedelem:** $P_{90} - P_{10}$

Összefoglalás

Ebben a fejezetben megvitattuk:

- **a *z* értékeket**
- ***z* értékeket és szokatlan értékek**
- **Kvartilisek**
- **kvantilisek**
- **A kvantilisek konvertálása adatértékekre és vissza**
- **Más jellemzők**

2-7. fejezet
Exploratív adatanalízis (EDA)

Definíció

- Exploratív adatanalízis a statisztkai módszerek (mint ábrázolás, a centrum és a variabilitás meghatározása) alkalmazásának a folyamata, amit azért végzünk, hogy megismerjük az adatok legfontosabb statisztikai jellemzőit

Definíció

- Az outlier egy olyan érték, ami nagyon távol esik a többi adat többségétől.

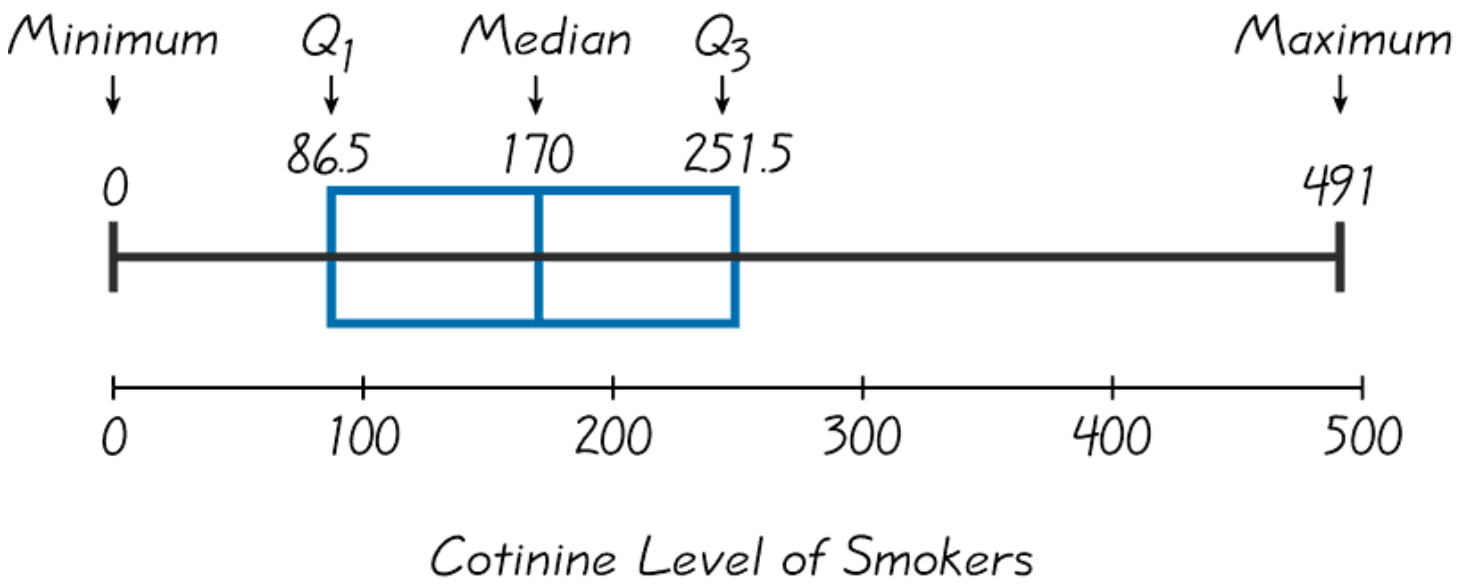
Fontos elvek

- Egy outlier-nek drámai hatása lehet az átlagra
- Egy outlier-nek drámai hatása lehet a szórásra
- Egy outlier-nek drámai hatása lehet a hisztogramra, ami miatt az eloszlás teljesen zavaros lesz

Definíciók

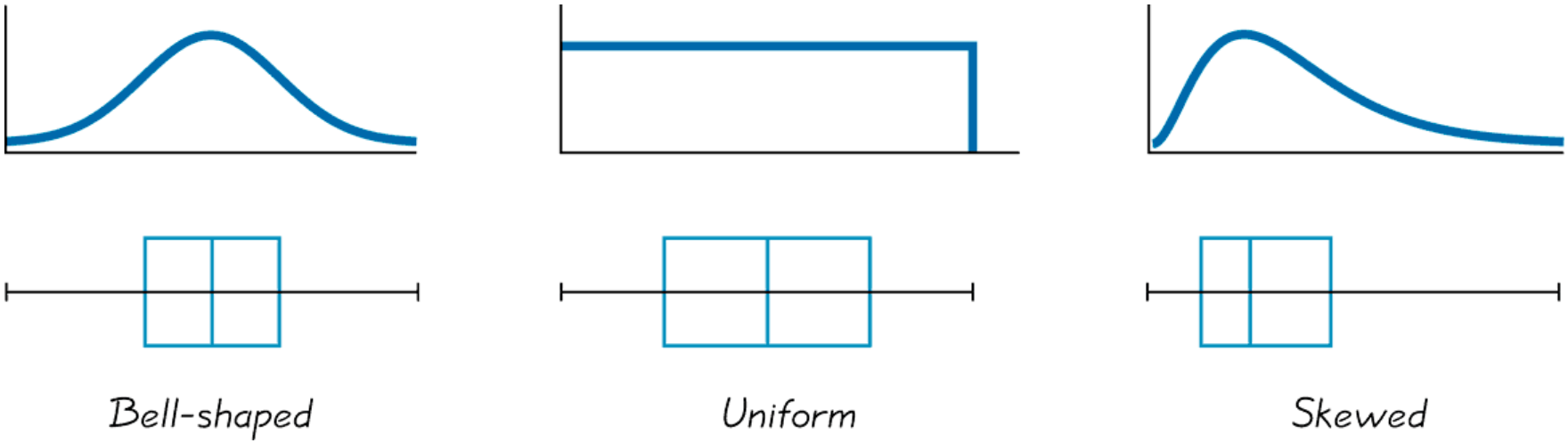
- Egy adathalmazra vonatkozóan, az 5-szám összesítő a minimum értékből; a Q_1 első kvartilisből; a mediánból (Q_2); a harmadik kvartilisből, Q_3; és a maximum értékből áll.
- A boxplot egy a minimumtól a maximumig terjedő vonalból áll, valamint egy dobozból, amiben függőleges vonal húzódik az alsó kvartilisnél, Q_1; a mediánnál; és a felső kvartilisnél, Q_3.

Boxplot



2-16.
ábra

Boxplot-ok



2-17. ábra

Módosított boxplot

- Outlier, ha Q_3 –at 1.5 X IQR-el meghaladja
- Outlier, ha Q_1 –nél 1.5 X IQR-el kisebb

- Ezeket kihagyjuk és csak jelöljük (csillaggal), a maradékra csinálunk boxplotot.

Összefoglalás

Ebben a fejezetben áttekintettük:

- **Exploratív adatanalízist**
- **Az outlier-ek hatását**
- **5-szám összesítőt és a boxplot-ot**

4. előadás
Valószínűség

4-1 Áttekintés

4-2 Alapok

4-3 Addíciós szabály

4-4 Multiplikációs szabály: Alapok

4-5 Multiplikációs szabály: Komplementer és feltételes valószínűség

4-6 A valószínűségek meghatározása szimulációval

4-7 Kombinatorikus szabályok

4-1 fejezet
Áttekintés

Áttekintés

A ritka esemény szabály a következtető statisztikában:

Ha, valamilyen feltevések mellett valamilyen megfigyelt esemény valószínűsége kicsi, akkor arra következtetünk, hogy a feltevés nem igaz.

A statisztikusok a ritka esemény szabályt használják következtetési szabályként (a logikai következtetés helyett).

Példa: egy adott módszer használata mellett 98 lány és 2 fiú születik

4-2. fejezet
Alapok

Kulcsfogalmak

Ebben a fejezetben az események valószínűségének alapfogalmát vezetjük be. Három különböző módszert mutatunk be a valószínűség értékeinek meghatározására.
A legfontosabb célkitűzésünk, hogy megtanuljuk, hogyan kell interpretálni a valószínűség számértékeit.

Definíciók

- **Esemény**

valamilyen folyamat vagy procedúra (továbbiakban véletlen kísérlet) eredményeinek vagy kimeneteinek gyűjteménye

- **Elemi esemény**

egy olyan esemény, amit nem lehet egyszerűbb komponensekre bontani

- **Esemény tér**

a lehetséges elemi események összessége; minden lehetséges kimenet, amit nem lehet tovább bontani

Példák

- esemény/folyamat: egyszerű (nem iker) szülés
- esemény: lány (elemi esemény)
- teljes eseménytér [fiú, lány]

- esemény/folyamat: három szülés
- esemény: 2 lány és egy fiú (nem elemi, mert: llf,lfl,fll)
- teljes eseménytér [fff,ffl,flf,lff,fll,lfl,llf,lll] 8 elemi esemény

Jelölések

P **– jelöli a valószínűséget**

A, B, és C **– adott eseményeket jelöl.**

P (A) **- jelöli annak a valószínűségét, hogy az *A* esemény bekövetkezik.**

A valószínűség kiszámításának szabályai

1. szabály: A valószínűség közelítése a relatív gyakorisággal

Végezz el egy kísérletet (vagy figyelj meg egy folyamatot), és számold meg, hányszor történik meg az *A* esemény. Ezeken a konkrét eseményeken alapulva, *P*(*A*) a következő módon becsülhető:

$$P(A) = \frac{A \text{ bekövetkezéseinek száma}}{\text{hányszor ismétlődött a kísérlet összesen}}$$

Példa: Mi a vsz.-e annak, hogy egy rajzszög a talpára esik?

- Dobjuk le 1000-szer és számoljuk meg hányszor esik talpra.
- Hasonló feladat macskával …

A vsz. kiszámításának szabályai

2. szabály: Klasszikus/kombinatorikus megközelítés (Egyformán valószínű kimeneteket feltételez)

Tegyük fel, hogy egy véletlen kísérletnek *n* különböző elemi esemény a kimenetele és minden egyes kimenet bekövetkezésének ugyanakkora az esélye. Ha egy *A* esemény *s* esetben következhet be az *n* kimenet közül, akkor

$$P(A) = \frac{s}{n} = \frac{A \text{ bekövetkezésének estei}}{\text{az összes elemi események száma}}$$

Példa: Mi a vsz.-e, hogy a dobókockával 6-ost dobunk

- Ideális kocka vagy valódi kocka?
- Elemi események: 1-est, 2-est, 3-ast, 4-est, 5-öst, 6-ost dobunk
- Ha mindegyiknek ugyanakkora a vsz.-e, akkor

  $$P(1)=P(2)=P(3)=P(4)=P(5)=P(6)=1/6$$

- Hasonló problémák: urna golyókkal, lottószámok, kártyajátékok, ….

Szokásos hiba

- Azért, mert nem tudjuk egy esemény vsz.-ét, még nem jelenti azt, hogy 50% - 50% hogy az megtörténik vagy sem:
- Átmegyek-e az elemi statisztika vizsgán?
- Milyen idő lesz holnap?
- Szeret? Nem szeret?

A vsz. kiszámításának szabályai

3. szabály: Szubjektív valószínűség

***P*(*A*), az *A* esemény valószínűségét a releváns körülmények figyelembevételével becsüljük.**

A nagy számok törvénye

Ha a véletlen kísérletet újra és újra megismételjük, a relatív gyakoriságból kapott (1. szabály) valószínűség az esemény valódi valószínűségét közelíti meg.

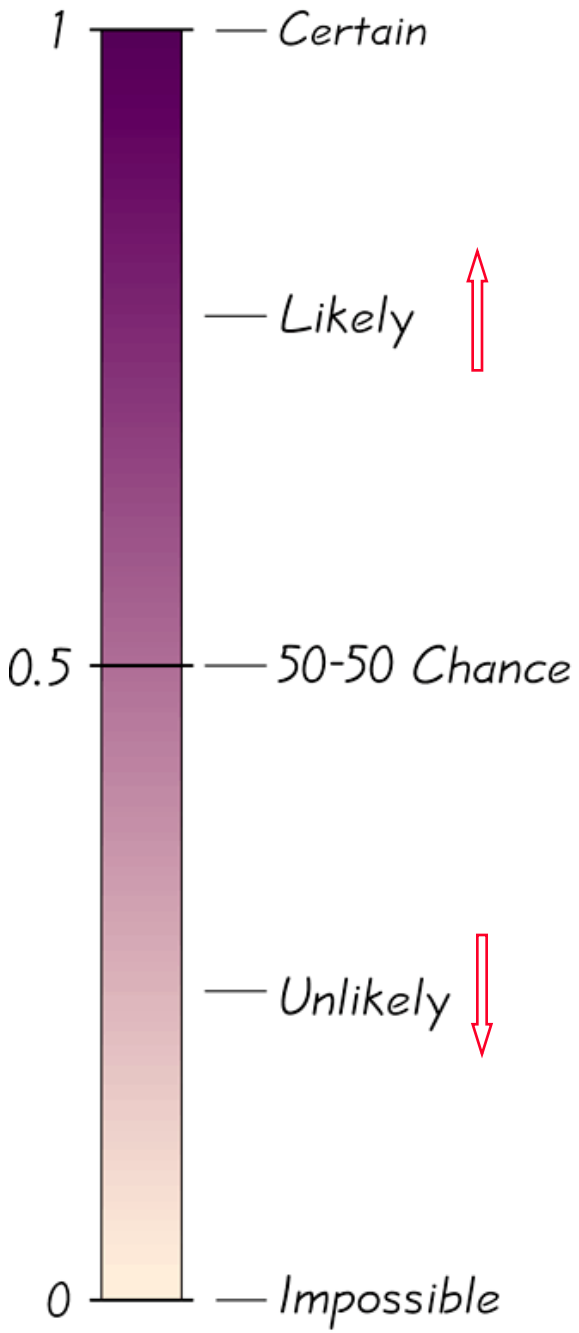
A valószínűség határai

- **A lehetetlen esemény valószínűsége 0.**
- **A bizonyosan bekövetkező esemény valószínűsége 1.**
- **Minden *A* eseményre, *A* vsz.-ge 0 és 1 közé esik, beleértve a határokat is.**
  Vagyis, $0 \leq P(A) \leq 1$.

Magyarázat

- Összes kísérlet száma: N
- Amiben A bekövetkezett: N_A
- A vsz. becslése $N_A/N \rightarrow P(A)$
- $0 \leq N_A/N \leq 1 \rightarrow 0 \leq P(A) \leq 1$

A valószínűség lehetséges értéke



Definíció

Az *A* esemény komplementerét $\overline{A}$ jelöli, ami mindazokból az eseményekből áll, melyekben *A* nem következik be.

Példa

- A valóságban több fiú születik, mint lány. 205 újszülött közül 105 fiú. Mi a valószínűsége annak, hogy egy véletlenül kiválasztott újszülött nem fiú.

  P(nem fiú)=P(lány)=100/205=0.488

Definíciók

Az igazi esélyek az *A* esemény megtörténése ellenében $P(A)/P(\overline{A})$, általában a:b alakban kifejezve (vagy "a a b-hez"), ahol *a* és *b* egész számok (közös osztó nélkül).

Az igazi esélyek az A esemény megtörténése mellett az előbbi reciproka. Ha *A* ellenében a:b az esély, akkor *A* mellett b:a.

A nyerési esély az *A* eseménnyel szemben a nettó profit (ha nyersz) viszonya a feltett összeghez.

nyerési esély egy A eseménnyel szemben

A = (nettó profit) : (feltett összeg)

Példa

- A kaszinóban tegyünk a 13-as számra 5$-t.
- A nyerés vsz.-e: 1/38
- A kaszinó 35:1-arányban fogad.
- Mekkora az igazi esély?
- a 13-assal szemben az esély=P(nem 13)/P(13)=37/38 / 1/38 = 37 vagyis 37:1

Összefoglalás

Ebben a fejezetben megvitattuk:

- **A ritka események szabályát**
- **A valószínűség szabályait.**
- **A nagy számok törvényeit.**
- **A komplementer eseményt.**
- **Esélyeket.**

4-3. fejezet
Addíciós szabály

Kulcsfogalmak

A fejezet célja, hogy bemutassuk az addíciós szabályt ami egy jó eszköz arra, hogy vele olyan vsz.-eket számítsunk ki melyek *P*(*A* vagy *B*) alakúak, azaz annak a vsz.-e hogy vagy *A* esemény bekövetkezik, vagy *B* esemény bekövetkezi (esetleg mindkettő) a véletlen kísérlet kimeneteként.

Halálos áldozatok gyalogos gázolásnál

Ittasság	Gyalogos igen	Gyalogos nem
Vezető igen	59	79
Vezető nem	266	581

Példa

- Mi a vsz.-e annak, hogy vagy a vezető vagy a gyalogos ittas volt?
- Összes eset 985
- Ittas volt valaki: 404/985 = 41%

Definíció

Összetett esemény

bármely 2 vagy többe elemi eseményből összetett esemény

Jelölés

***P*(*A* vagy *B*) = *P*(*A* + *B*) = *P* (egy kísérletben, *A* esemény vagy *B* esemény vagy mindkettő bekövetkezik)**

Az összetett események vsz.-ének általános szabálya

Ha ki akarjuk számítani annak a vsz.-ét, hogy *A* bekövetkezik vagy *B* bekövetkezik, meg kell számolni, hogy *A* hányszor következik be és hogy *B* hányszor következik be, de nem szabad több mint egyszer megszámolni a lehetséges kimeneteket.

Példa

- Mekkora annak a valószínűsége, hogy a vezető vagy a gyalogos ittas volt?
- Vezető ittas: 138
- Gyalogos ittas: 325
- Összesen 463
- de, kétszer számoltuk azt az 59 esetet, amikor mindketten ittasak voltak 463-59=404

Összetett esemény

Formális összeadási szabály:

$$P(A \text{ vagy } B) = P(A+B) = P(A) + P(B) - P(A \text{ és } B)$$

ahol $P(A$ és $B)$ jelenti annak a vsz.-ét, hogy A és B mindketten egyszerre bekövetkeznek a kísérlet kimeneteként.

Intuitív szabály:

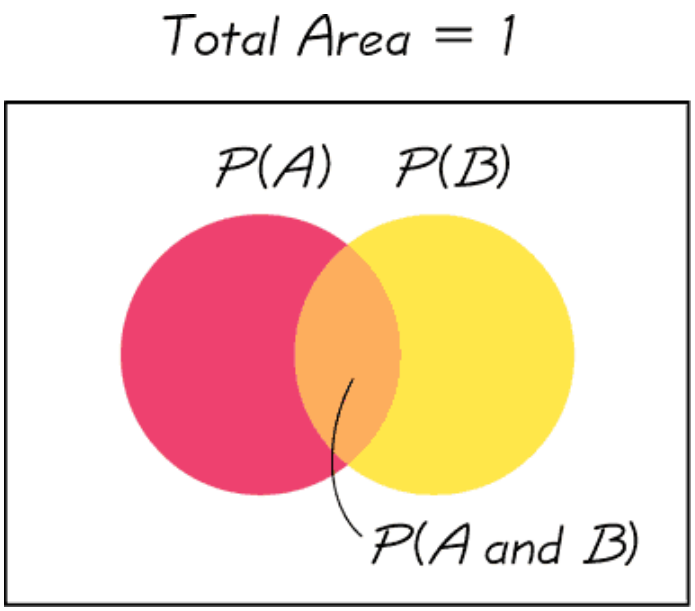
$$N_{A+B} = N_A + N_B - N_{A \text{ és } B}$$

$$N_{A+B}/N = N_A/N + N_B/N - N_{A \text{ és } B}/N$$

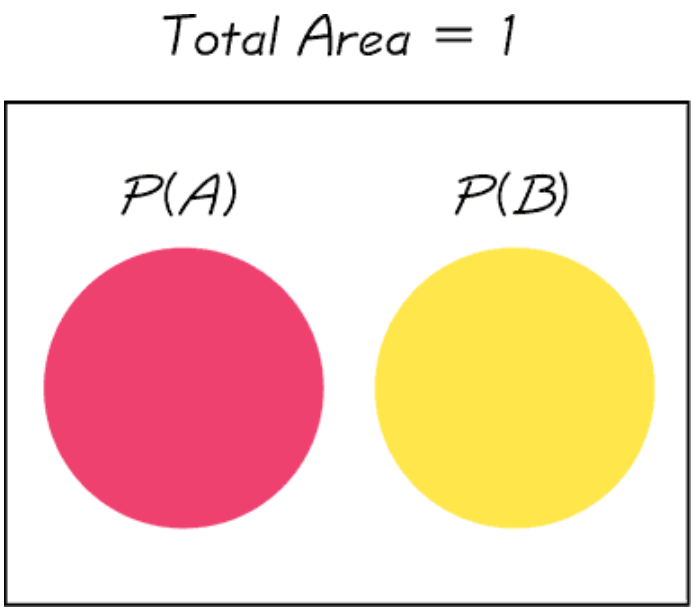
$$P(A+B) = P(A) + P(B) - P(A \text{ és } B)$$

Definíció

Az *A* és *B* események diszjunktak (vagy kölcsönösen kizárók) ha nem történhetnek meg egyszerre. (Vagyis, diszjunkt események nem fedhetnek át egymással.)



Nem diszjunkt események Venn diagrammja



Diszjunkt események Venn diagrammja

Komplementer események

A és $\bar{A}$
diszjunkt események

Egy esemény és a komplementere nem következhetnek be egyszerre.

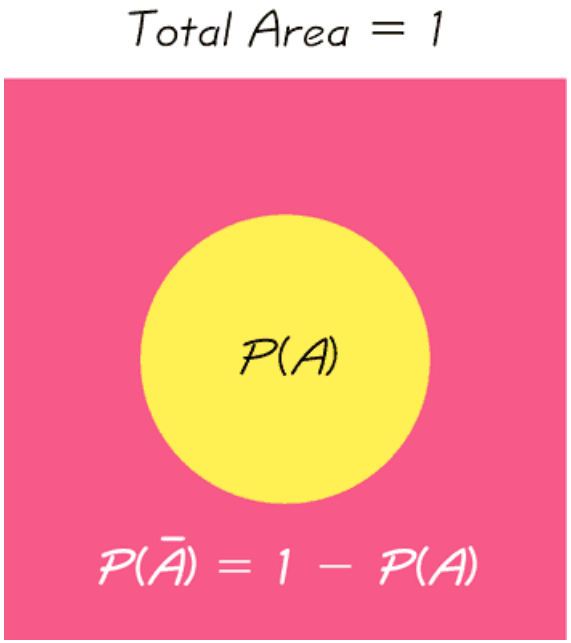
Komplementer események szabályai

$$P(A) + P(\bar{A}) = 1$$

$$P(\bar{A}) = 1 - P(A)$$

$$P(A) = 1 - P(\bar{A})$$

A és komplementerének Venn diagrammja



Összefoglalás

Ebben a fejezetben tárgyaltuk az:

- **Összetett eseményeket.**
- **A formális összeadási szabályt.**
- **Az intuitív összeadási szabályt.**
- **Diszjunkt eseményeket.**
- **Komplementer eseményeket.**

4-4. fejezet
Multiplikációs szabály

Kulcsfogalmak

Ha az első *A* esemény kimenete valahogy befolyásolja a második *B* esemény kimenetét, fontos hogy a *B* esemény vsz.-ének kiszámításakor figyelembe tudjuk venni hogy *A* bekövetkezett.

***P*(*A* és *B*) = P(A*B) kiszámításának szabályát multiplikációs szabálynak nevezzük.**

Feltételes valószínűség

A második esemény *B* vsz.-ében figyelembe kell vennünk, hogy *A* bekövetkezett.

A feltételes vsz. jelölése

$P(B|A)$ jelöli annak a vsz.-ét, hogy a B esemény bekövetkezik, feltéve hogy A esemény már bekövetkezett ($B|A$ mint "B feltéve, hogy A.")

Példa

- Mi a valószínűsége annak, hogy a vezető ittas volt (A esemény)?
- P(A)=138/985 = 14%
- Mi a valószínűsége annak, hogy a gyalogos ittas volt (B esemény)?
- P(B)=325/985=33%
- Mi a valószínűsége annak, hogy a gyalogos ittas volt, ha tudjuk, hogy a vezető ittas volt?
- Vezető ittas 138 esetben, ebből 59 esetben a gyalogos is.
- P(B|A)=59/138=43%

Definíció

Egy esemény feltételes valószínűsége az a valószínűség, amit akkor kapunk, ha figyelembe vesszük, hogy egy másik esemény már megtörtént. $P(B \mid A)$ jelöli B esemény feltételes vsz.-ét, feltéve, hogy A bekövetkezett. Kiszámítása:

$$P(B \mid A) = \frac{P(A \text{ és } B)}{P(A)}$$

Példa (tovább)

- P(A és B)= 59/985
- P(A)=138/985
- P(B|A)=P(A és B)/P(A)=59/138 mint előbb.

Definíció

Független események

Két esemény, *A* és *B* függetlenek ha az egyik bekövetkezése nem befolyásolja a másik bekövetkezésének valószínűségét. (Több esemény hasonló módon független, ha bármelyikük bekövetkezése nem befolyásolja a többiek bekövetkezésének valószínűségét.) Ha *A* és *B* nem függetlenek, akkor egymástól függőnek nevezzük őket.

Formális szorzási szabály

- $P(A \text{ és } B) = P(A) \bullet P(B \mid A)$
- **Ha A és B független események, akkor** $P(B \mid A) = P(B)$.

Intuitív szorzási szabály

$N_{A \text{ és } B} = (N_{A*B}/N_A)N_A$

$N_{A \text{ és } B}/N = (N_{A*B}/N_A)N_A/N$

$P(A*B)=P(B|A)P(A)$

$P(A*B)=P(B)*P(A)$, ha A és B függetlenek

Összefoglalás:

- Feltételes valószínűséget.
- Formális szorzási szabályt.
- Intuitív szorzási szabályt.

4-6. fejezet
Valószínűségek kiszámítása szimulációval

Kulcsfogalmak

Ebben a fejezetben egy másik módszert mutatunk be a valószínűségek kiszámítására, amivel az előző fejezetekben bevezetett formális módszerek nehézségeit ki lehet kerülni.

Definíció

Egy folyamat szimulációja egy olyan másik folyamat, ami ugyanúgy viselkedik, és így hasonló eredményeket produkál mint az első.

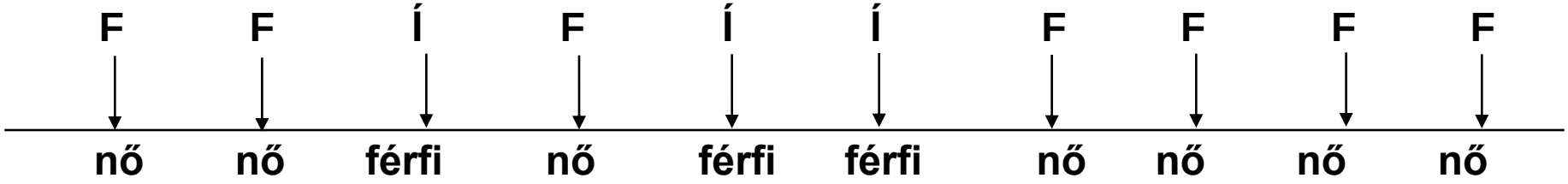
Szimulációs példa (nagyon egyszerű példa)

Nemek (F,N) szelekciója **Ha valamilyen nemi szelekciós módszert tesztelünk, akkor tudnunk kell, hogy mi a valószínűsége annak, hogy 100 újszülött közül legalább 60 lány. Feltéve, hogy a fiú és lány születések egyforma gyakoriak (vagy nem). Találjunk ki egy egyszerű szimulációt, amivel ki tudjuk számítani ezt a valószínűséget.**

Példa: Generáljuk 100 újszülött nemét

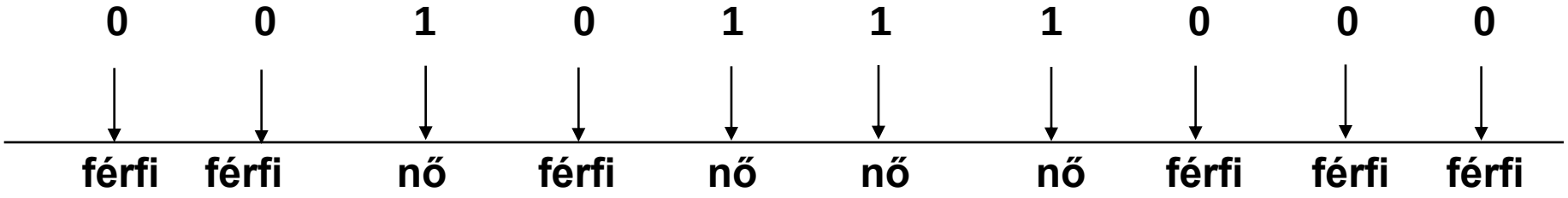
1. megoldás:

- Dobjunk fel 100-szor egy érmét és fej = nő
  írás = férfi



2. megoldás:

- Generáljunk 0’ és 1’ sorozatokat egy számítógéppel, ahol
  0 = férfi
  1 = nő



- Generáljunk nagyon sokszor (N alkalommal) 100 db véletlen 0 vagy 1 számot 50-50% valószínűséggel (vagy pl. 51,12% - 48,88 %).
- Számoljuk meg hányban van 60 vagy több 1-es (N_{60} alkalommal).
- P(60 vagy több lány 100 születésből)=N_{60}/N

Véletlen számok

Sok szimulációban, véletlen számokat használunk a valóságos események szimulációjára. Különböző véletlen szám generálási módszerek:

- **Véletlen számok táblázata**
- **Excel (VÉL() függvény, 0 és 1 között egyenltesen)**
- **C (y=random(100)) véletlen 0 és 100 közötti egész egyenletesen**

Összefoglalás

Ebben a fejezetben megvitattuk a:

- **Szimulációkat.**
- **Véletlen szám generálást.**

5. előadás
Valószínűség eloszlások

5-1 Áttekintés

5-2 Véletlen változók

5-3 A binomiális eloszlás

5-4 A binomiális eloszlás átlaga, varianciája és szórása

5-5 A Poisson eloszlás

6-1 Áttekintés

6-2 A normális eloszlás

5-1. fejezet
Áttekintés

Áttekintés

Ezen az előadáson

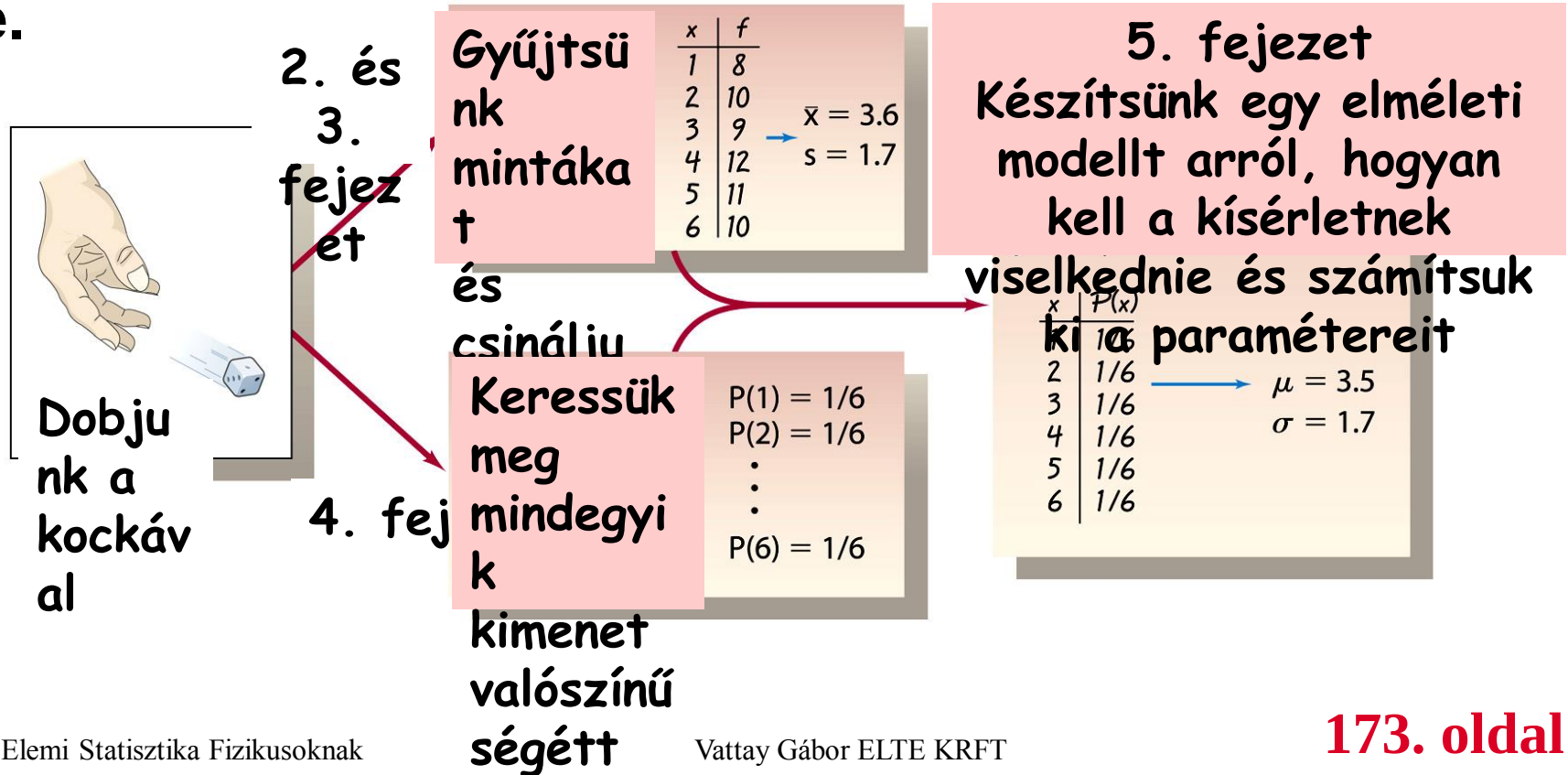
diszkrét valószínűség eloszlások

tulajdonságaival foglalkozunk a 2.-3. előadáson bemutatott leíró statisztika és a 4. előadáson bemutatott valószínűség tárgyalása során használt módszerek kombinálásával.

A valószínűség eloszlások azt írják le, hogy valószínűleg mi fog történni és nem azt, hogy valójában mi történt.

A leíró módszerek és a valószínűség kombinálása

Ebben a fejezetben valószínűség eloszlásokat konstruálunk, amik a lehetséges kimeneteket és a hozzájuk tartozó várható relatív gyakoriságukat mutatják be.



5-2. fejezet
Véletlen változók

Kulcsfogalmak

Ebben a fejezetben bevezetjük a valószínűségi eloszlás fogalmát, ami megadja egy változó véletlen által meghatározott értékeinek a valószínűségét.

Figyelembe veszi, hogy egy adott kimenet gyakran következik-e be, vagy pedig egy szokatlan értékkel van dolgunk, ami ritkán fordul elő véletlenül.

Definíciók

- **Véletlen változó**

  egy változó (tipikusan *x* jelöli) aminek az egyes számértékeit a véletlen kísérlet véletlenszerű kimenetei határoznak meg

- **Valószínűség eloszlás**

  egy olyan leírás, ami a véletlen változó minden egyes értékéhez hozzárendeli annak valószínűségét; gyakran grafikonként vagy táblázatként vagy képlettel van kifejezve

Definíciók

- **Diszkrét véletlen változó**

  vagy véges sok, vagy megszámlálhatóan sok számú értéket vehet fel

- **Folytonos véletlen változó**

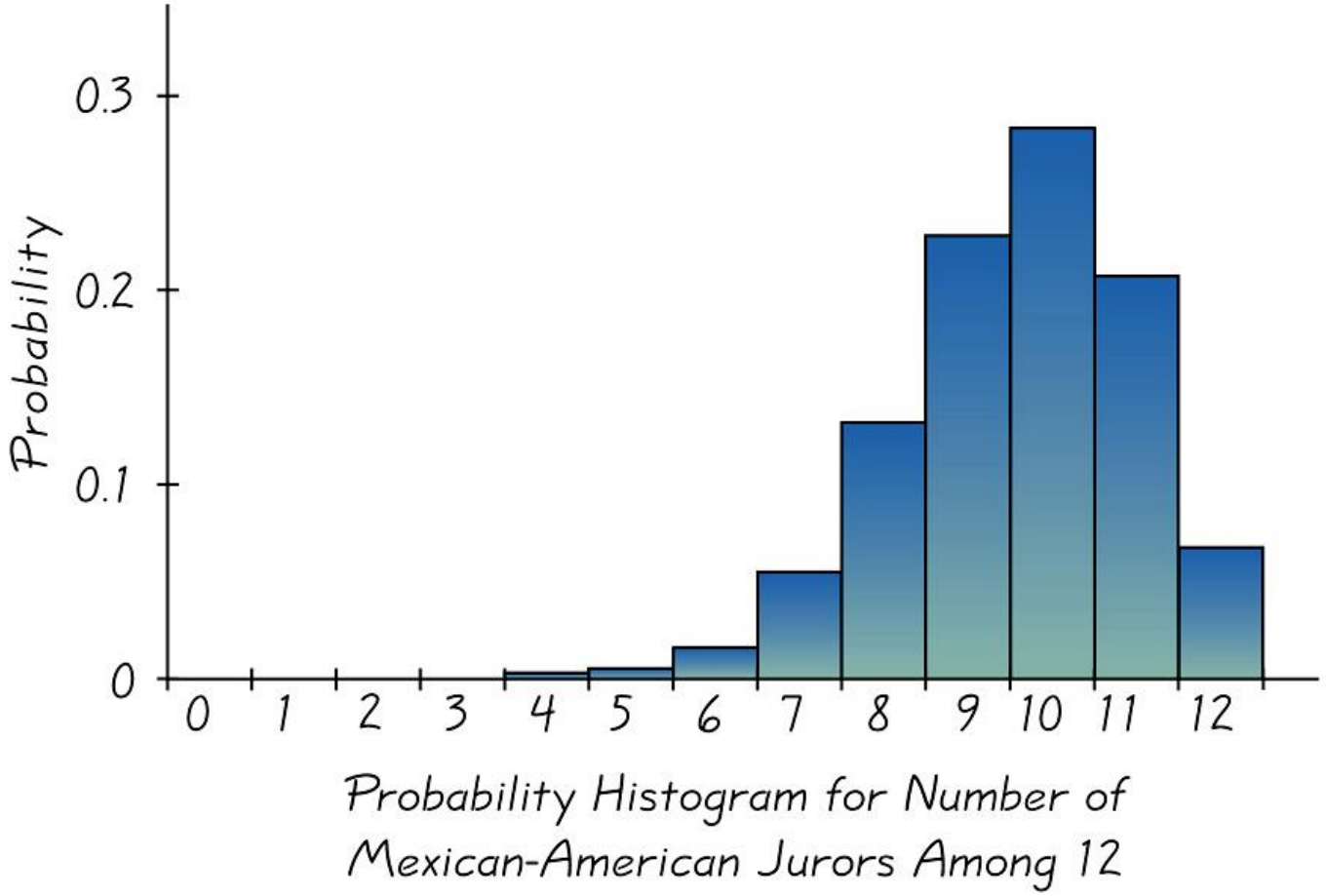
  végtelen sok értéket vehet fel, melyek valamilyen folytonos skálán megadható mérés eredményeiként adódnak, és nem tartalmaznak hiányokat vagy szakadásokat

Példa

- Texasban, Hidalgo járásban annak a valószínűsége, hogy az esküdtszék 12 tagja közül hány Mexikói-Amerikai. A lakosság 80%-a Mexikói-Amerikai.
- Szokatlan-e, hogy egy esküdtszék 7 tagja Mexikói-Amerikai vagy nem?

Grafikonok

A valószínűség hisztogram nagyon hasonló a relatív gyakoriság hisztogramhoz, de a függőleges skála most a valószínűségeket mutatja.



A valószínűség eloszlás fontos tulajdonságai

$$\sum P(x) = 1$$

ahol *P* pozitív értékeket vehet fel.

$$0 \leq P(x) \leq 1$$

minden *x értékre.*

A valószínűség eloszlások átlaga, varianciája és szórása

$$\mu = \Sigma\,[x \cdot P(x)]$$ **Átlag**

$$\sigma^2 = \Sigma\,[(x-\mu)^2 \cdot P(x)]$$ **Variancia**

$$\sigma^2 = [\Sigma\, x^2 \cdot P(x)] - \mu^2$$ **Variancia (rövidített)**

$$\sigma = \sqrt{\Sigma\,[x^2 \cdot P(x)] - \mu^2}$$ **Szórás**

Ritkán előforduló értékek azonosítása

Az értékek nagy része az átlag 2 (3) szórásnyi környezetébe esik. Ezen kívül találhatók a ritka értékek.

A "szokatlan" értékek az alábbi határokon kívülre esnek:

A szokásos értékek maximuma = $\mu + 2\sigma$

A szokásos értékek minimuma = $\mu - 2\sigma$

A ritka értékek azonosítása

Ritka esemény szabály

Ha bizonyos feltevés mellett (mint pl. hogy egy érme szabályos) egy bizonyos bekövetkező esemény megfigyelése (mint pl. 992 fej 1000 dobásból) nagyon kicsi, akkor arra következtetünk, hogy a feltevés nem igaz.

- **Szokatlanul sok**: ***x*** **siker** ***n*** **próbálkozásból**
  szokatlanul sok ha *P*(*x* vagy több siker) ≤ 0.05 (0.003).
- **Szokatlanul kevés**: ***x*** **siker** ***n*** **próbálkozásból**
  szokatlanul kevés ha *P*(*x* vagy kevesebb siker) ≤ 0.05 (0.003).

Definíció

A diszkrét véletlen változó várható értékét általában *E* jelöli, ami a kimenetek átlaga. Értékét úgy kaphatjuk meg, ha kiszámítjuk a $\Sigma [x \bullet P(x)]$ kifejezés értékét.

$$E = \sum [x \bullet P(x)]$$

Összefoglalás

Ebben a fejezetben megvitattuk:

- **A leíró statisztika és a valószínűségek kombinálását.**
- **Véletlen változókat és eloszlásukat.**
- **Valószínűség histogrammokat.**
- **A valószínűség eloszlások tulajdonságait.**
- **Átlagot, varianciát és szórást a vsz. eloszlás esetén.**
- **A különös esetek azonosítását.**
- **A várható értéket.**

5-3. fejezet
Binomiális Eloszlás

Kulcsfogalmak

Ebben a fejezetben bemutatjuk a binomiális eloszlás definícióját és a valószínűségek értékeinek kiszámítási módját.

A binomiális eloszlást akkor tudjuk használni, ha a kimeneteket két csoportra lehet osztani, mint elfogadható/nem elfogadható, túlélő/elpusztult stb.

Definíciók

A binomiális eloszlás akkor lép fel, ha a véletlen kísérletre a következő feltételek teljesülnek:

1. Mindig fixen rögzített számú kísérletet végzünk .

2. A kísérletek függetlenek. (Bármely egyes kísérlet kimenetele nem befolyásolja a többit.)

3. Minden kísérlet kimeneteleit két csoportba lehet sorolni (általában sikeres és sikertelen).

4. A siker valószínűsége állandó a különböző kísérletekben.

Jelölések a binomiális eloszlással kapcsolatban

S és F (success és failure) jelöli a két lehetsges kimenet csoportot; *p* és *q* jelöli az S és F valószínűségeit, azaz

$P(S) = p$ (*p* = a siker valószínűsége)

$P(F) = 1 - p = q$ (*q* = a sikertelenség vsz.-e)

Jelölések (folyt.)

n	jelöli a próbálkozások fix számát.
x	jelöli n próbálkozás közül a sikeresek számát, így x bármely egész szám lehet 0 és n között, beleértve a határokat is.
p	jelöli a siker valószínűségét egy-egy kísérletben.
q	jelöli a sikertelenség valószínűségét egy-egy kísérletben.
$P(x)$	jelöli annak valószínűségét, hogy pontosan x próbálkozás lesz sikeres n próbálkozás közül.

A binomiális eloszlás képlete

$$P(x) = \frac{n!}{(n-x)!x!} \cdot p^x \cdot q^{n-x}$$

ahol $x = 0, 1, 2, \ldots, n$

és

n = a kísérletek száma

x = a sikerek száma az n próbálkozásból

p = a siker valószínűsége egy-egy kísérletben

q = a sikertelenség valószínűsége ($q = 1 - p$)

A képlet indoklása

$$P(x) = \frac{n!}{(n-x)!x!} \cdot p^x \cdot q^{n-x}$$

A pontosan x sikert tartalmazó kimenetek száma az n kísérlet esetén

Indoklás (folyt.)

$$P(x) = \frac{n!}{(n-x)!x!} \cdot p^x \cdot q^{n-x}$$

A pontosan x sikert tartalmazó kimenetek száma n kísérlet esetén

bármilyen sorrendben bekövetkező x siker valószínűsége az n kísérlet esetén

Összefoglalás

Ebben a fejezetben bemutattuk a:

- **A binomiális eloszlás definícióját.**
- **Jelölések.**
- **A képlet indoklása.**

5-4. fejezet
A binomiális eloszlás átlaga, varianciája, és szórása

Kulcsfogalmak

Ebben a fejezetben a binomiális eloszlás fontosabb tulajdonságait tekintjük át, kiszámítjuk az átlagát, a varianciáját és szórását.

Ugyanúgy mint eddig, a cél nem az, hogy ezeket kiszámítsuk, hanem hogy interpretáljuk és megértsük .

Diszkrét eloszlásokra vonatkozó képletek:

Átlag $\mu = \Sigma[x \bullet P(x)]$

Variancia $\sigma^2 = [\Sigma\, x^2 \bullet P(x)\,] - \mu^2$

Szórás $\sigma = \sqrt{[\Sigma\, x^2 \bullet P(x)\,] - \mu^2}$

A binomiális eloszlásra vonatkozó képletek:

Átlag $\mu = n \cdot p$

Variancia $\sigma^2 = n \cdot p \cdot q$

Szórás $\sigma = \sqrt{n \cdot p \cdot q}$

Ahol

n **= a kísérletek rögzített száma**

p **= a siker valószínűsége**

q **= a sikertelenség valószínűsége**

Összefoglalás

Ebben a fejezetben megvitattuk az:

- **A binomiális eloszlás átlagát, varianciáját és szórását.**
- **Az eredmény interpretálását.**

5-5. fejezet
A Poisson eloszlás

Kulcsfogalmak

A Poisson eloszlás azért fontos, mert nagyon gyakran használjuk ritka (kis valószínűségű) események eloszlásának leírására.

Definíció

A Poisson eloszlás egy diszkrét eloszlás, ami bizonyos események előfordulásának számát adja meg egy adott intervallumban. Az *x* véletlen változó az események előfordulási száma abban az intervallumban. Az intervallum lehet idő, távolság, terület, térfogat vagy hasonló.

Képlete:

$$P(x) = \frac{\mu^x \bullet e^{-\mu}}{x!} \quad \text{ahol } e \approx 2.71828$$

Példa

- Rutherford és Geiger (1910)
- Polonium radioaktív bomlása során az alfa-részecskék számát mérték
- 10.097 alfa részecske 52.16 óra alatt
- 0.0538 alfa részecske/másodperc

A Poisson eloszlás feltételei

- Az *x* véletlen változó bizonyos események előfordulásának számát adja meg egy adott intervallumban.
- Az előfordulásoknak véletlenszerűeknek kell lenniük.
- Az előfordulásoknak függetleneknek kell lenniük egymástól.
- Az előfordulásoknak egyenletesen kell eloszlaniuk az intervallumon belül.

Paraméterek

- Az átlaga μ.
- A szórás $\sigma = \sqrt{\mu}$.

Eltérés a binomiálishoz képest

A Poisson és a binomiális között a következő fontos különbségek vannak:

- **A binomiális eloszlás külön-külön függ a minta *n* méretétől és a *p* valószínűségtől, miközben a Poisson csak a *μ* átlagtól.**

- **A binomiális esetén az *x* lehetséges értékei 0, 1, . . . *n*, míg a Poisson eloszlásnál *x* lehetséges értékei 0, 1, . . . , felső határ nélkül.**

A binomiális közelítése Poissonnal

A Poisson eloszlással jól közelíthető a binomiális, ha *n* nagy és *p* kicsi.

Ökölszabály

- $n \geq 100$
- $np \leq 10$

A binomiális eloszlás közelítése Poissonnal - μ

- $n \geq 100$
- $np \leq 10$

μ kifejezése

$$\mu = n \cdot p$$

Összefoglalás

Ebben a fejezetben megvitattuk a:

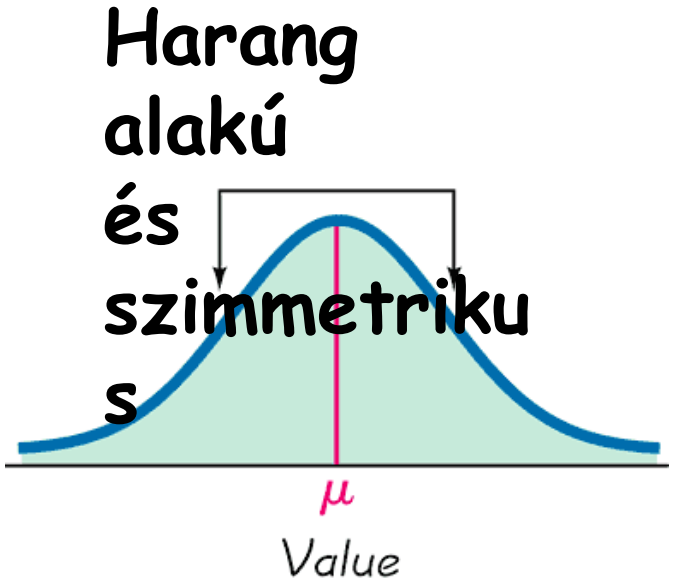
- **Poisson eloszlás definícióját.**
- **A Poisson eloszlás feltételeit.**
- **A Poisson és a binomiális közötti különbséget.**
- **A binomiális Poisson közelítését.**

6-1. fejezet
Áttekintés

Áttekintés

A következő fejezetek a:

- **Folytonos változókról**
- **Normális eloszlásról szólnak**



6-1 ábra

$$f(x) = \frac{e^{\frac{-1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}}{\sigma\sqrt{2\pi}}$$

6-1 képlet

6-2. fejezet
A standard normális eloszlás

Kulcsfogalmak

Ebben a fejezetben a standard normális eloszlást mutatjuk be, aminek három fő tulajdonsága van:

1. Harang alakú.

2. Átlaga 0.

3. Szórása 1.

Nagyon fontos, hogy megtanuljuk, hogyan kell kiszámítani a standard normális eloszlás különböző részei alatti területeket (valószínűségeket vagy relatív gyakoriságot).

Definíció

- **Egy folytonos véletlen változó eloszlása egyenletes eloszlás, ha értékei egyenletesen oszlanak el valamilyen intervallumban. Az egyenletes eloszlás téglalap formájú.**

Definíció

- **Sűrűség függvény** **egy folytonos valószínűség eloszlás görbéje. A következő tulajdonságokkal rendelkezik:**

1. **A görbe alatti teljes terület 1.**
2. **A görbe minden pontja 0 vagy annál nagyobb. (A görbe soha nem eshet az x tengely alá.)**

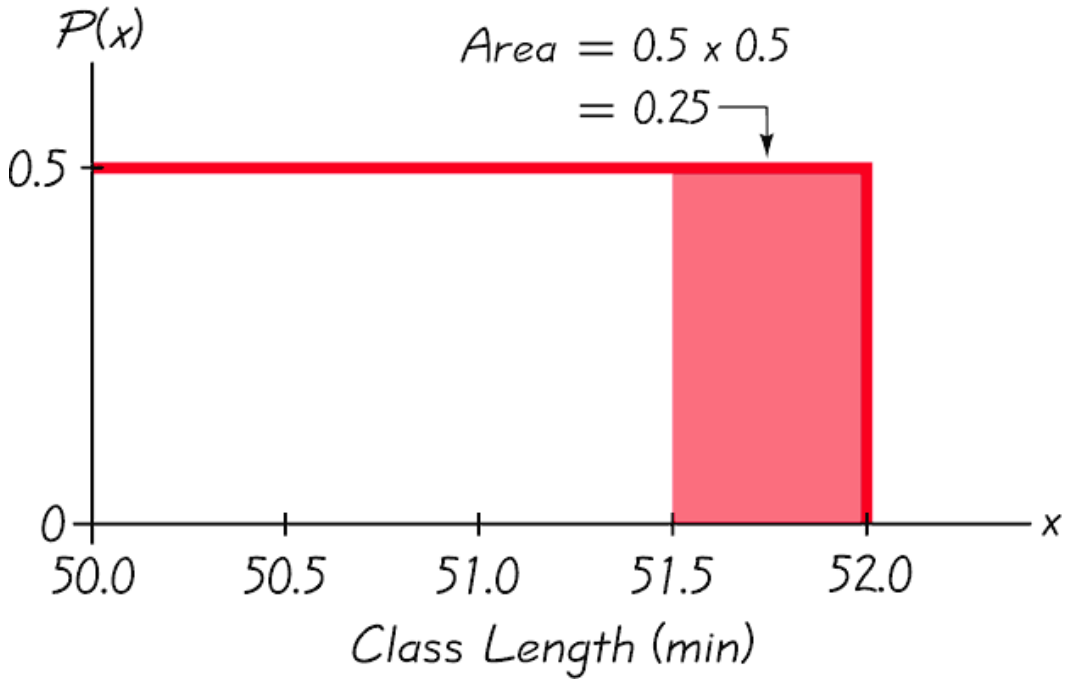
Terület és valószínűség

Mivel a görbe alatti terület 1, kapcsolat van a terület és a valószínűség között.

Példa

- Mivel az elemi statisztika előadások olyan izgalmasak, hosszuk 50 és 52 perc közötti egyenletes eloszlást mutat ☺.
- Neked 51.5 percnél el kell menned. Mi a valószínűsége annak, hogy lekésed a 6-os villamost?

A valószínűség kiszámítása a területből



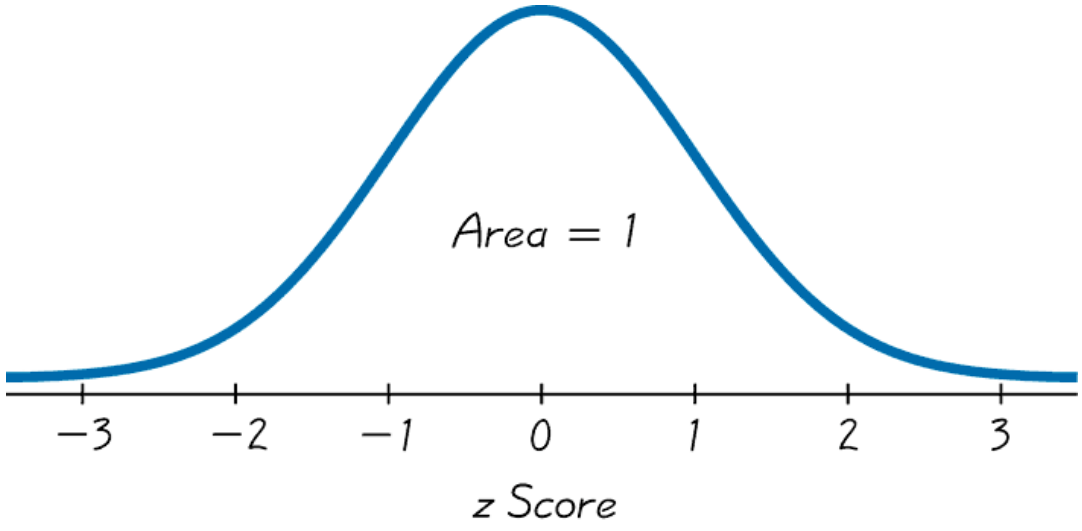
6-3. ábra

Példa - Hőmérők

Legyenek hőmérőink, amelyek átlagban 0-t mutatnak 1 fok szórással ha fagypontban lévő vízbe helyezzük őket. Számítsuk ki, mi a valószínűsége, hogy egy ilyen hőmérő kevesebb mint 1.58 fokot mutat, ha fagypontban lévő vízbe helyezzük.

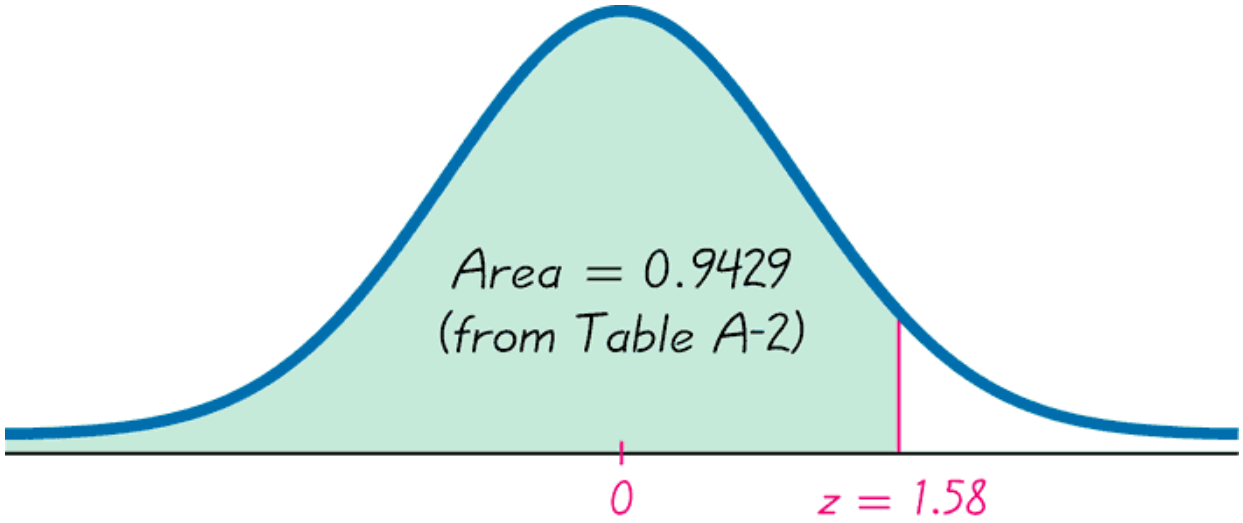
Definíció

- **A standard normális eloszlás egy folytonos valószínűség eloszlás, aminek 0 az átlaga, szórása 1 és a sűrűség függvénye alatti terület is 1.**



Példa – folyt.

P(z < 1.58) =



6-6. ábra

Standard Normál Eloszlás Táblázat

TABLE A-2 Standard Normal (z) Distribution: Cumulative Area from the LEFT

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.50 and lower	.0001									
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
−3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
−2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
−2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
−2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
−2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
−2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	* .0049	.0048
−2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
−2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
−2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
−2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
−2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
−1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
−1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
−1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
−1.6	.0548	.0537	.0526	.0516	.0505	* .0495	.0485	.0475	.0465	.0455
−1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559

TABLE A-2 (*continued*) Cumulative Area from the LEFT

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	* .9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890

A táblázat használata

z érték (z score)

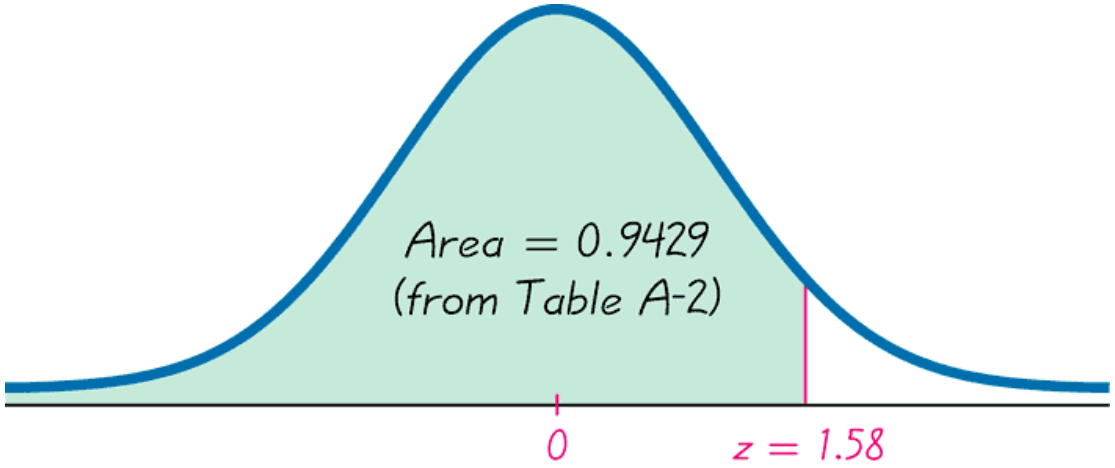
Távolság a standard normál eloszlás vízszintes skáláján a baloldali oszlopban és a legfelső sorban.

Terület (area)

A görbe alatti terület baloldalról mérve a táblázat belsejében levő értékek.

Példa – folyt.

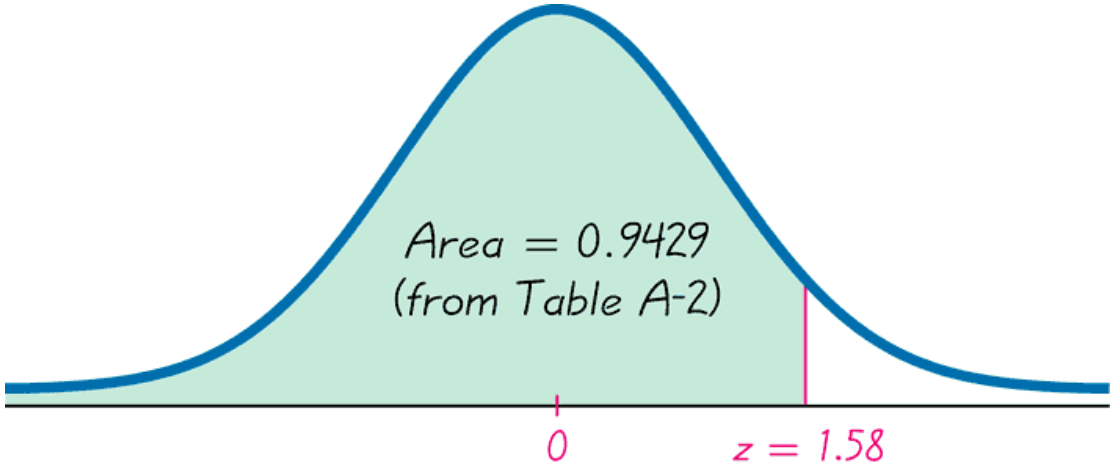
$P(z < 1.58) = 0.9429$



6-6. ábra

Példa – folyt.

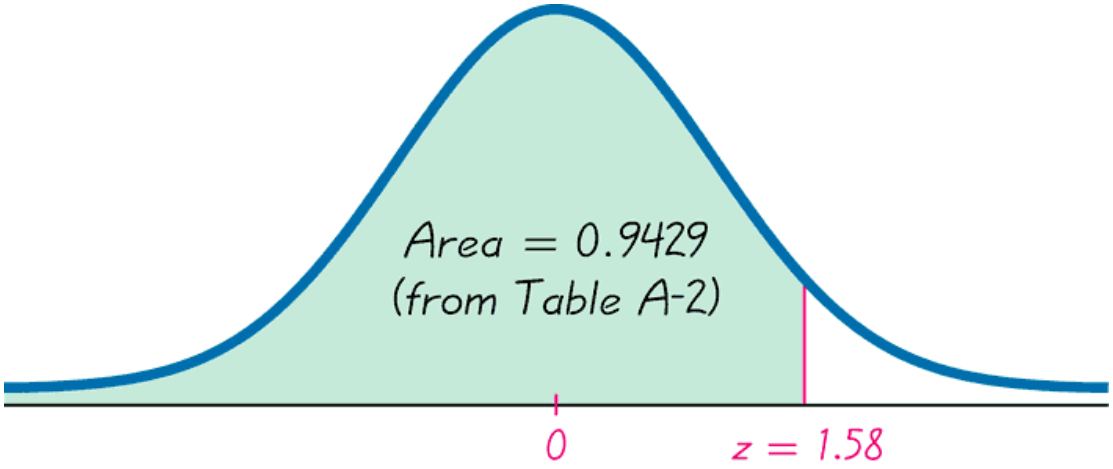
$$P(z < 1.58) = 0.9429$$



Annak a valószínűsége, hogy az egyik hőmérő kevesebb mint 1.58 fokot mutat 0.9429.

Példa – folyt.

$P(z < 1.58) = 0.9429$

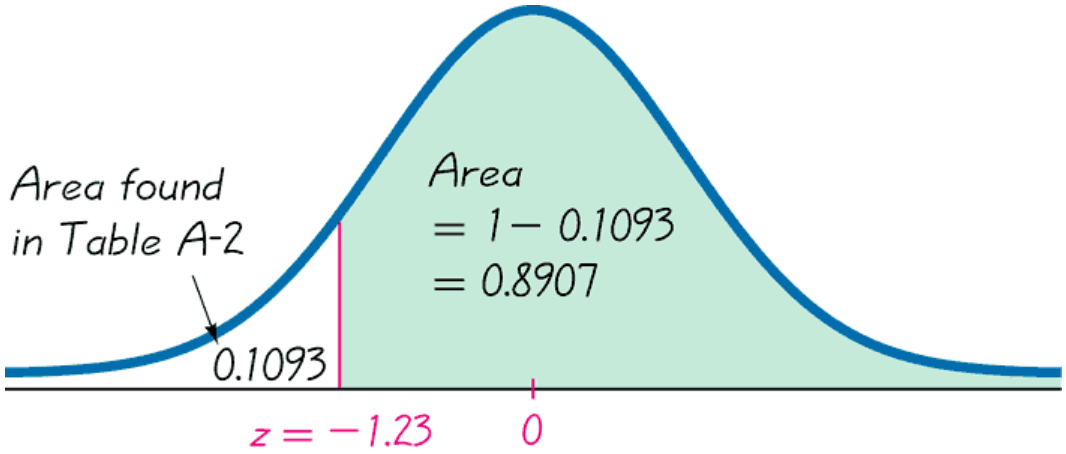


94.29%-a a hőmérőknek kevesebbet mutat mint 1.58 fok.

Példa – folyt.

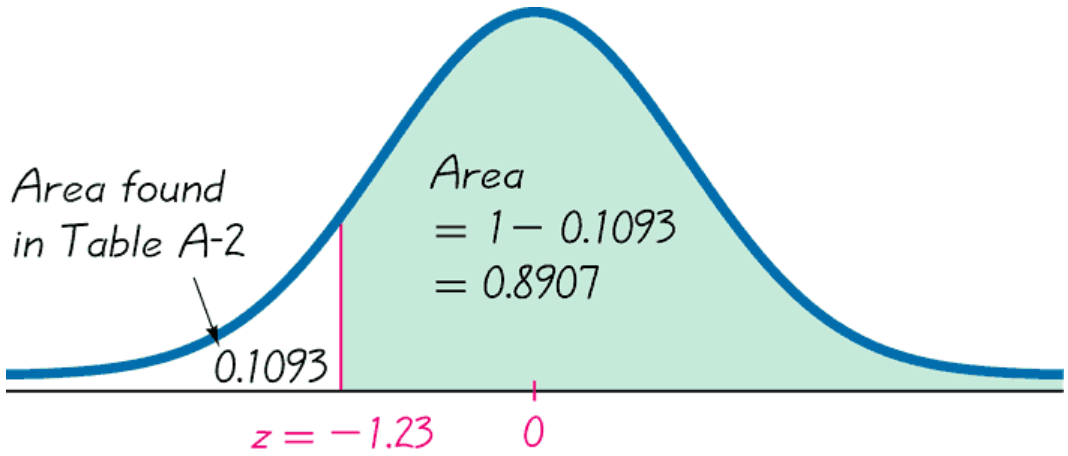
Ugyanolyan hőmérők esetén mi a vsz.-e, hogy egy véletlenül választott hőmérő többet mutat mint –1.23 fok.

$$P(z > -1.23) = 0.8907$$



Példa – folyt.

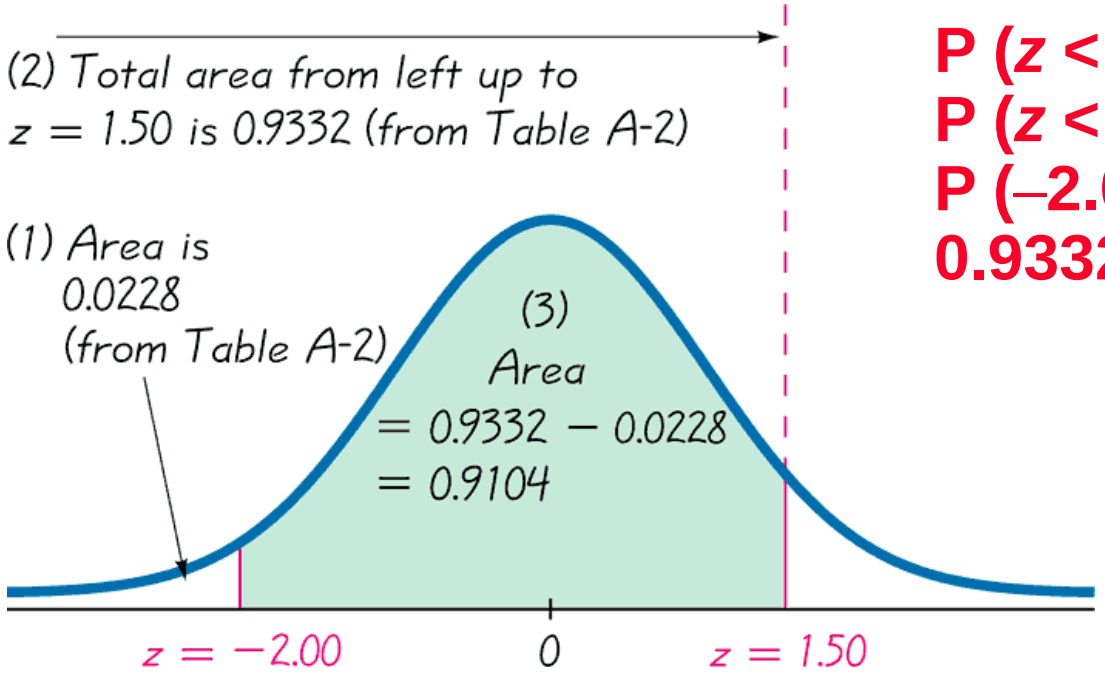
$$P(z > -1.23) = 0.8907$$



89.07%-a a hőmérőknek többet mutat mint –1.23 fok.

Példa – folyt.

Mi a vsz.-e, hogy egy véletlenül választott hőmérő –2.00 és 1.50 fokok közötti értéket mutat.



P (z < –2.00) = 0.0228
P (z < 1.50) = 0.9332
P (–2.00 < z < 1.50) =
0.9332 – 0.0228 = 0.9104

Annak a vsz.-e hogy a hőmérő – 2.00 és 1.50 fokok közötti értéket mutat 0.9104.

Jelölés

$$P(a < z < b)$$

jelöli annak a valószínűségét, hogy a *z* érték *a* és *b* közé esik.

$$P(z > a)$$

jelöli annak a valószínűségét, hogy egy *z* érték nagyobb mint *a*.

$$P(z < a)$$

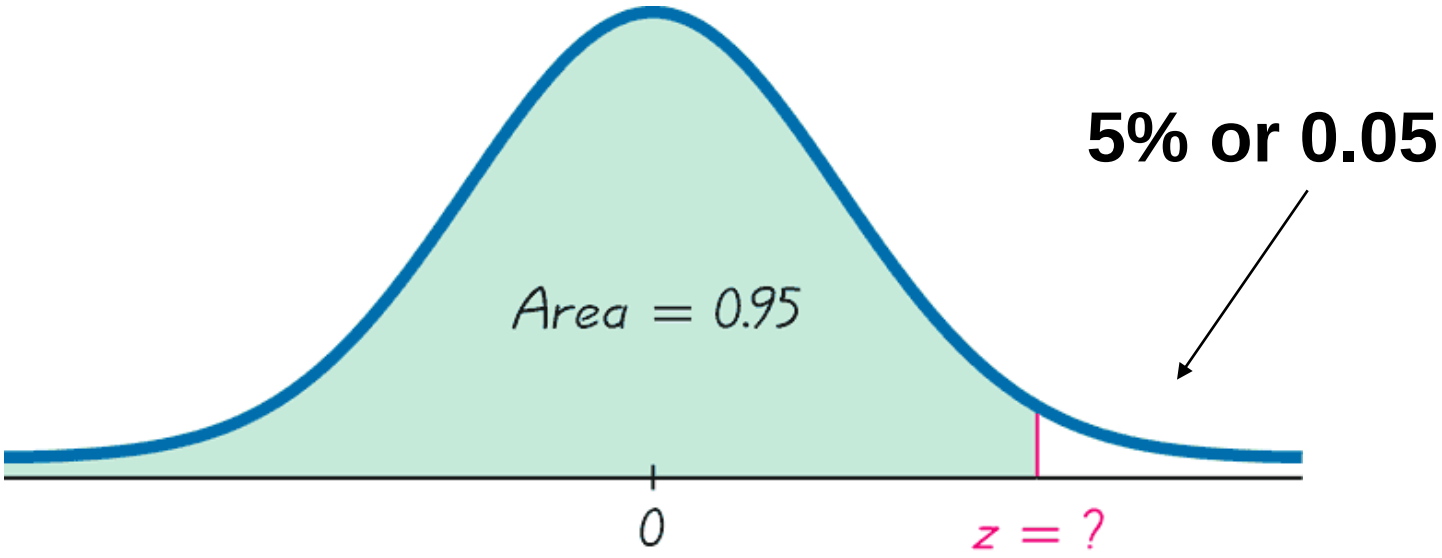
jelöli annak a valószínűségét, hogy egy *z* érték kisebb mint *a*.

A z érték meghatározása a valószínűségből

1. **Rajzolj egy haranggörbét és határozd meg az a területet, ami egy adott valószínűséghez tartozik. Ha ez nem egy baloldalról kumulált terület lenne, akkor vezesd vissza valahogy a problémát ilyenre!**

2. **Keresd meg a táblázat belsejében a megfelelő balról kummulált valószínűséget, és keresd ki hozzá a z értéket.**

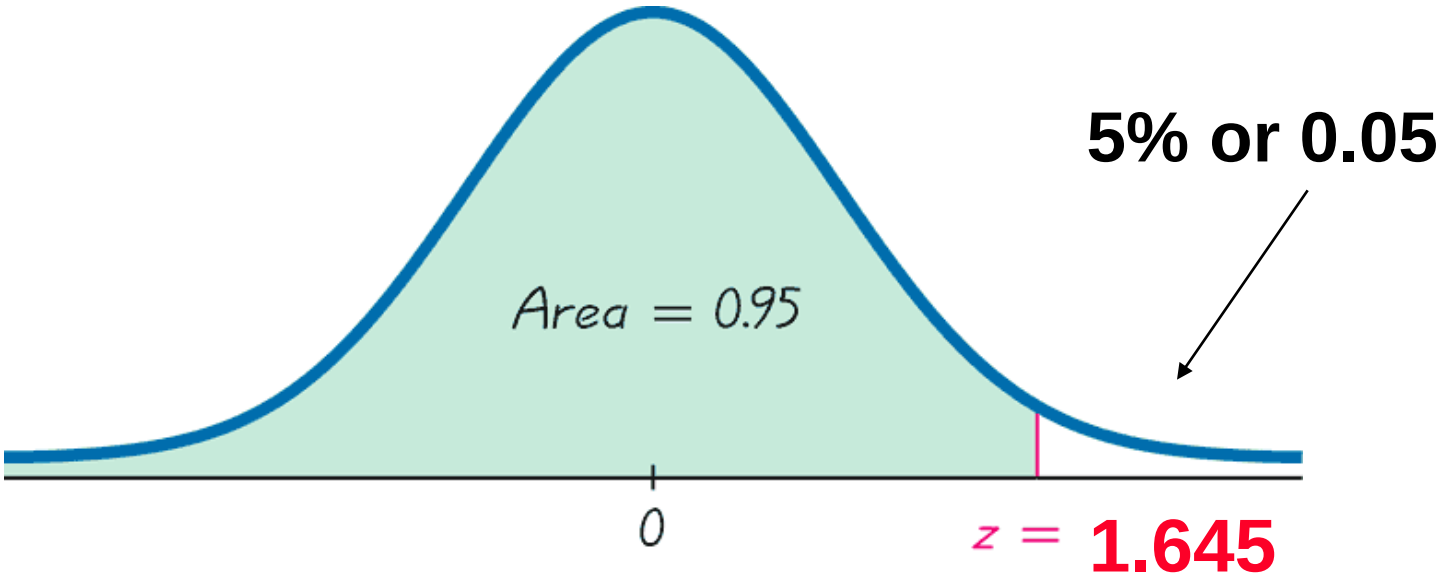
z érték meghatározása a valószínűséghez



(*z* érték pozitív lesz)

6-10. ábra
A 95. Percentilis meghatározása

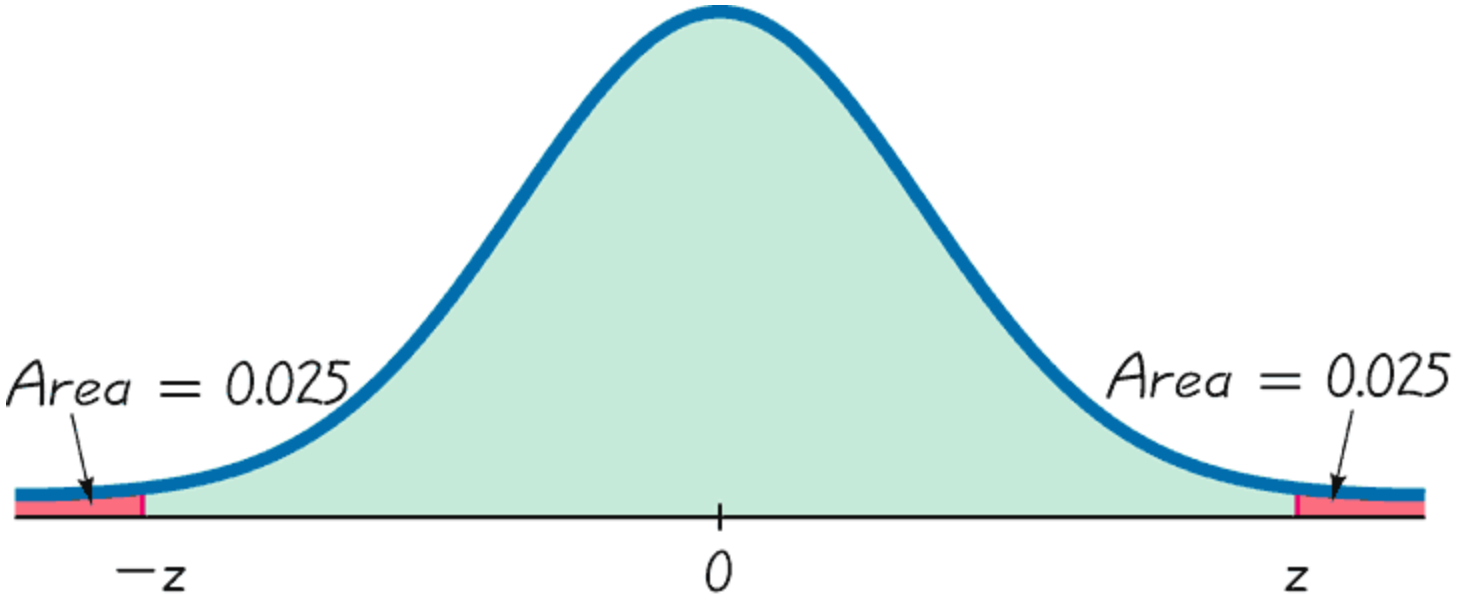
z érték meghatározása a valószínűséghez



(*z* érték pozitív lesz)

6-10. ábra
A 95. Percentilis meghatározása

z érték meghatározása a valószínűséghez

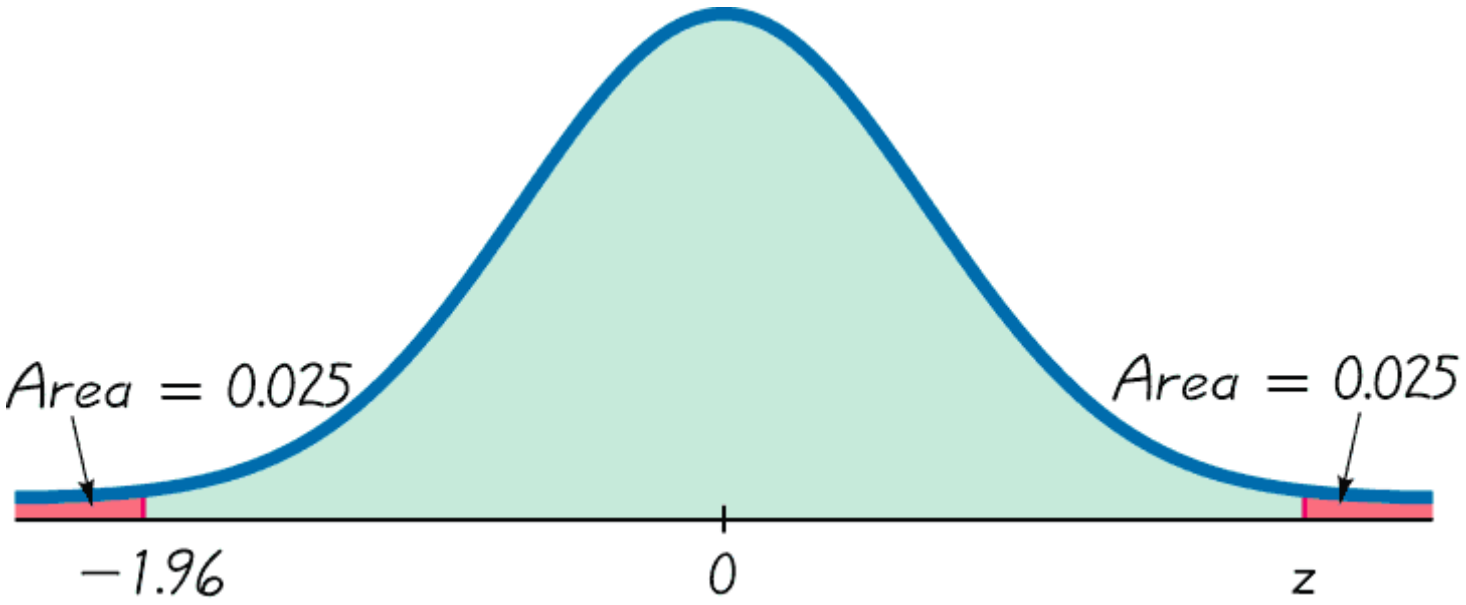


(Az egyik *z* érték negatív, a másik pozitív lesz)

6-11. ábra

Az alsó 2.5% és a felső 2.5% meghatározása

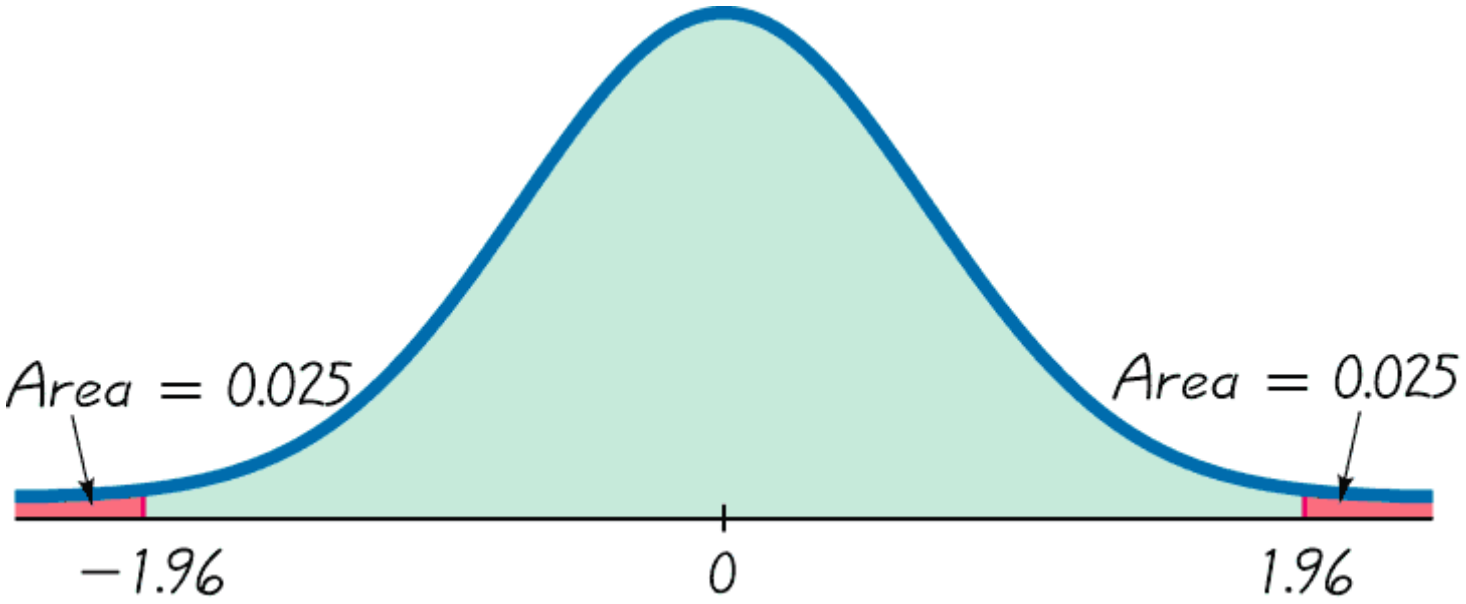
z érték meghatározása a valószínűséghez



(Az egyik *z* érték negatív, a másik pozitív lesz)

6-11. ábra
Az alsó 2.5% és a felső 2.5% meghatározása

z érték meghatározása a valószínűséghez



(Az egyik *z* érték negatív, a másik pozitív lesz)

6-11. ábra

Az alsó 2.5% és a felső 2.5% meghatározása

Összefoglalás

Ebben a fejezetben megvitattuk:

- **A sűrűség függvényt.**
- **A terület és a valószínűség közti kapcsolat**
- **Standard normális eloszlás.**
- **A táblázatok használata.**

6. Előadás
A normális eloszlás

6-3 A normális eloszlás alkalmazásai

6-4 Statisztikák eloszlása és becslő függvények

6-5 A központi határeloszlás törvénye

6-6 A binomiális eloszlás közelítése normálissal

6-7 A normalitás vizsgálata

A fejezet példája:

Nemrég Baltimore belső kikötőjében elsüllyedt egy vízitaxi.

A 25 rajta tartózkodó ember közül 5-en meghaltak, 16-an

megsebesültek. A vizsgálat kimutatta, hogy a biztonságos

össz utas tömeg 1600 kg lett volna. Feltéve, hogy egy utas átlagos tömege 64 kg, 25 utas felvétele volt engedélyezve. A 64 kg-os átlagot 44 évvel ezelőtt állapították meg, amikor az emberek sokkal könnyebbek voltak. (Az elsüllyedt hajó 25 utasának átlagos tömege 76 kg volt.) Az eset után az USA-ban a közlekedési

6-3. fejezet
A normális eloszlás alkalmazásai

Kulcsfogalmak

Ebben a fejezetben átnézzük, hogy hogyan kell olyan normális eloszlásokkal dolgozni, amelyek nem 0 az átlaguk és nem 1 a szórásuk.

A legfontosabb, hogy egyszerűen átkonvertálhatunk egy nem standard eloszlást úgy, hogy az eredmény standard normális eloszlás legyen és így a korábban használt módszereket alkalmazni tudjuk.

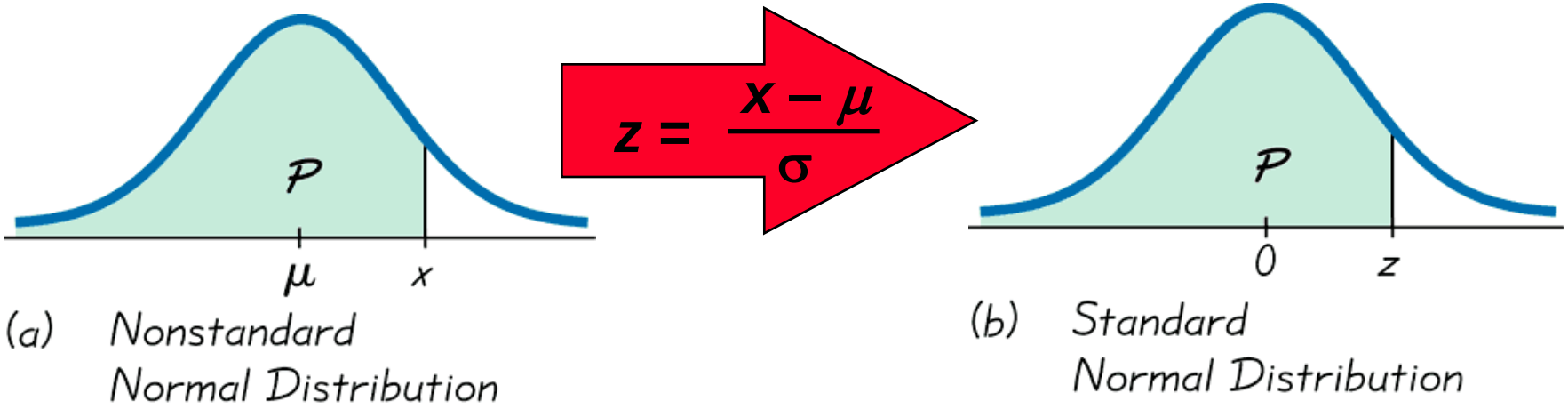
Konverziós formula (standardizálás)

6-2. képlet

$$z = \frac{x - \mu}{\sigma}$$

$$x = \mu + \sigma \cdot z$$

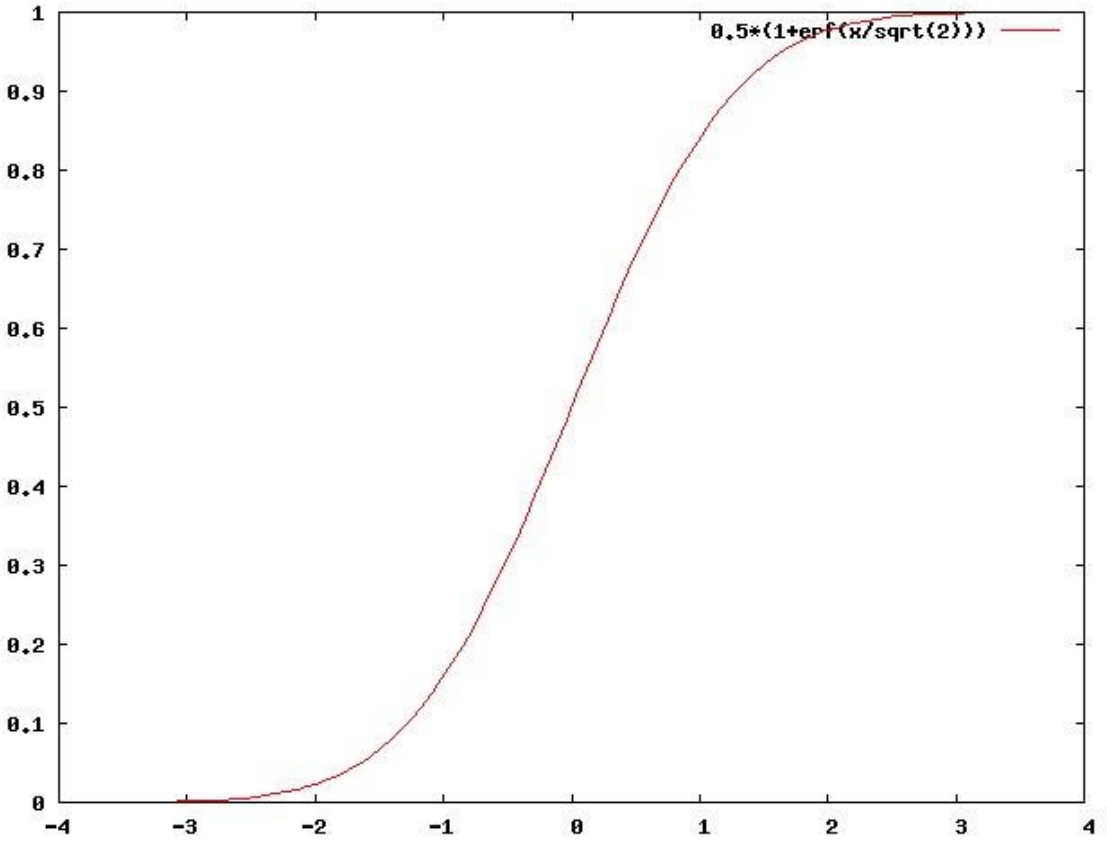
Konvertálás nem-standardból standardba



6-12. ábra

A hiba függvény

$$\Pr(X \leq x) = \Phi\left(\frac{x-\mu}{\sigma}\right) = \frac{1}{2}\left(1 + \operatorname{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right).$$



Példa – a vízitaxi utasainak súlyeloszlása

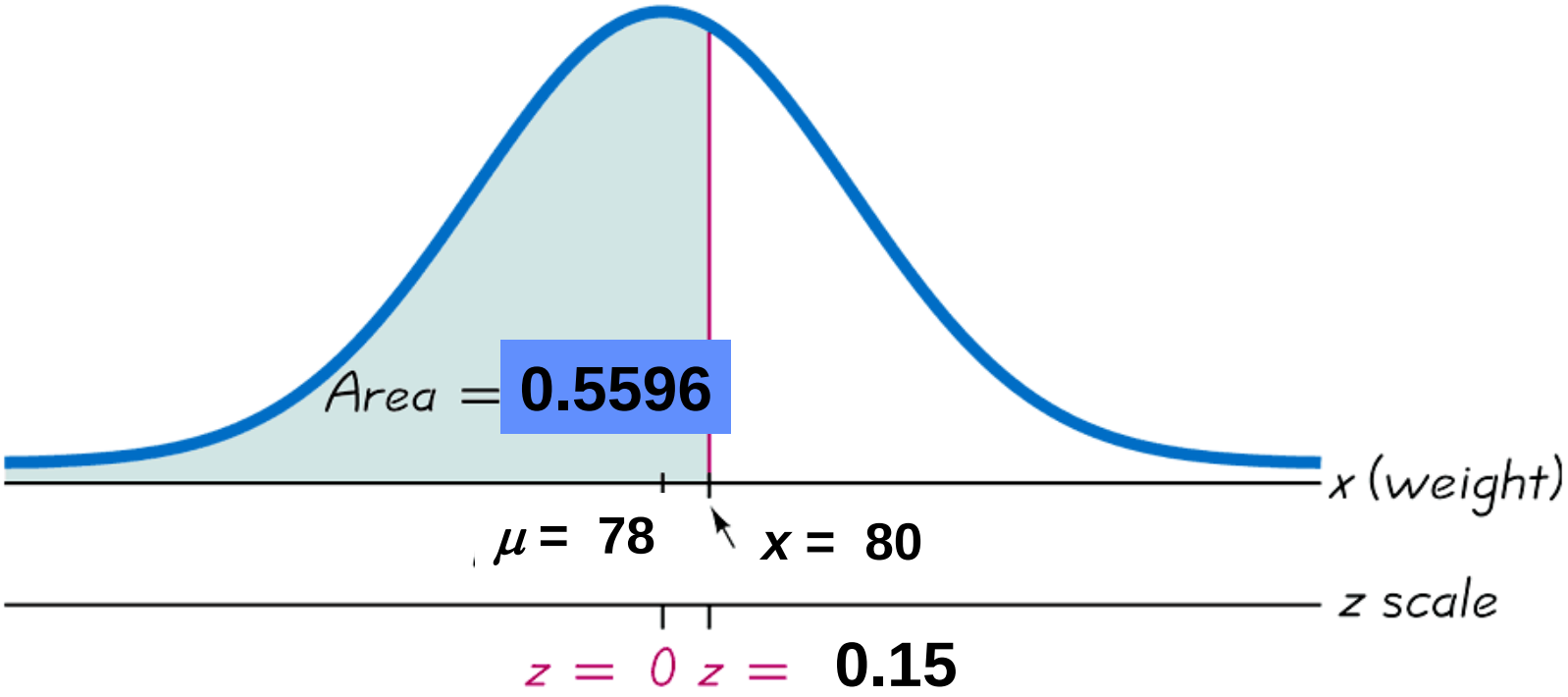
A fejezet elején a vízitaxi megengedett utas tömege 1600 kg volt és az átlagos utas tömegét 64 kg-nak feltételezték. Tegyük fel a legrosszabb esetet, hogy az összes utas férfi. És tegyük fel, hogy a férfiak tömege normális eloszlást követ 78 kg-os átlaggal és 13 kg szórással. Ha véletlenül választunk egyet, mi a valószínűsége annak, hogy tömege kisebb mint 80 kg?

Példa - folyt

$\mu = 78$

$\sigma = 13$

$$z = \frac{80 - 78}{13} = 0.15$$



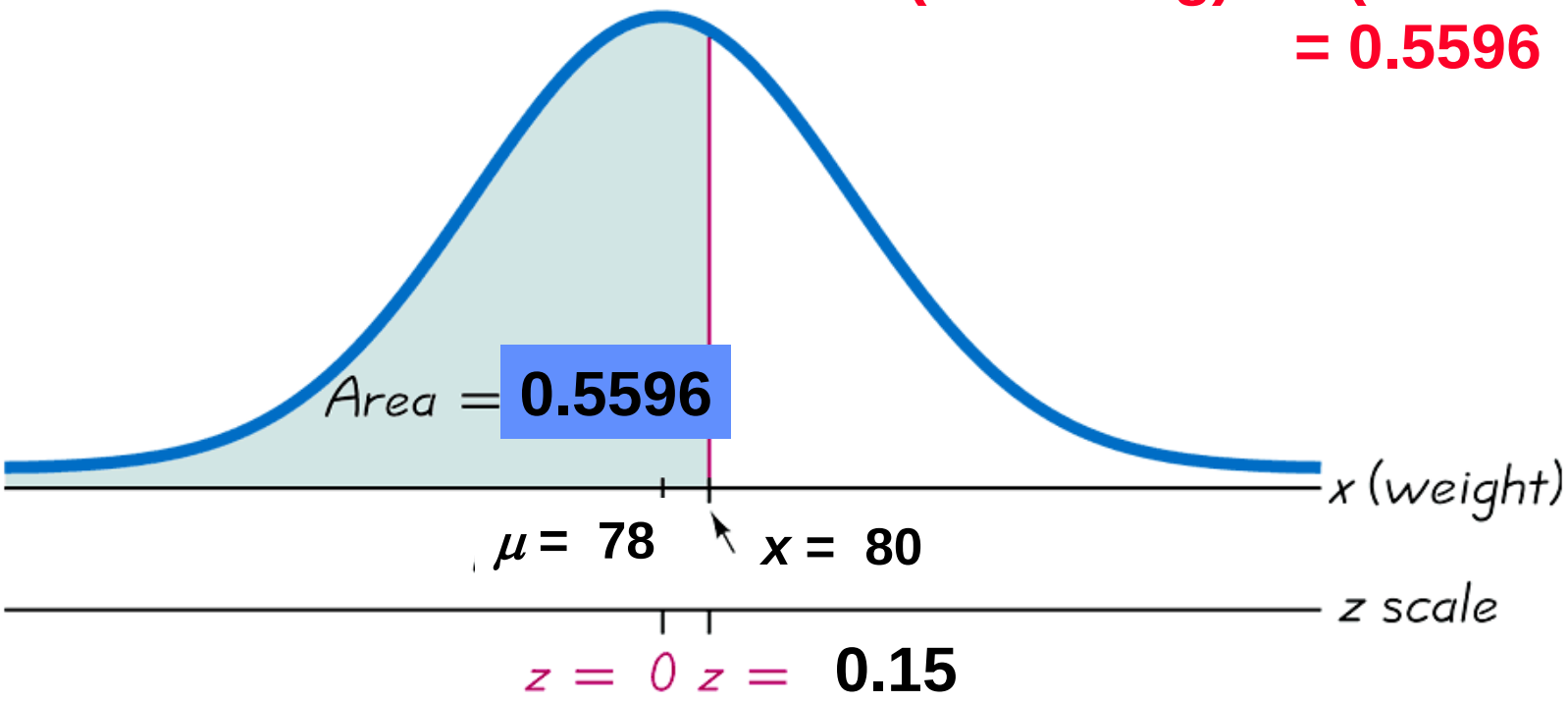
6-13. ábra

Példa - folyt

$\mu = 78$

$\sigma = 13$

$P(x < 80\ kg) = P(z < 0.15)$
$= 0.5596$



6-13. ábra

A változó értékeinek megtalálása 6-2. képlet segítségével

1. **Rajzolj egy normális eloszlás görbét, rajzold be, hogy hol és milyen valószínűségeket vagy százalékokat keresel, és rajzold be a keresett x értékeket!**

2. **A táblázatot használva keressük meg azt a *z* értéket, amelyik az x-től balra eső területhez tartozik. A táblázat belsejében keresd ki a területet és abból a z értéket!**

3. **A 6-2. képletet használva, írd be μ, σ, értékét és a *z értéket* és számítsd ki x-et:**

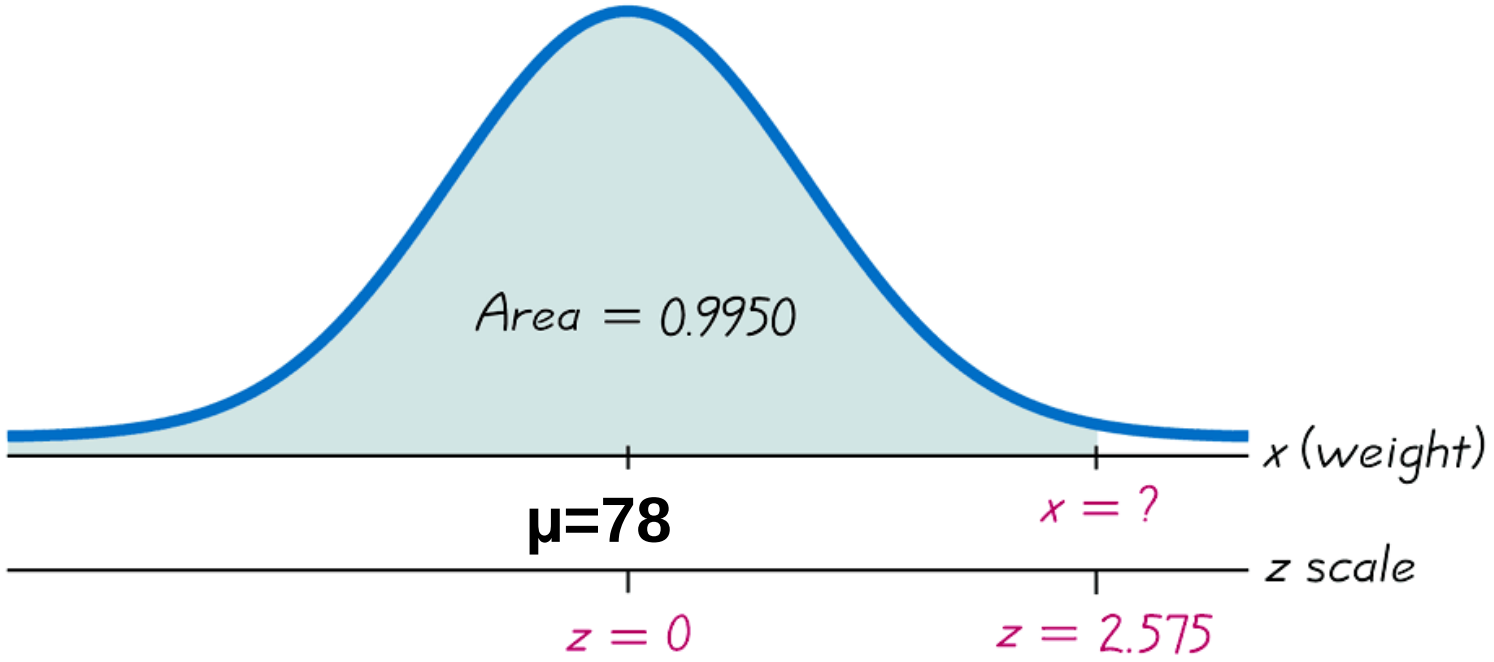
$$x = \mu + (z \bullet \sigma)$$ **(6-2. másik alakja)**

(Ha *z* a haranggörbe baloldalán van, akkor z negatív a képletben.)

4. **Nézd meg az eredeti ábrán, hogy értelmes-e az eredmény.**

Példa – A legkönnyebb és a legnehezebb

A példa adatait használva határozzuk meg mekkora az a súly, ami a legkönnyebb 99.5%-ot elválasztja a legnehezebb 0.5%-tól?

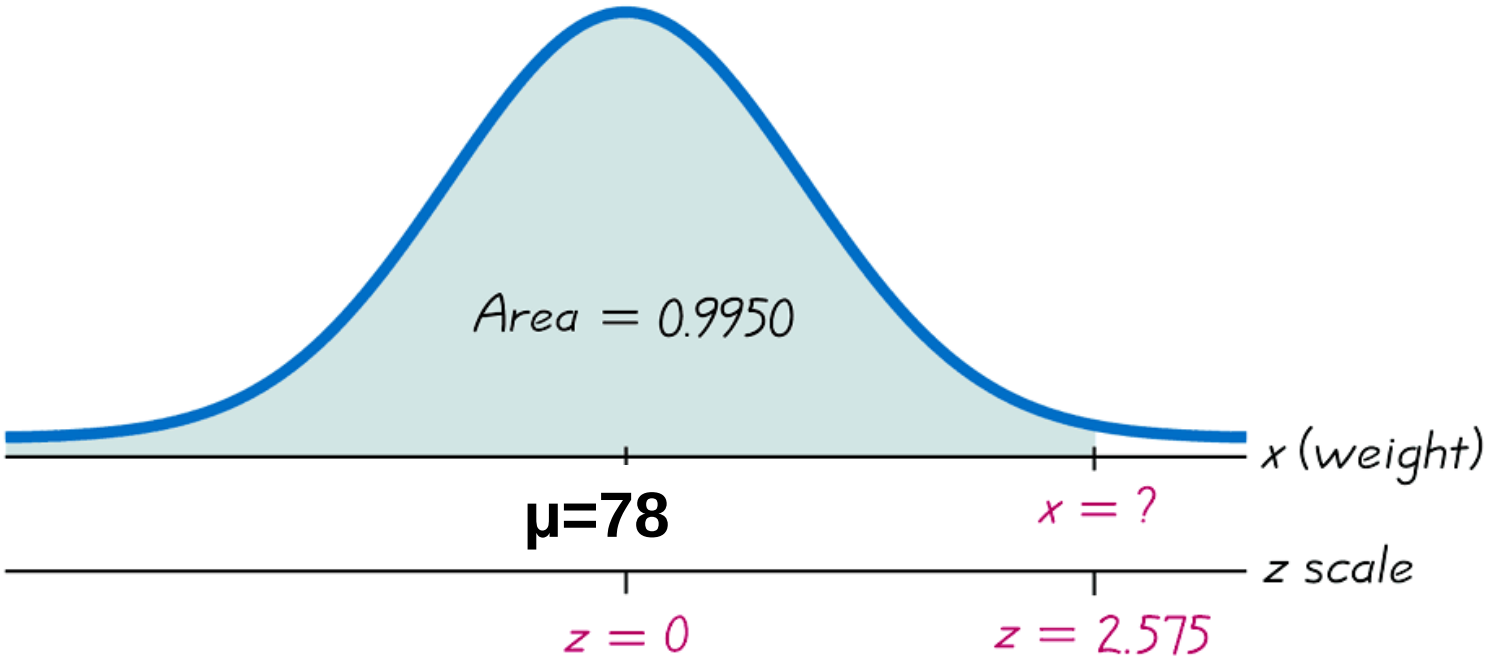


Példa – folyt

$x = \mu + (z \bullet \sigma)$

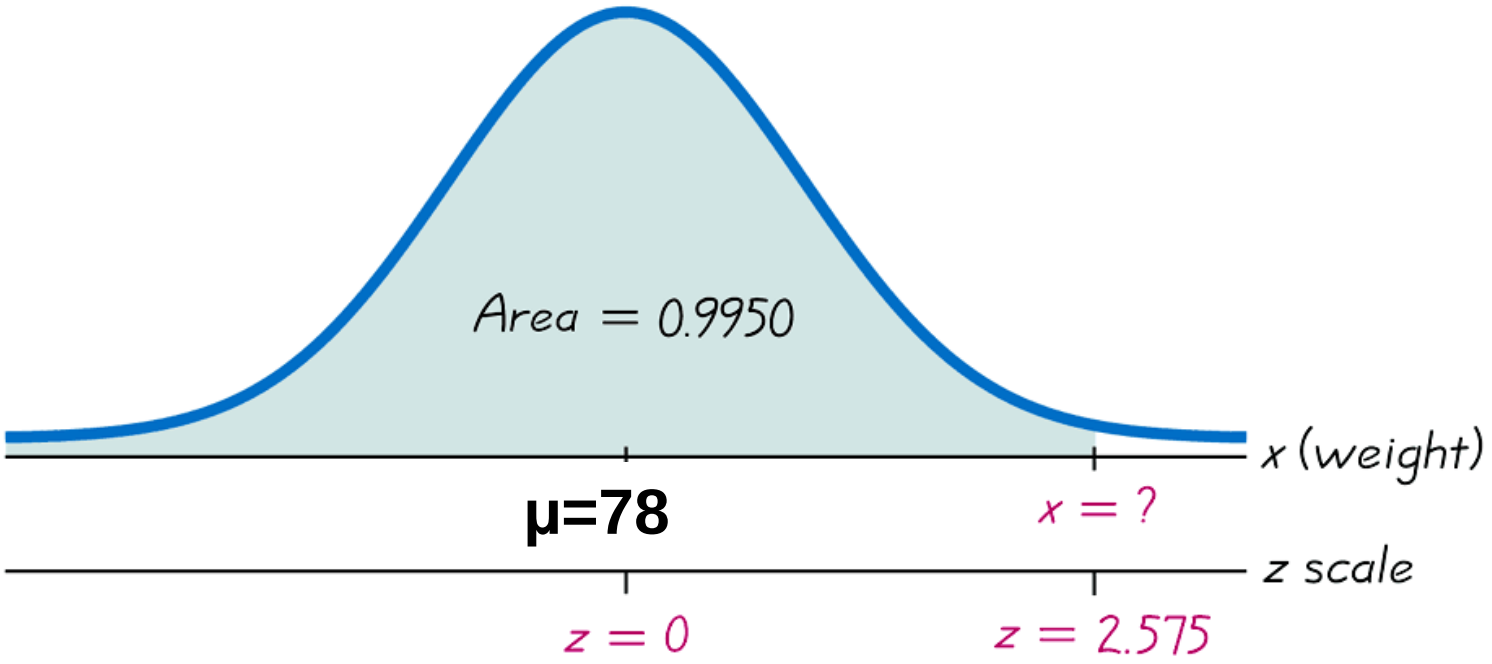
$x = 78 + (2.575 \bullet 13)$

$x = 111{,}475$



Példa – folyt.

Kb. 111 kg a választópont a 99.5% legkönnyebb és a 0.5% legnehezebb között.



Összefoglalás

Ebben a fejezetben megvitattuk:

- **A nem standard normális eloszlást.**
- **A standard normálisba konvertálást.**

6-4. fejezet
A statisztikák eloszlásai és becslések

Kulcsfogalmak

A fejezet célja, hogy bevezessük a statisztika eloszlását, ami az adott statisztika értékeinek eloszlása abban az esetben, amikor az értékeket a populációból kiválasztott minden lehetséges adott elemszámú mintára kiszámítjuk.

Látni fogjuk, hogy bizonyos statisztikák jobbak mint mások a populáció paramétereinek becslésére.

Definíció

❖ **A statisztika eloszlása (mint például a minta arány vagy a minta átlag eloszlása) a statisztika minden lehetséges értékének eloszlása abban az esetben, amikor értékét a populáció minden lehetséges *n* elemszámú mintájára kiszámítjuk.**

Definíció

❖ Az **arány eloszlása** valami mintabeli arányának eloszlása, a populáció minden lehetséges *n* elemszámú mintájában.

Tulajdonságok

- A minta arányok a populációs arányhoz tartanak. (Azaz a lehetséges minták arányainak átlaga egyenlő az „igazi” populációs aránnyal.)
- Bizonyos feltételek mellett a mintabeli arányok eloszlása normális eloszlással közelíthető.

Definíció

❖ Az **átlag eloszlása** a minták átlagainak eloszlása abban az esetben, ha a populációból vett összes lehetséges *n* elemszámú mintát vesszük. (Az átlag eloszlását általában táblázatosan megadott valószínűség eloszlásként, hisztogramként vagy képlettel prezentáljuk.)

Definíció

❖ **A statisztika értéke, mint például a minta átlag $\bar{x}$, függ a mintába kerülő konkrét értékektől, és általában mintáról mintára változik. A statisztikának ezt a variabilitását minta variabilitásnak nevezzük.**

Becslő függvények (becslések)

Bizonyos statisztikák sokkal jobbak, mint mások a populáció paramétereinek becslésére. A következő példa ezt mutatja be.

Példa

A populáció álljon az 1, 2, és 5 értékekből. Véletlenülszerűen, visszatevéssel választunk 2 elemszámú mintákat. Összesen 9 minta lehetséges.

a. Minden mintára megkeressük az átlagot, a mediánt, a terjedelmet, a varianciát és a szórást.

b. Mindegyik statisztikára számítsuk ki ezek átlagát.

Table 6-7 Sampling Distributions of Statistics (for Samples of Size 2 Drawn with Replacement from the Population 1, 2, 5)

Sample	Mean $\bar{x}$	Median	Range	Variance s^2	Standard Deviation s	Proportion of Odd Numbers	Probability
1, 1	1.0	1.0	0	0.0	0.000	1	1/9
1, 2	1.5	1.5	1	0.5	0.707	0.5	1/9
1, 5	3.0	3.0	4	8.0	2.828	1	1/9
2, 1	1.5	1.5	1	0.5	0.707	0.5	1/9
2, 2	2.0	2.0	0	0.0	0.000	0	1/9
2, 5	3.5	3.5	3	4.5	2.121	0.5	1/9
5, 1	3.0	3.0	4	8.0	2.828	1	1/9
5, 2	3.5	3.5	3	4.5	2.121	0.5	1/9
5, 5	5.0	5.0	0	0.0	0.000	1	1/9
Mean of Statistic Values	8/3	8/3	16/9	26/9	1.3	2/3	
Population Parameter	8/3	2	4	26/9	1.7	2/3	
Does the sample statistic target the population parameter?	Yes	No	No	Yes	No	Yes	

Interpretáció

Láthatjuk, hogy bizonyos statisztikák jók abban az értelemben, hogy a populáció paramétereihez tartanak. Az ilyen statisztikákat torzítatlan becsléseknek nevezik.

Olyan statisztikák, melyek a populációs paraméterekhez tartanak: átlag, variancia, részarány

Olyan statisztikák, melyek nem tartanak a populáció paramétereihez: medián, terjedelem, szórás

Összefoglalás

Ebben a fejezetben megvitattuk:

- **Statisztika eloszlását.**
- **Az arány eloszlását.**
- **Az átlag eloszlását.**
- **A minta variabilitását.**
- **Becsléseket.**

6-5. fejezet
A központi határeloszlás tétel

Kulcsfogalmak

Ebben a fejezetben megalapozzuk a populáció paramétereinek becslését és a hipotézis vizsgálatokat, melyről a következő előadások szólnak majd.

Központi határeloszlás tétel

Adott:

1. **Az x véletlen változónak μ átlaga és σ szórással rendelkező eloszlása van (ami vagy normális vagy sem).**
2. **Egyszerű n elemszámú véletlen mintákat választunk a populációból. (A mintákat úgy választjuk, hogy bármely n elemszámú mintát ugyanazzal az eséllyel választunk ki.)**

Központi határeloszlás tétel – folyt.

Konklúziók:

1. A minta átlag $\bar{x}$, ahogy a minta méretét növeljük, a **normális** eloszláshoz tart.

2. A minta átlagok átlaga μ.

3. A minta átlagok szórása pedig $\sigma/\sqrt{n}$.

Általános gyakorlati tanácsok

1. **Általában ha a minta *n* mérete nagyobb mint 30, akkor a minta átlagok eloszlását meglehetősen jól lehet normális eloszlással közelíteni. A közelítés egyre jobb, ahogy *n* növekszik.**
2. **Ha az eredeti populáció maga is normális eloszlású, akkor a minta átlagok eloszlása mindig normális bármely *n-re* (nem csak a 30-nál nagyobb értékek esetén).**

Jelölés

a minta átlagok átlaga

$$\mu_{\bar{x}} = \mu$$

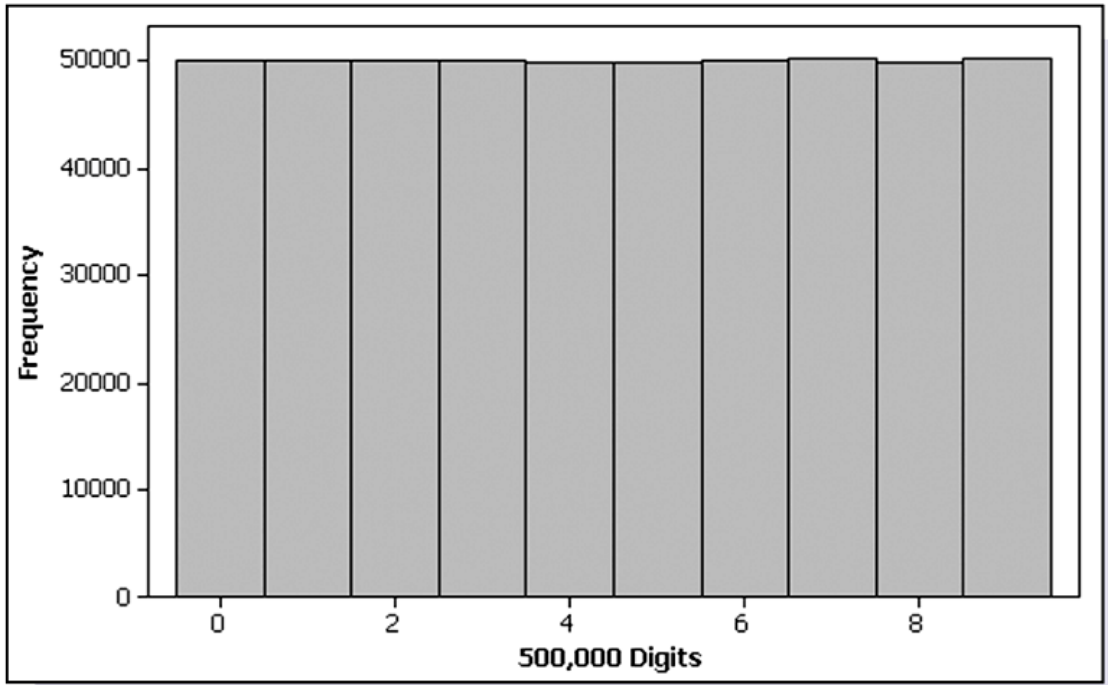
a minta átlagok szórása

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

(gyakran az átlag standard hibájának is nevezik)

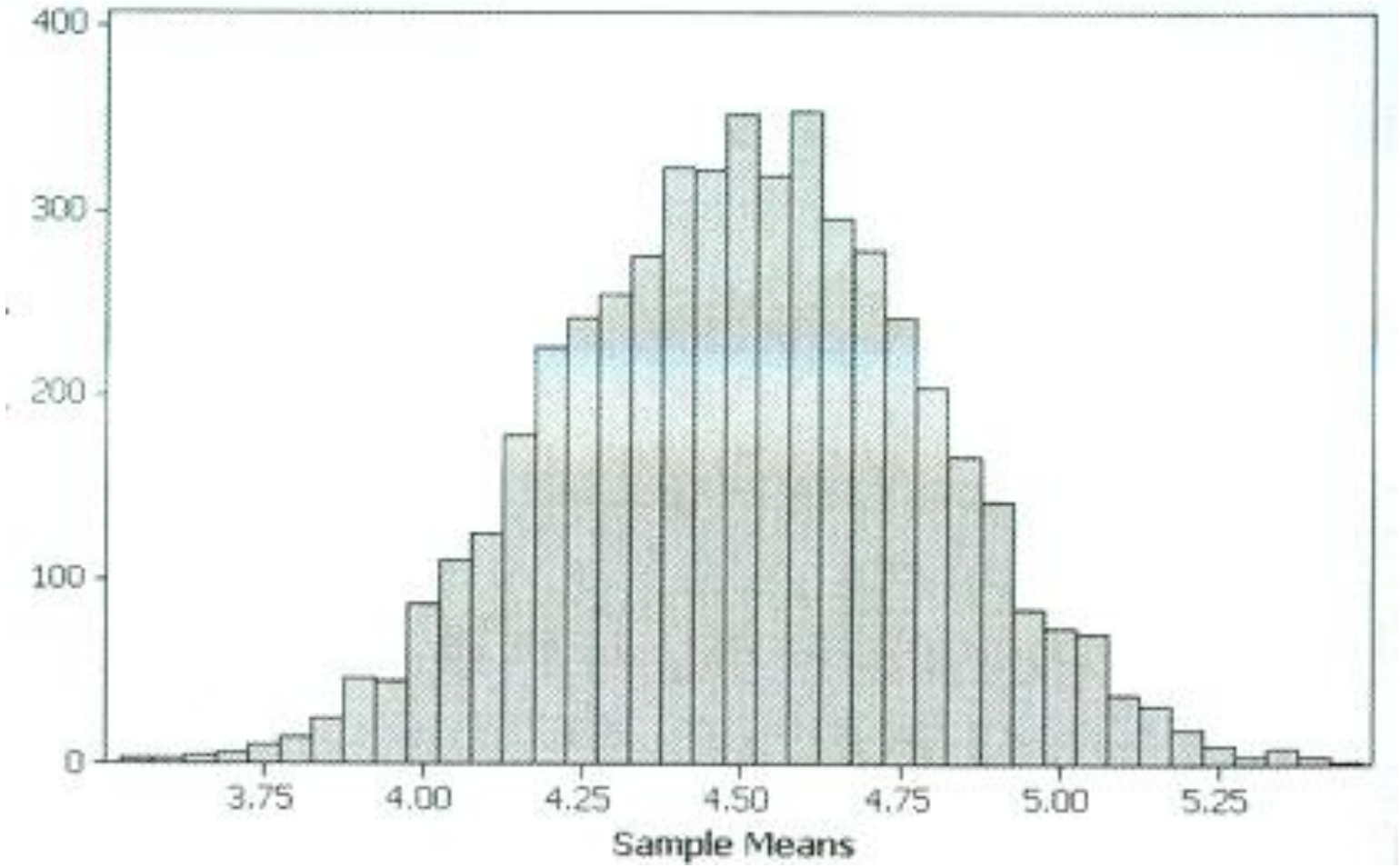
Szimuláció véletlen számokkal

Generáljunk 500,000 véletlen 0 és 9 közötti egész számot, csoportosítsuk 5000 mintába, mindegyikben 100 számmal. Keresd meg mindegyik minta átlagát.



Annak ellenére, hogy az eredeti 500,000 szám egyenletesen oszlik el, az 5000 minta átlag eloszlása normális eloszlás lesz!

5000 db 100 elemű minta átlagainak eloszlása



Fontos felismerés

Ahogy a minta nagyság nő, a minta átlag eloszlása egyre inkább normális lesz.

Példa – vízitaxi biztonság

A férfiak egy adott populációjának tömege normális eloszlású, átlagosan 78 kg a súlya 13 kg szórással,

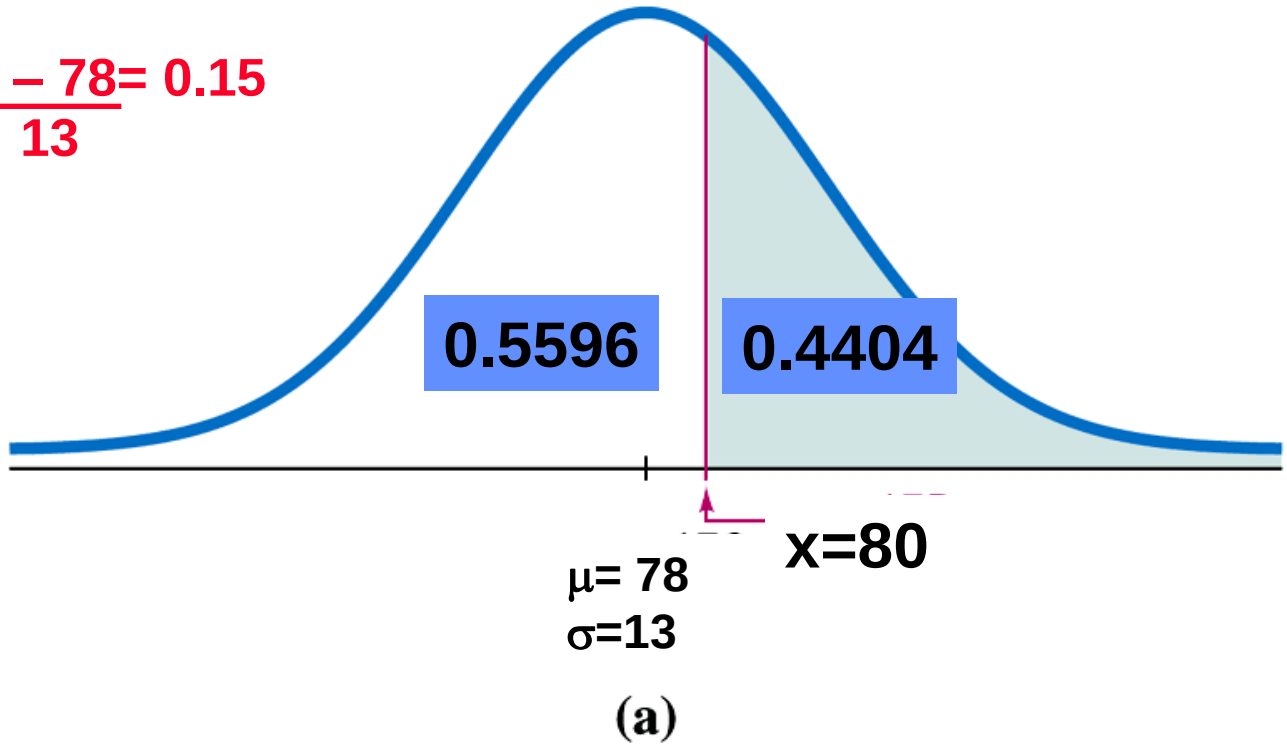
a) ha kiválasztunk egy férfit, mi a valószínűsége annak, hogy a tömege több mint 80 kg.

b) ha 20 különböző férfit véletlenül választunk, számítsuk ki, hogy mi annak a valószínűsége, hogy átlagsúlyuk meghaladja a kritikus 80 kg-ot.

Példa – folyt.

a) egy embert kiválasztva határozzuk meg, hogy mi a valószínűsége annak, hogy tömege több mint 80 kg.

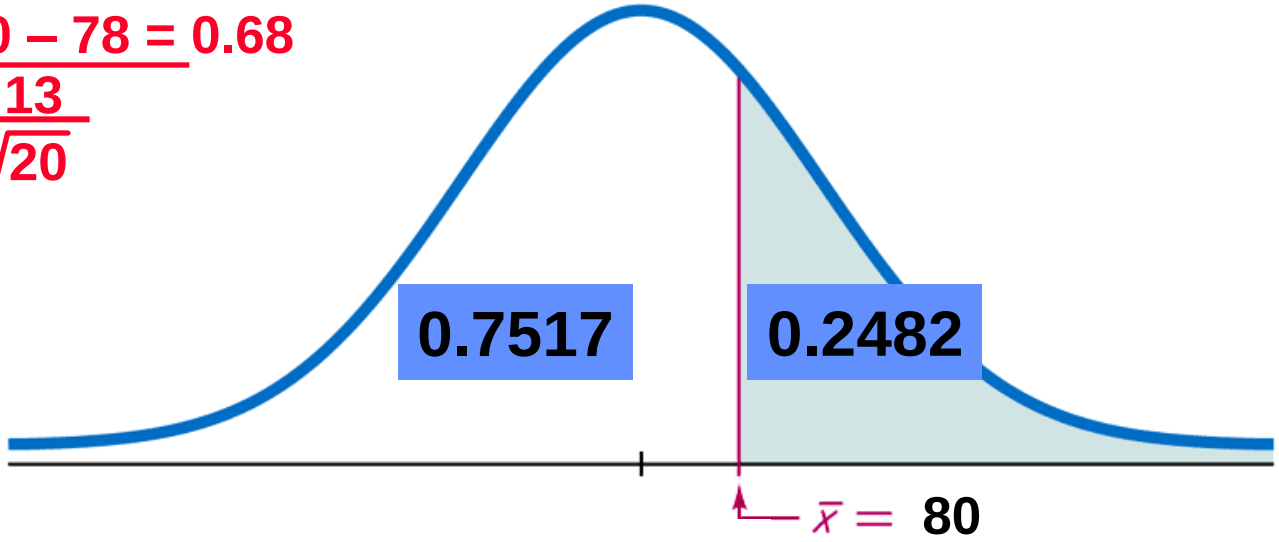
$$z = \frac{80 - 78}{13} = 0.15$$



Példa – folyt

b) ha 20 különböző férfit választunk véletlenül, számítsuk ki annak a valószínűségét, hogy átlagsúlyuk több mint 80 kg.

$$z = \frac{80 - 78}{\frac{13}{\sqrt{20}}} = 0.68$$



$\mu_{\bar{x}} = 78$

$(\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = 13/\sqrt{20}=2,906$

(b)

Példa – folyt.

a) egy véletlenül kiválasztott férfinál annak a valószínűsége, hogy 80 kg-nál nehezebb

$$P(x > 80) = 0.4404$$

b) véletlenül kiválasztott 20 férfi esetén annak a valószínűsége, hogy átlagosan nehezebbek mint 80 kg

$$P(\bar{x} > 80) = 0.2482$$

Egyvalaki esetén sokkal valószínűbb, hogy 80 kg-nál nagyobb, mint hogy 20 férfi esetében az átlaguk nagyobb, mint 80 kg.

Az eredmények értelmezése

Ha a biztonságos kapacitás 1600 kg, akkor elég nagy esélye van annak (24%-os valószínűsége), hogy 20 férfi tömege ezt meg fogja haladni!

Véges populációs korrekció

Ha visszatevés nélkül mintavételezünk, és a minta *n* mérete nagyobb mint 5%-a a véges *N* elemű populációnak, akkor a minta szórást korrigálnunk kell az alábbi faktorral:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

véges populációs korrekciós faktor

Összefoglalás

Ebben a fejezetben megvitattuk:

- **A központi határeloszlás tételt.**
- **Praktikus megfontolásokat.**
- **A mintaméret hatását.**
- **Véges populációs korrekciót.**

6-6. fejezet
A binomiális közelítése normálissal

Kulcsfogalmak

Ebben a fejezetben megmutatjuk, hogy hogyan lehet egy binomiális eloszlást normális eloszlással közelíteni.

Ha az $np \geq 5$ és az $nq \geq 5$ feltételek egyszerre teljesülnek, akkor a binomiális eloszlást egy $\mu = np$ átlagú és $\sigma = \sqrt{npq}$ szórású normális eloszlással jól közelíthető.

Példa

- Egy Boeing 767-300 repülőn 213 ülőhely van.
- A nők átlag tömege 65 kg, a férfiaké 78 kg.
- Ha 122 férfinél több van, akkor vigyázni kell az utasok ültetésére
- Tegyük fel, hogy 50-50% a férfi és nő utasok valószínűsége
- Mi annak a valószínűsége, hogy legalább 122 férfi utas van a gépen.
- Az eloszlás binomiális, de nekünk most 92 esetre kellene kiszámítanunk …

Áttekintés

Binomiális eloszlás

1. A véletlen kísérletek száma **állandó.**
2. A kísérletek **függetlenek.**
3. Minden kísérletnek **két kimenete van.**
4. A siker valószínűsége állandó a kísérletek során.

.

A binomiális közelítése normális eloszlással

$$np \geq 5$$

$$nq \geq 5$$

ekkor $\mu = np$ és $\sigma = \sqrt{npq}$

és a véletlen változó

eloszlása  .

(normal)

A binomiális normálissal való közelítése

1. **Bizonyosodj meg, hogy $np \geq 5$ és $nq \geq 5$ tényleg fennáll.**

2. **Számítsd ki a μ és σ paraméterek értékeit a $\mu = np$ és $\sigma = \sqrt{npq}$ képlettel.**

3. **Azonosítsd x diszkrét értékeit (a sikerek számát). A diszkrét x értéket helyettesítsük az $x - 0.5$ -től $x + 0.5$ –ig intervallummal. (Ld. folytonossági korrekciók még ebben a fejezetben.) Rajzoljuk meg a normális görbét μ , σ, paraméterekkel.**

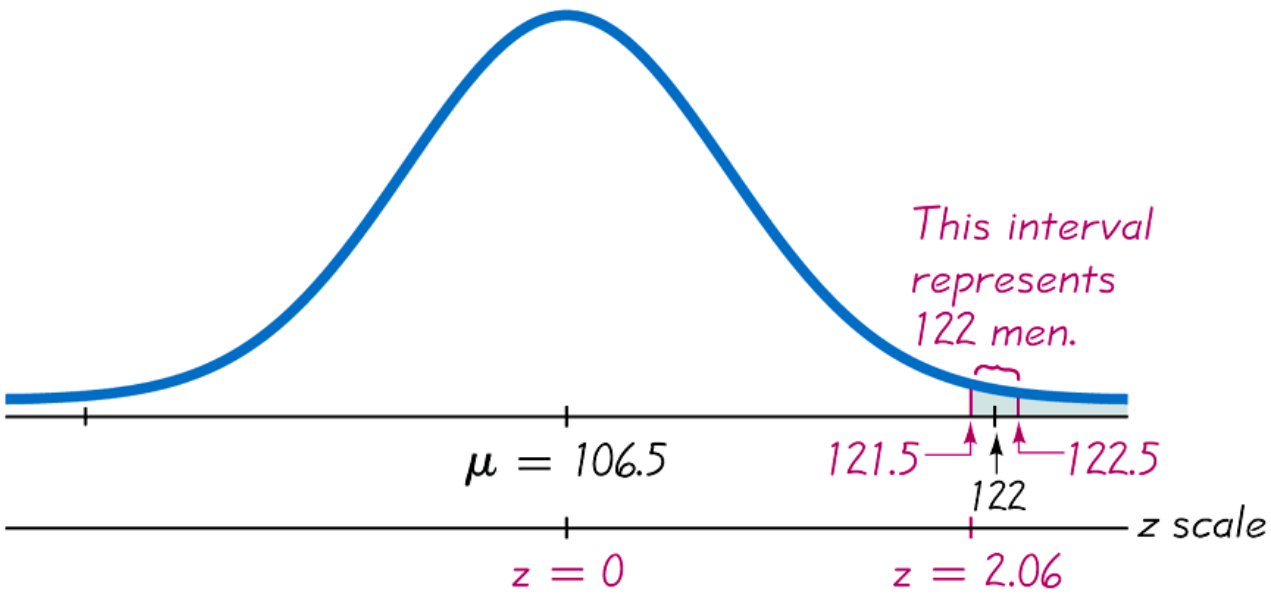
A binomiális normálissal való közelítése Folyt.

4. **Helyettesítsük *x* –et vagy *x* – 0.5 –el, vagy *x* + 0.5 -el, a feladatnak megfelelően.**

5. **Az *x* – 0.5 vagy *x* + 0.5 értéket (a feladatnak megfelelően) használva *x helyett*, keresd meg a kívánt valószínűséget úgy, hogy először a megfelelő *z* értékhez kikeresed a tőle balra fekvő területet.**

Példa – A férfiak száma az utasok között

A “legalább 122 férfi” valószínűségének meghatározása 213 utas esetén



6-21. ábra

Definíció

Amikor a normális eloszlást használjuk (ami egy folytonos eloszlás) arra, hogy a binomiálist közelítsük (ami pedig diszkrét), egy folytonossági korrekciót kell végrehajtanunk és a diszkrét egész *x-et* a

$$x - 0.5 \text{ -tól } x + 0.5 \text{ –ig}$$

intervallummal kell helyettesíteni

(hozzá kell adni és levonni 0.5-öt).

A folytonossági korrekció menete

1. **Ha a binomiálist normálissal közelíted, mindig használd a folytonossági korrekciót.**

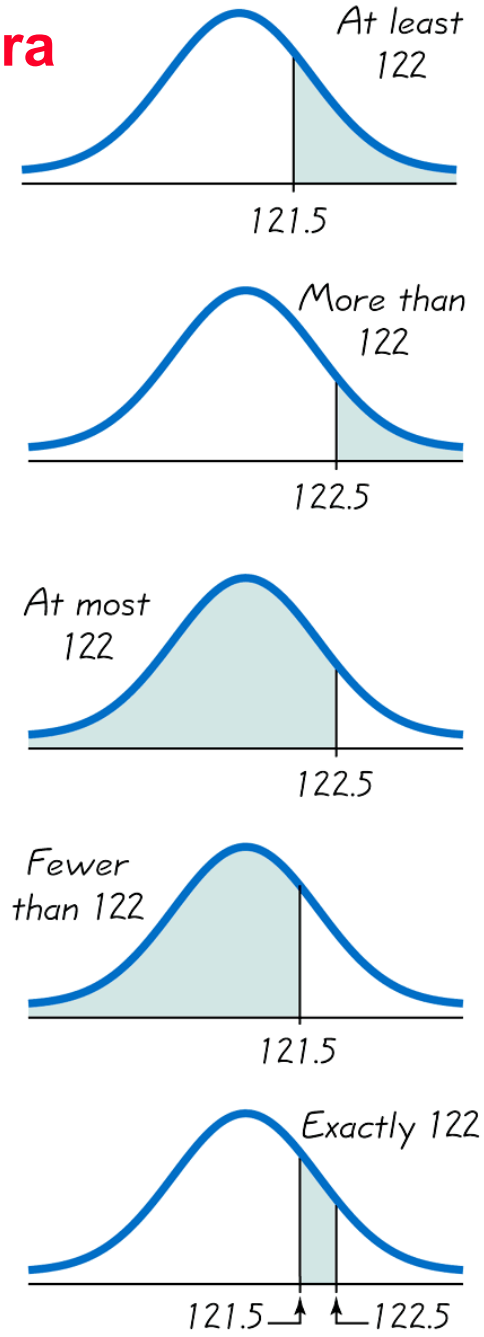
2. **Először keresd meg a diszkrét egész *x-et* a binomiális problémánál.**

3. **Rajzolj egy normális eloszlást, μ átlag köré, és rajzolj egy függőleges *x*-re centrált sávot *x* – 0.5 és *x* + 0.5 határokkal. Példánkban *x* = 122, rajzoljunk be egy sávot 121.5-nél és 122.5-nél. A berajzolt terület reprezentálja a diszkrét egész *x* érték valószínűségét.**

- folyt.

4. **Aztán gondold meg, hogy *x* maga benne van-e abban a valószínűségben, amit ki akarsz számítani. Utána gondold meg, hogy a „legalább *x*”, „legfeljebb *x*”, „több mint *x*”, „kevesebb mint *x*”, vagy „pontosan *x*” valószínűségére van-e szükséged. Satírozd be a sávtól balra vagy jobbra eső területet és a sávot magát is akkor, és csak akkor ha *x* maga is benne van. A teljes besatírozott terület adja a keresett valószínűséget, amit keresünk.**

6-22. ábra



X = legalább 122
(tartalmazza 122-t és felette)

X = több mint 122
(nincs benne a 122)

X = legfeljebb 122
(tartalmazza 122-t és alatta)

X = kevesebb mint 122
(nem tartalmazza 122-t)

X = pontosan 122

Összefoglalás

Ebben a fejezetben megvitattuk:

- **A binomiális normálissal való közelítését.**
- **A normális közelítés procedúráját.**
- **A folytonossági korrekciókat.**

6-7. fejezet
A normalitás vizsgálata

Kulcsfogalmak

Ebben a fejezetben meghatározzuk hogy valamilyen eloszlás mikor tekinthető normálisnak.

A kritériumok a hisztogram vizuális megfigyelése és a haranggörbével való összehasonlításától az outlierek azonosításan keresztül a normális kvantilis-kvantilis plot bevezetéséig fognak terjedni.

Definíció

- **Normál QQ plot (vagy normál valószínűség plot) egy pontokból (*x*,*y*) álló gráf, ahol az *x* érték az eredeti minta adatokból áll és az *y* érték a megfelelő *z* érték, ami a standard normális eloszlásból származó kvantilis érték.**

Módszerek az adatok normalitásának vizsgálatára

1. **Hisztogram:** Készíts hisztogramot. Ha eltér a haranggörbétől, akkor vesd el a normalitást.

2. **Outlierek:** Keresd meg az outliereket. Ha több mint egyet találsz, vesd el a normalitást.

3. **Normál QQ plot:** Ha a hisztogram alapvetően szimmetrikus, és legfeljebb egy outlier van, készítsd el a normál QQ plotot az alábbi módon:

- folyt

3. Normál QQ plot

a. Rendezd sorba az adatokat a legkisebbtől a legnagyobbik irányában.

b. A *n* elemű minta esetén, minden érték a minta 1/*n*-ed részét jelenti. Használva az *n értékét*, határozzuk meg az 1/2*n*, 3/2*n*, 5/2*n*, 7/2*n*, … területeket. Ezek lesznek a megfelelő minta értéktől balra esés valószínűségei.

c. Felhasználva a standard normális eloszlást (táblázat , szoftver vagy kalkulátor) számítsuk ki a fenti területekhez tartozó z értékeket.

- folyt

d. Párosítsd a kiszámított z értékeket az x értékekkel, majd készítsd el az (*x*, *y*) grafikont, ahol *x* az eredeti adatok és *y* a megfelelő *z* érték.

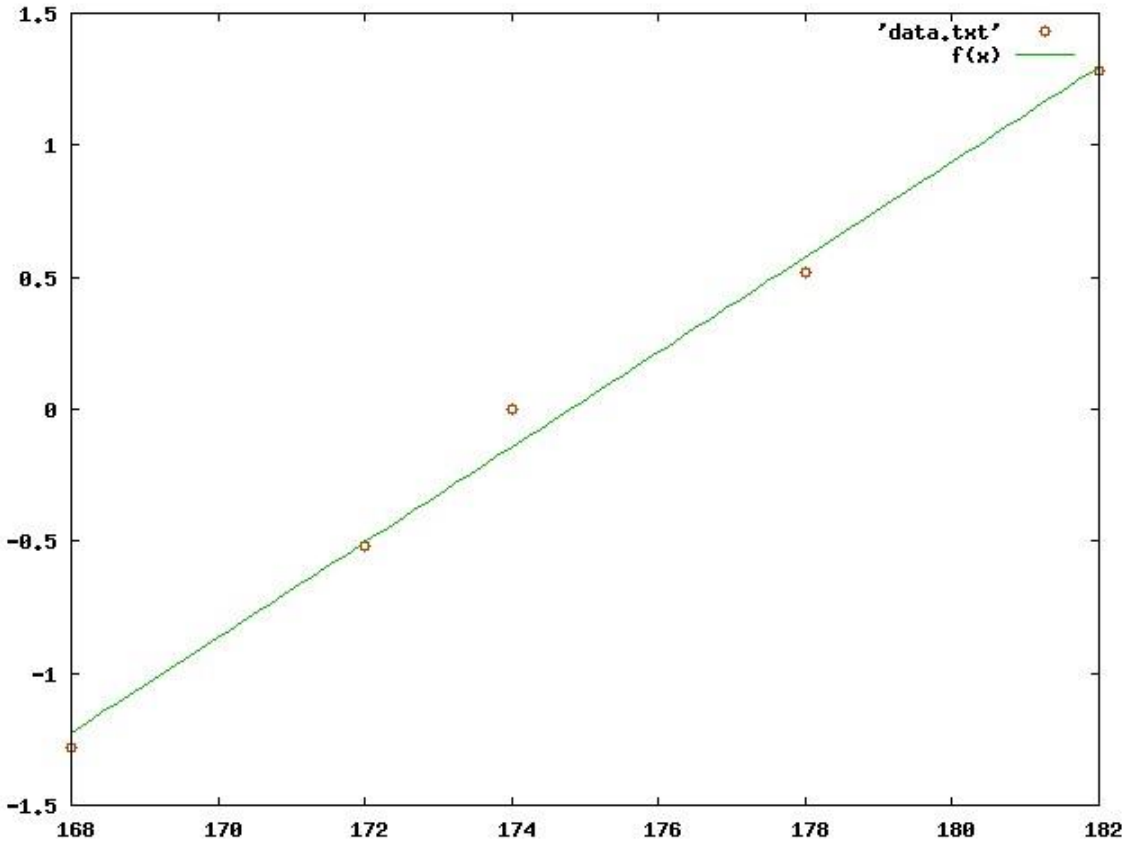
e. Vizsgáld meg az így készített QQ plotot az alábbi kritériumok alapján:

Ha az adatok nem fekszenek egy egyenesen, vagy valamilyen szisztematikus, de nem egyenes alakzatot öltenek, akkor az adatok nem normális eloszlással rendelkező populációból származnak. Ha az adatok elfogadhatóan közel vannak egy egyeneshez, akkor a populáció normálisnak tűnik.

Példa

- Vegyünk emberek magasságának adatait
- Elég pl. 5-öt 178, 168, 182, 172, 174
- n=5 minden adat 1/5-öde a teljesnek
- területek: 0.1,0.3,0.5,0.7 és 0.9
- z= -1.28, -0.52, 0, 0.52 és 1.28
- (x,y)= (168, -1.28) (172, -0.52) (174,0) (178,0.52) (182, 1.28)

Példa



Interpretáció: Mivel a pontok elfogadhatóan közel vannak egy egyeneshez és nem látszik bennük semmilyen más szisztematikus eltérés, arra következtetünk, hogy az eredeti adatok egy normális populációból származnak.

Összefoglalás

Ebben a fejezetben megvitattuk:

- **A normál QQ plotot.**
- **Azt a procedúrát, amivel eldönthetjük, hogy az adatok normális eloszlásúak-e.**

7. előadás
Becslések és minta elemszámok

7-1 Áttekintés

7-2 A populáció arány becslése

7-3 A populáció átlag becslése: σ ismert

7-4 A populáció átlag becslése: σ nem ismert

7-5 A populáció varianciájának becslése

7-1. fejezet
Áttekintés

Áttenkintés

Ebben a fejezetben elkezdjük a következtető (induktív) statisztika tárgyalását.

- **A következtető statisztika két legfontosabb alkalmazása, amikor a minta adatokat arra használjuk hogy (1) megbecsüljük a populáció valamelyik paraméterének értékét, illetve hogy (2) teszteljünk valamilyen a populációra vonatkozó állítást (hipotézist).**
- **Módszereket mutatunk be a populáció legfontosabb paramétereinek becslésére: arány, átlag és variancia.**
- **Meghatározzuk azokat a minta**

7-2. fejezet
A populáció arány becslése

Kulcsfogalmak

Ebben a fejezetben bemutatjuk, hogy a populáció arányt hogyan becsülhetjük a minta arányból, és hogyan adhatjuk meg a konfidencia intervallumot. Bemutatjuk azt is, hogy a becsléshez mekkora minta elemszám szükséges.

A populáció arány becslésének feltételei

1. **A minta egy egyszerű véletlen minta.**

2. **A binomiális eloszlás feltételei fennállnak.**

3. **Van legalább 5 sikeres és 5 sikertelen eset (a binomiálisnál bevezetett értelemben).**

Jelölések

p = **populáció arány**

$\hat{p} = \frac{x}{n}$ **minta arány**

↑ **(kimondva ‘p-kalap’)**

az *x* sikernek egy *n* elemű mintában

$\hat{q} = 1 - \hat{p}$ = **minta arány**

a sikertelen eseteknek egy *n*
elemű mintában

Definíció

Egy pontbecslés egy számérték (vagy pont), amivel a populáció paraméter értékét becsüljük.

Definíció

A minta arány $\hat{p}$ a legjobb pontbecslése a populáció aránynak p.

Példa:

Energia átadás kézzel (Emily Rosa, 9 éves, „A close look at the therapeutic touch”, Journal of the American Medical Association, Vol. 279, No. 13)

21 terapeuta, 280 kísérlet, 123 siker. Általában egy terapeuta milyen arányban találja el a helyes kezet?

Mivel a minta arány a legjobb pontbecslés a populáció arányra, ezért a legjobb pontbecslésünk p=123/280=0.44 .

Definíció

A konfidencia intervallum (vagy intervallumbecslés) egy tartománya (vagy intervalluma) az értékeknek, amivel a populáció paraméterének értékét becsüljük. (KI-vel rövidítjük néha.)

Definíció

A konfidencia szintje az az 1- α valószínűség (gyakran százalékban megadva), ami megadja, azon esetek arányát, ahányszor a konfidencia intervallum valójában tartalmazza a populáció paraméter értékét, ha a becslést sokszor megismételjük. (A konfidencia szintet a megbízhatóság fokának vagy szintjének is nevezzük.)

A leggyakoribb értékek 90%, 95% és 99%.

(α = 10%), (α = 5%), (α = 1%)

Példa: Adjuk meg az előző példánál azt a 95%-os konfidencia intervallumot, amibe a populáció arány beleesik.

“ 95%-ban biztosak vagyunk abban, hogy a 0.381 től 0.497-ig intervallum tartalmazza a *p* igazi értékét.”

Ez azt jelenti, hogy ha sok különböző 280 elemű mintát választanánk, és megkonstruálnánk hozzájuk a konfidencia intervallumokat, akkor 95%-uk tartalmazná a *p* igazi értékét.

Kritikus érték

1. **Tudjuk, hogy bizonyos feltételek mellett (központi határeloszlás tétel) az arány minta eloszlását normális eloszlással lehet közelíteni, mint ahogy azt a következő 7-2. ábrán látjuk.**

2. **A minta aránynak kicsi az esélye arra, hogy a 7-2. ábrán a piros részbe essen.**

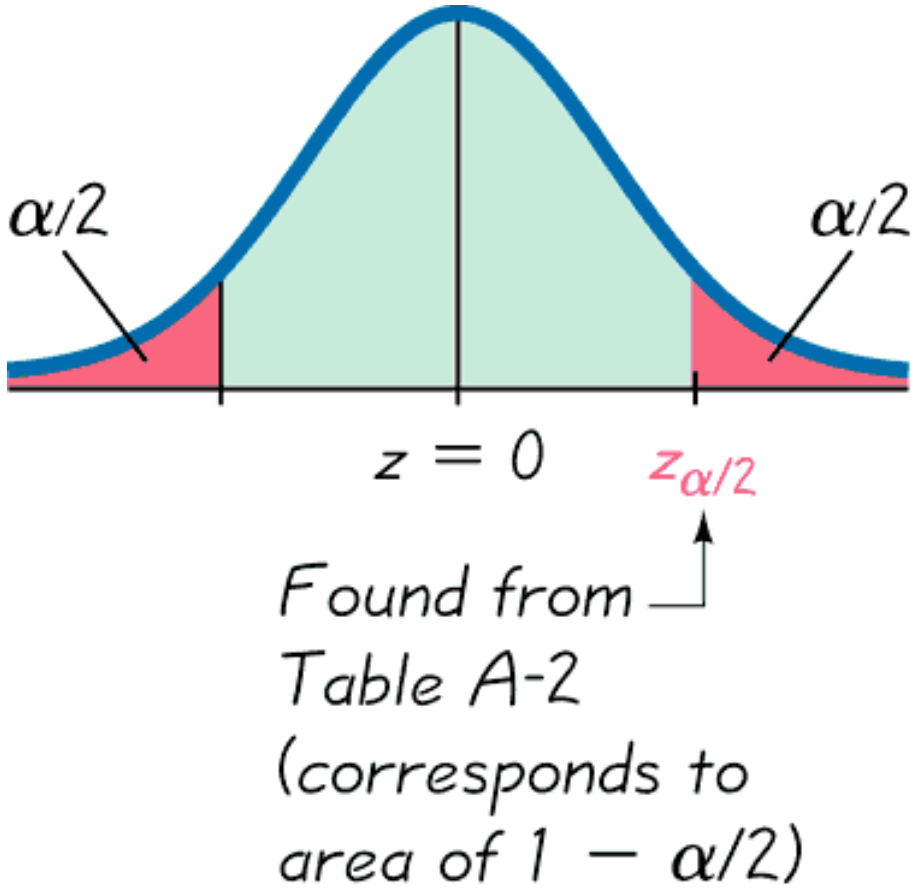
3. **Annak a valószínűsége, hogy bármelyik farok részbe esik a minta arány, összesen α.**

Kritikus érték

4. **Annak a valószínűsége, hogy a minta arány a zöld, belső részére esik 1-α a 7-2. ábrán.**

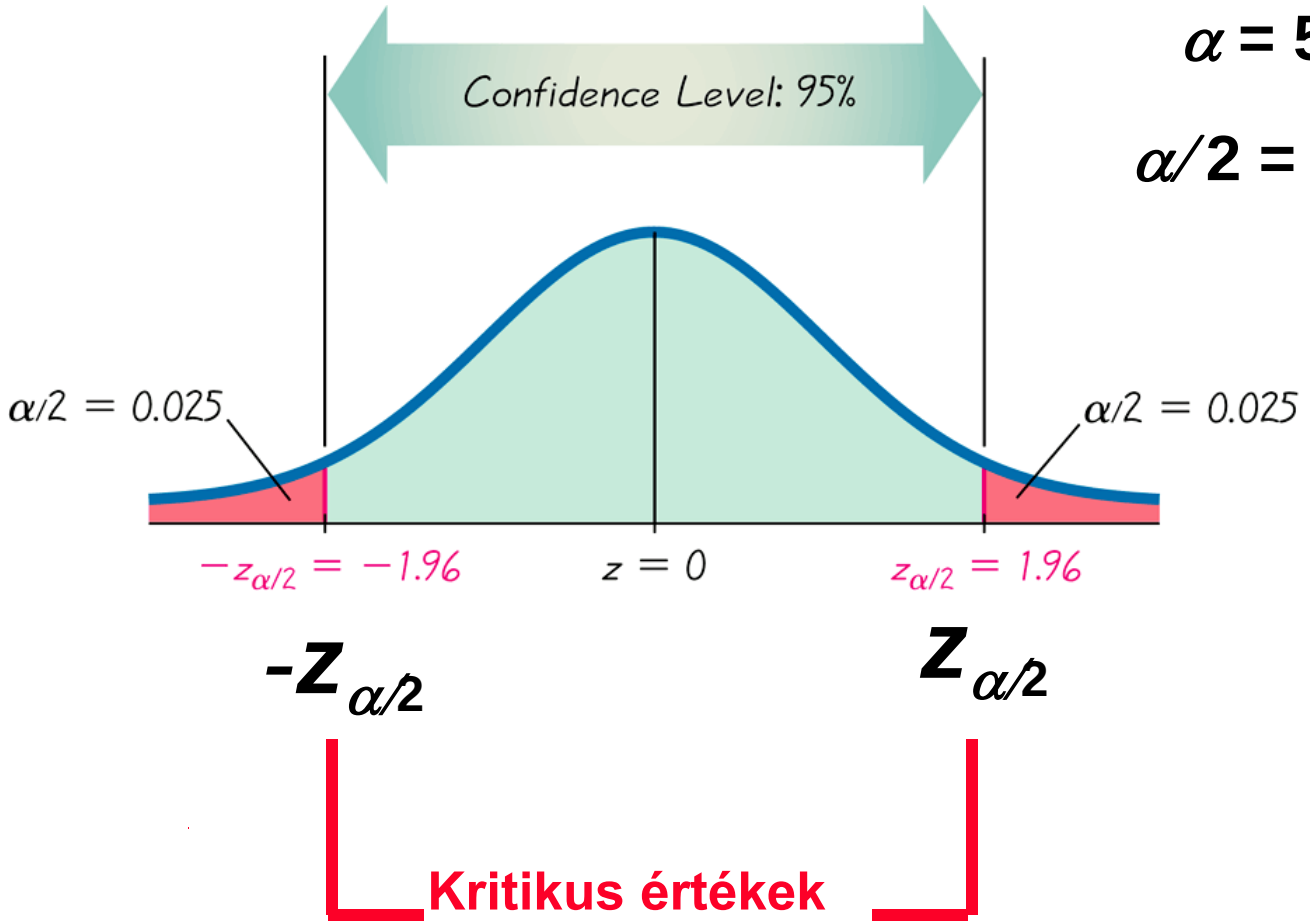
5. **Azt a *z* értéket, ami elválasztja a jobb farok részt $z_{\alpha/2}$-val jelöljük és kritikus értéknek nevezzük, mivel azon a határon van, ami elválasztja a valószínű és a nemvalószínű értékeket.**

Kritikus érték $z_{\alpha/2}$



7-2. ábra

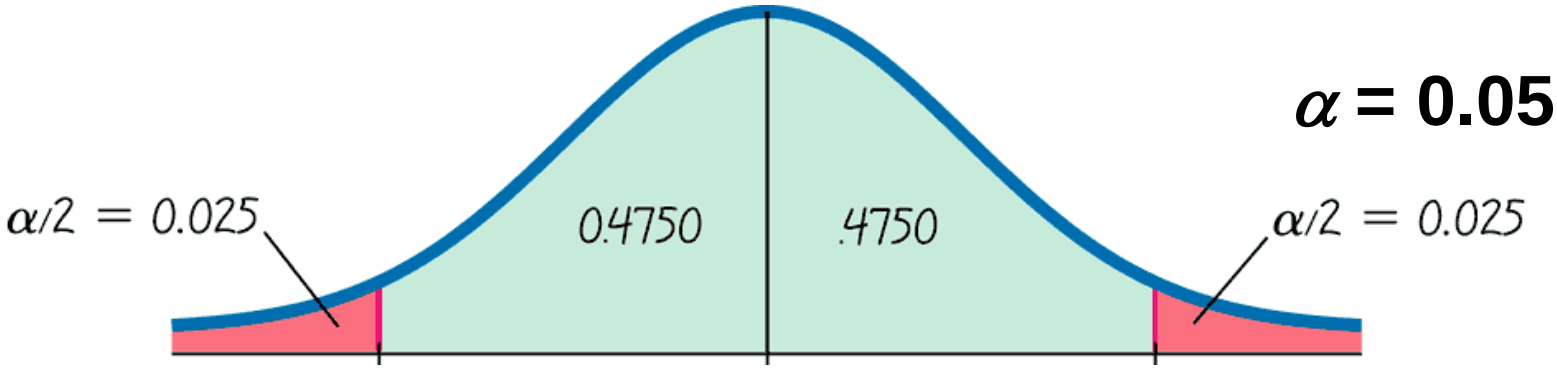
A $z_{\alpha/2}$ meghatározása a 95%-os konfidencia szinthez



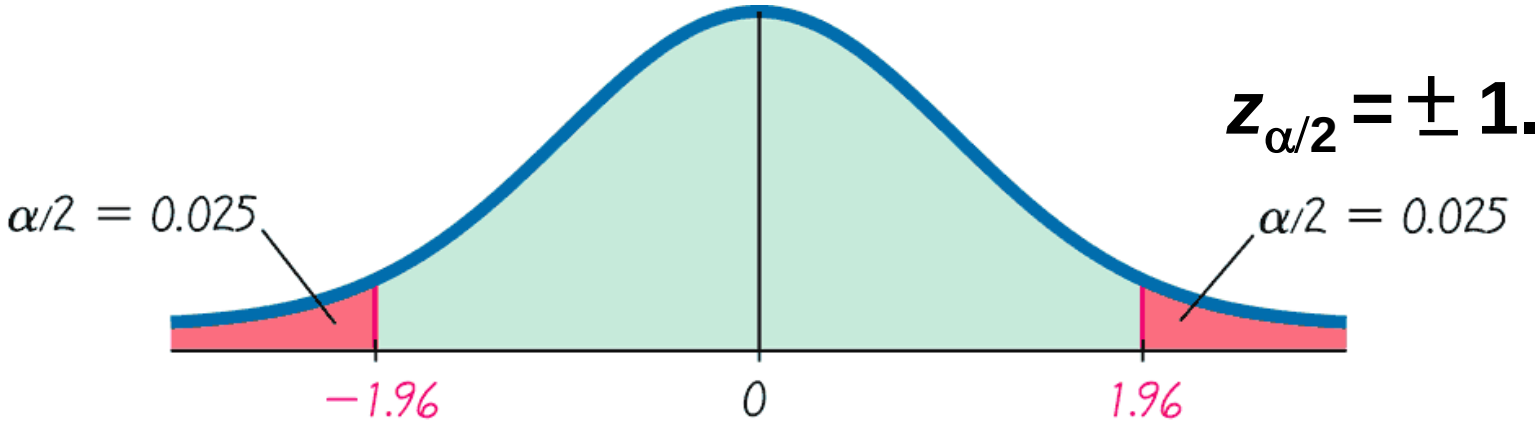
$\alpha = 5\%$

$\alpha/2 = 2.5\% = .025$

A $z_{\alpha/2}$ meghatározása a 95%-os konfidencia szinthez - folyt



$\alpha = 0.05$



$z_{\alpha/2} = \pm 1.96$

Néhány fontosabb kritikus érték

Konfidencia szint	α	Kritikus érték $z_{\alpha/2}$
90%	0.1	1.645
95%	0.05	1.96
99%	0.01	2.575

Definíció

Amikor egy egyszerű véletlen mintából becsüljük a populáció arányt (*p*-t), a hiba, amit *E*-vel jelölünk, a maximális eltérés (1 – α valószínűséggel) a megfigyelt *p* arány és az igazi populációs arány (*p) között*. A hibát (*E*-t) a becslés maximális hibájának is nevezik. Értékét a kritikus érték és az arány szórásának szorzataként kapjuk a következő 7-1. képlet szerint.

A *p* becslésének hibája

7-1. képlet

$$E = z_{\alpha/2}\sqrt{\frac{\hat{p}\,\hat{q}}{n}}$$

A populáció arány konfidencia intervalluma

$$\hat{p} - E < p < \hat{p} + E \ ,$$

ahol

$$E = z_{\alpha/2} \sqrt{\frac{\hat{p}\,\hat{q}}{n}}$$

A populáció arány konfidencia intervalluma

$$\hat{p} - E < p < \hat{p} + E$$

$$\hat{p} \pm E$$

$$(\hat{p} - E, \hat{p} + E)$$

A *p-re* vonatkozó konfidencia intervallum megkonstruálása

1. **Ellenőrizd, hogy a szükséges feltevések teljesülnek-e. (A minta egyszerű véletlen mintavételezésű, a binomiális feltételei fennállnak, a normális eloszlás használható a minta arányra, mivel $np \geq 5$ és $nq \geq 5$ is fennáll.)**

2. **A normális eloszlás táblázata segítségével határozzuk meg a $z_{\alpha/2}$ kritikus értéket.**

3. **Számítsd ki a hibát $E = z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$**

A *p-re* vonatkozó konfidencia intervallum megkonstruálása-folyt

4. **Felhasználva a hiba E értékét és a minta arányt $\hat{p}$, határozd meg $\hat{p} - E$ és $\hat{p} + E$ értékeit. Helyettesítsd be őket az általános konfidencia intervallum képletbe:**

$$\hat{p} - E < p < \hat{p} + E$$

Példa: ugyanaz

a) Keresd meg az E hibát 95%-os konfidencia szintnél.

Ellenőrizzük a feltételeket. $n\hat{p} = 123 \geq 5$, és $n\hat{q} = 157 \geq 5$.

Aztán kiszámítjuk. Azt találtuk, hogy $\hat{p} = 0.44$, $\hat{q} = 1 - 0.44 = 0.56$, $z_{\alpha/2} = 1.96$, és $n = 280$.

$$E = 1.96\sqrt{\frac{(0.44)(0.56)}{280}}$$

$$E = 0.058$$

Példa: ugyanaz

b) Határozzuk meg a 95%-os konfidencia intervallumot a populáció arányra *p*.

Behelyettesítve az előző értékeket:

$$0.439 - 0.058 < p < 0.439 + 0.058,$$

$$0.381 < p < 0.497$$

Példa: ugyanaz

c) Ennek alapján mit mondhatunk a módszer hatásosságáról?

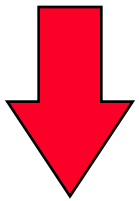
A kísérlet alapján 95%-os biztonsággal mondhatjuk, hogy a 38.1% és a 49.7% közti intervallum tartalmazza azt az arányt, ami esetén az energiaátvitelt a terapeuták érzékelik. Ez rosszabb, mint amit a véletlen próbálgatással (50%) kapnánk.

Minta elemszám

Tegyük fel, hogy adatokat gyűjtünk annak érdekében, hogy a populáció valamilyen tulajdonságát meghatározzuk. Kérdés, hogy hány mintát kell ehhez összegyűjteni?

A minta elemszám meghatározása

$$E = z_{\alpha/2}\sqrt{\frac{\hat{p}\,\hat{q}}{n}}$$



(oldjuk meg *n*-re)

$$n = \frac{(z_{\alpha/2})^2\,\hat{p}\,\hat{q}}{E^2}$$

Az *p* arány meghatározásához szükséges mintaszám

Ha van előzetes becslés $\hat{p}$-re :

$$n = \frac{(z_{\alpha/2})^2 \, \hat{p}\,\hat{q}}{E^2}$$

7-2. képlet

Ha nincs előzetes becslés $\hat{p}$-re:

$$n = \frac{(z_{\alpha/2})^2 \, 0.25}{E^2}$$

7-3. képlet

Example: Meg akarjuk határozni, hogy hány háztartásnak van Internet hozzáférése Magyarországon. Hány háztartást kell megkérdezni, ha 95%-os biztonsággal 4%-nál kisebb hibával akarjuk ezt meghatározni?

a) Korábbi eredmény felhasználása: 2004 decemberében, a háztartások 17%-ban volt Internet hozzáférés.

$$n = \frac{[z_{\alpha/2}]^2 \hat{p}\hat{q}}{E^2}$$

$$= \frac{[1.96]^2 (0.17)(0.83)}{0.04^2}$$

= 338 háztartás

Ha 95%-os biztonsággal igaz lesz, hogy a 338 háztartás megkérdezésével keletkező arány a valódi aránytól nem tér el jobban mint 4%.

Pontbecslés készítése a konfidencia intervallumból

A $\hat{p}$ pontbecslése:

$$\hat{p} = \frac{\text{(felső határ)} + \text{(alsó határ)}}{2}$$

Hiba:

$$E = \frac{\text{(felső határ)} - \text{(alsó határ)}}{2}$$

Összefoglalás

Ebben a fejezetben megvitattuk:

- **Pontbecslést.**
- **Konfidencia intervallumot.**
- **Konfidencia szintet.**
- **Kritikus érték.**
- **Hiba.**
- **Minta elemszám meghatározása.**

7-3. fejezet
Populáció átlag becslés:
σ ismert

Kulcsfogalmak

Ebben a fejezetben a populáció átlag pontbecslésére és konfidencia intervallumának meghatározása adunk módszert. Ebben a fejezetben feltesszük, hogy a populáció szórása ismert. (Ez a feltétel nem valószerű!)

Feltevések

1. A minta egyszerű véletlen mintavételezéssel lett kiválasztva. (Minden ugyanolyan hosszúságú minta kiválasztásának egyenlő az esélye.)

2. A populáció σ szórása ismert.

3. Egyik vagy mindkét alábbi feltétel igaz: A populáció normális eloszlású vagy $n > 30$.

A populáció átlag pontbecslése

A minta átlag $\bar{x}$ a populáció átlag μ legjobb pontbecslése.

Minta átlag

1. **Minden populáció esetén a minta átlag $\bar{x}$ torzítatlan becslése a populáció átlagnak μ, ami azt jelenti, hogy a μ populáció átlag körül csoportosul a minta átlagok eloszlása különböző minták esetén.**

2. **Sok populáció esetén a minta átlag $\bar{x}$ konzisztensebb (kisebb a változékonysága) mint más minta statisztikáknak.**

Példa: Egy vizsgálatban megvizsgálták 106 felnőtt testhőmérsékletét. A minta átlag 36.77 fok a szórás 0.34 fok volt. Keresd meg a populáció átlag μ legjobb pontbecslését!

Mivel a minta átlag $\bar{x}$ a legjobb pontbecslése a populáció átlagnak μ, ezért a legjobb pontbecslés 36.77° C.

Definíció

A hiba a minta átlag $\bar{x}$ és a populáció átlag μ valószínű eltéréseinek maximuma és $\bar{E}$-vel jelöljük.

Képlet

Hiba

$$E = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

7-4. képlet

Az átlag hibája (ismert *σ*-*t* feltételezve)

A μ populáció átlag konfidencia intervalluma (ismert σ szórás esetén)

$$\bar{x} - E < \mu < \bar{x} + E$$

vagy

$$\bar{x} \pm E$$

vagy

$$(\bar{x} - E, \bar{x} + E)$$

Definíció

Az $x - E$ és $x + E$ értékeket konfidencia intervallum határoknak hívjuk.

A μ konfidencia intervallumának megkonstruálása (ismert σ)

1. **Ellenőrizd, hogy a feltételek teljesülnek-e.**
2. **A normális eloszlás táblázatából határozd meg a $z_{\alpha/2}$ kritikus értéket.**
3. **Számítsd ki a hibát $E = z_{\alpha/2} \cdot \sigma / \sqrt{n}$.**
4. **Keresd meg az $\bar{X} - E$ és $\bar{X} + E$ *értékeket*. Helyettesítsd be az általános képletbe:**

$$\bar{x} - E < \mu < \bar{x} + E$$

Példa: ugyanaz. Keressük meg a hibát *E* és a 95%-os konfidencia intervallumot a *μ-re*.

$n = 106$

$\bar{x} = 36.77^{o}$

$s = 0.34^{o}$

$$E = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} = 1.96 \cdot \frac{0.34}{\sqrt{106}} = 0.064$$

$$\bar{x} - E < \mu < \bar{x} + E$$

$\alpha = 0.05$

$\alpha/2 = 0.025$

$z_{\alpha/2} = 1.96$

$$36.70^{o} < \mu < 36.83^{o}$$

$$36.77^{o} - 0.064 < \mu < 36.77^{o} + 0.064$$

A μ populációs átlag meghatározásához szükséges minta elemszám

$$n = \left[\frac{(z_{\alpha/2}) \cdot \sigma}{E} \right]^2$$

7-5. képlet

Ahol

$z_{\alpha/2}$ = a konfidencia szinthez tartozó kritikus z érték

E = megkívánt hiba

σ = a populáció szórása

Példa: Tegyük fel, hogy meg akarjuk határozni a fizika professzorok átlagos IQ értékét. Hány fizika professzort kell véletlenül kiválasztani a vizsgálatban ahhoz, hogy ha 95%-os biztonsággal és 2 IQ pont pontossággal akarjuk az értéket meghatározni? Tegyük fel, hogy $\sigma = 15$, ugyanúgy, mint az általános populációban.

$\alpha = 0.05$

$\alpha/2 = 0.025$

$z_{\alpha/2} = 1.96$

$E = 2$

$\sigma = 15$

$$n = \left[\frac{1.96 \bullet 15}{2}\right]^2 = 216.09 = 217$$

Egy 217 véletlen egyszerű mintavételezett fizika professzor IQ tesztjéből 95%-os biztonsággal 2 IQ pont hibával meg tudjuk határozni az igazi populáció átlagot, μ-t.

Összefoglalás

Ebben a fejezetben megbeszéltük a:

- **Hibát.**
- **Ismert σ esetén a konfidencia intervallumot.**
- **A μ meghatározásához szükséges minta elemszámot.**

7-4. fejezet
A populáció átlag becslése: σ nem ismert

Kulcsfogalmak

Ebben a fejezetben módszert adunk a konfidencia intervallum becslésére abban az esetben ha a populáció szórása nem ismert. Ha σ nem ismert, akkor a Student *t* eloszlást kell használnunk, bizonyos feltételek teljesülése esetén.

Feltevések σ ismeretlen esetben

1) A minta véletlen egyszerű.

2) A minta vagy normális populációból származik, vagy $n > 30$.

A Student *t* eloszlás

Ha a populáció eloszlása lényegében normális, akkor a következő mennyiség eloszlását

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

a Student *t* eloszlás adja meg *n* elemszámú minták esetén. Gyakran *t* eloszlásnak hívják és kritikus értékeit $t_{\alpha/2}$ jelöli.

Definíció

A szabadsági fokok számát egy minta adataira vonatkozóan azon adatok száma adja, amelyek szabadon változhatnak, miközben az adatok összességének valamilyen feltételnek eleget kell tenniük (ilyen pl. az hogy átlaguk legyen egy megadott érték).

szabadsági fokok száma = n – 1

ebben a fejezetben.

Kritikus t értékek táblázata

Conf. Level	50%	80%	90%	95%	98%	99%
One Tail	0.250	0.100	0.050	0.025	0.010	0.005
Two Tail	0.500	0.200	0.100	0.050	0.020	0.010
df	.	.	.	.	.	.
1	1.000	3.078	6.314	12.706	31.821	63.657
2	0.816	1.886	2.920	4.303	6.965	9.925
3	0.765	1.638	2.353	3.182	4.541	5.841
4	0.741	1.533	2.132	2.776	3.747	4.604
5	0.727	1.476	2.015	2.571	3.365	4.032
6	0.718	1.440	1.943	2.447	3.143	3.707
7	0.711	1.415	1.895	2.365	2.998	3.499
8	0.706	1.397	1.860	2.306	2.896	3.355
9	0.703	1.383	1.833	2.262	2.821	3.250
10	0.700	1.372	1.812	2.228	2.764	3.169
11	0.697	1.363	1.796	2.201	2.718	3.106
12	0.695	1.356	1.782	2.179	2.681	3.055
13	0.694	1.350	1.771	2.160	2.650	3.012
14	0.692	1.345	1.761	2.145	2.624	2.977
15	0.691	1.341	1.753	2.131	2.602	2.947
16	0.690	1.337	1.746	2.120	2.583	2.921
17	0.689	1.333	1.740	2.110	2.567	2.898
18	0.688	1.330	1.734	2.101	2.552	2.878
19	0.688	1.328	1.729	2.093	2.539	2.861
20	0.687	1.325	1.725	2.086	2.528	2.845
21	0.686	1.323	1.721	2.080	2.518	2.831
22	0.686	1.321	1.717	2.074	2.508	2.819
23	0.685	1.319	1.714	2.069	2.500	2.807
24	0.685	1.318	1.711	2.064	2.492	2.797

Az E hiba (σ nem ismert)

7-6. képlet

$$E = t_{\alpha/2} \frac{s}{\sqrt{n}}$$

ahol $t_{\alpha/2}$ n – 1 szabadsági fokkal rendelkezik

s a minta szórása

Konfidencia intervallum μ-re (σ nem ismert)

$$\bar{x} - E < \mu < \bar{x} + E$$

ahol $$E = t_{\alpha/2} \frac{s}{\sqrt{n}}$$

A μ konfidencia intervallumának megkonstruálása (σ ismeretlen)

1. **Ellenőrizzük, hogy a feltételek teljesülnek.**

2. **Az n - 1 szabadsági fokhoz keressük ki a Student eloszlás táblázatából a kritikus $t_{\alpha/2}$ értéket a kívánt konfidencia szinthez.**

3. **Számítsd ki a hibát $E = t_{\alpha/2} \bullet s / \sqrt{n}$.**

4. **Keresd meg az $\bar{x} - E$ és $\bar{x} + E$ értékeket. Helyettesítsük be a konfidencia intervallum általános képletébe:**

$$\bar{x} - E < \mu < \bar{x} + E$$

Példa: A testhőmérséklet példában határozzuk meg a μ 95%-os konfidencia intervallumát.

$n = 106$

$\bar{x} = 36.77^o$

$s = 0.34^o$

$\alpha = 0.05$

$\alpha/2 = 0.025$

$t_{\alpha/2} = 1.984$

$$E = t_{\alpha/2} \cdot \frac{s}{\sqrt{n}} = 1.984 \cdot \frac{0.34}{\sqrt{106}} = 0.065$$

$$\bar{x} - E < \mu < \bar{x} + E$$

$$36.70^o < \mu < 36.83^o$$

A Student *t* eloszlás tulajdonságai

1. **A Student *t* eloszlás más-más különböző minta elemszámokra.**

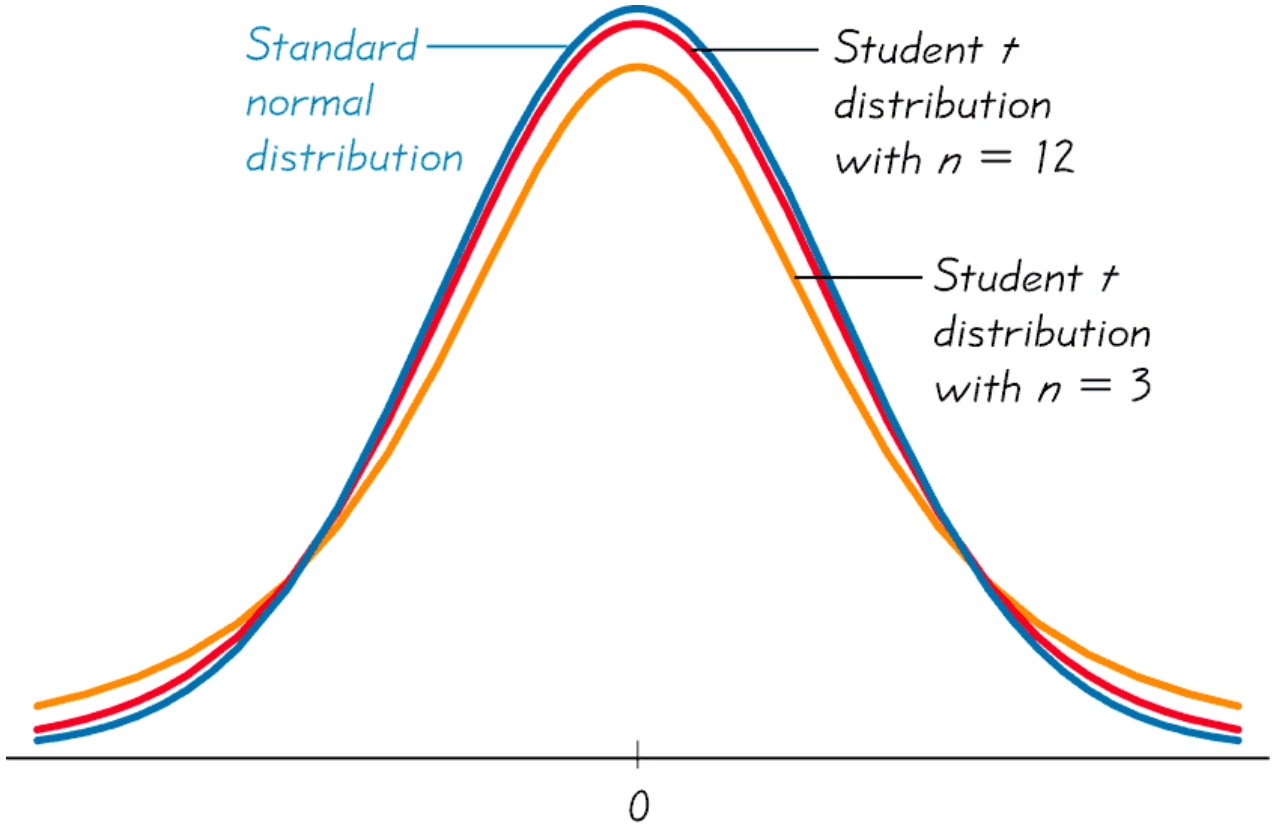
2. **A Student *t* eloszlás szimmetrikus és harang szerű görbe, de sokkal nagyobb variabilitása van, mint a normális eloszlásnak kis minta számok esetén.**

3. **A Student *t* eloszlás átlaga $t = 0$ (ugyanúgy, mint a standard normális eloszlás esetén az átlag $z = 0$).**

4. **A Student *t* eloszlás szórása változik a minta elemszámmal és nagyobb mint 1 (ellentétben a standard normális eloszlással, ahol $\sigma = 1$).**

5. **A minta elemszám növelésével *n* egyre nagyobb lesz, és a Student *t* eloszlás egyre közelebb kerül a normál**

Student *t* eloszlás
$n = 3$ és $n = 12$



7-5. ábra

Összefoglalás

Ebben a fejezetben tárgyaltuk:

- **A Student *t* eloszlást.**
- **A szabadsági fokok számát.**
- **A hibát.**
- **A μ konfidencia intervallumát ismeretlen σ esetén.**

7-5. fejezet
A populáció variancia becslése

Kulcsfogalmak

Ebben a fejezetben módszereket mutatunk be a (1) konfidencia intervallum meghatározására a populáció szórására és varianciájára (2) a szükséges minta elemszám meghatározására.

Bevezetjük a χ -négyzet (khí négyzet, chi-square) eloszlást, ami a konfidencia intervallum meghatározásához kell σ ill. σ^2 esetén.

Feltételek

1. A minta legyen egyszerű véletlen.

2. A populációnak normális eloszlásúnak kell lennie (nem elég, hogy a minta nagy legyen).

Khí-négyzet eloszlás

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

7-7. képlet

ahol

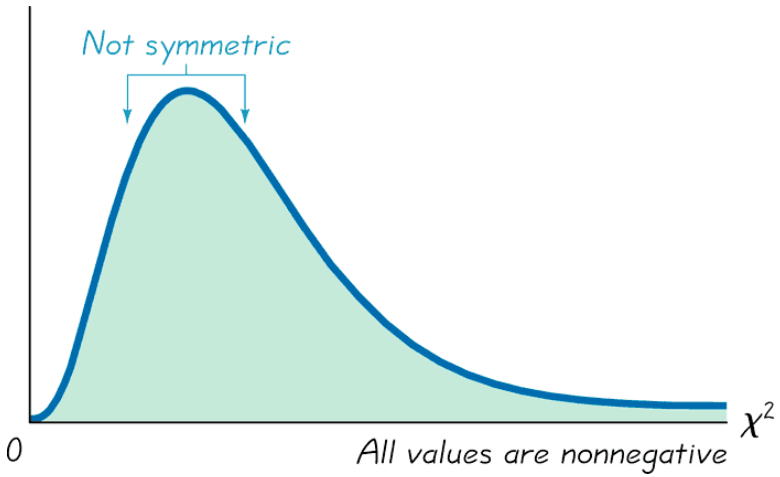
n = minta elemszám

s^2 = minta variancia

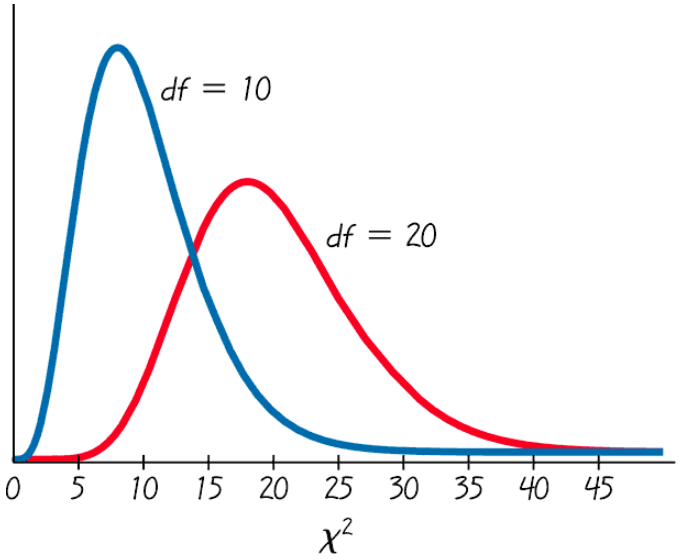
σ^2 = populáció variancia

A khi-négyzet statisztika tulajdonságai

1. **A khi-négyzet eloszlás nem szimmetrikus, ellentétben a normál és a Student eloszlásssal. A szabadsági fokok számának növekedésével egyre szimmetrikusabb lesz.**



7-8. ábra Khi-négyzet eloszlás



7-9. ábra Khi-négyzet eloszlás df = 10 és df = 20

Khi-négyzet táblázat

	Left Tail					Right Tail				
df	**0.005**	**0.01**	**0.025**	**0.05**	**0.10**	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**
1	---	---	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.121	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.336
30	13.787	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154	79.490
60	35.534	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425	104.215
80	51.172	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329	116.321
90	59.196	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116	128.299
100	67.328	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807	140.169

A khi-négyzet statisztika tulajdonságai- folyt

2. **A khi-négyzet eloszlás értékei nem lehetnek negatív számok.**

3. **A khi-négyzet eloszlás különbözik minden szabadsági fokra, amely df = n – 1 ebben a fejezetben. A szabadsági fokok növelésével megközelíti a normális eloszlást.**

Példa:

Határozzuk meg χ^2 kritikus értékeit, amelyekhez mindkét farokban 0.025 terület tartozik. Legyen a minta elemszáma 10, és a szabadsági fokok száma 10 – 1=9.

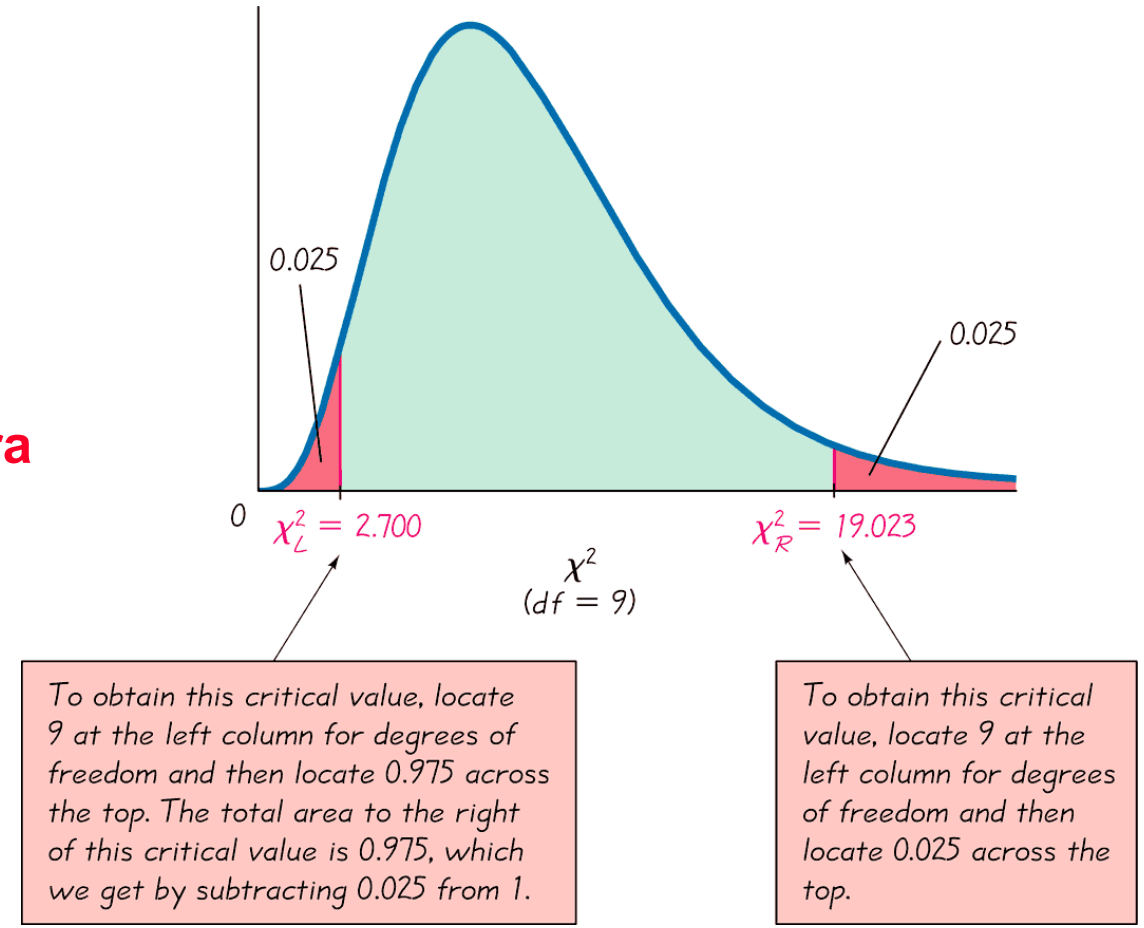
$$\alpha = 0.05$$

$$\alpha/2 = 0.025$$

$$1 - \alpha/2 = 0.975$$

A khi-négyzet statisztika kritikus értékei

7-10. ábra



A variancia becslései

A minta variancia s^2 a legjobb pontbecslése a populáció varianciájának σ^2.

Konfidencia intervallum (vagy intervallum becslés) a populáció varianciára σ^2

$$\frac{(n-1)s^2}{\chi^2_R} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_L}$$

Jobb-farok kritikus érték

Bal-farok kritikus érték

Konfidencia intervallum a σ-ra

$$\sqrt{\frac{(n-1)s^2}{\chi^2_R}} < \sigma < \sqrt{\frac{(n-1)s^2}{\chi^2_L}}$$

A σ vagy σ^2 –re vonatkozó konfidencia intervallum konstruálása

1. **Ellenőrizzük, hogy a feltételek fennállnak-e.**
2. **$n-1$ szabadsági fok esetén a táblázatból keressük meg a kritikus értékeket χ^2_R és χ^2_L, amely a kívánt konfidencia szinthez tartozik.**
3. **Az alábbi képlettel határozzuk meg a konfidencia intervallumot:**

$$\frac{(n-1)s^2}{\chi^2_R} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_L}$$

4. **σ konfidencia intervalluma ugyanez, csak gyököt kell vonni.**

Példa:

A testhőmérsékletes példában keressük meg a 95%-os konfidencia intervallumot σ-ra.

$n = 106$

$\bar{x} = 36.77^o$

$s = 0.34^o$

$\alpha = 0.05$

$\alpha/2 = 0.025$

$1 - \alpha/2 = 0.975$

$\chi^2_R = 129.561, \chi^2_L = 74.222$

$$\frac{(106-1)(0.34)^2}{129.561} < \sigma^2 < \frac{(106-1)(0.34)^2}{74.222}$$

$$0.093 < \sigma^2 < 0.16$$

$$0.30 < \sigma < 0.40$$

95%-ban bizonyosak vagyunk, hogy a 0.30°C és 0.40°C intervallum tartalmazza a σ igazi értékét. 95%-os biztonsággal állíthatjuk, hogy az egészséges emberek testhőmérsékletének szórása 0.30°C és 0.40°C között van.

A minta elemszám meghatározása

Table 7-2

Sample Size for σ^2		Sample Size for σ	
To be 95% confident that s^2 is within	of the value of σ^2, the sample size n should be at least	To be 95% confident that s is within	of the value of σ, the sample size n should be at least
1%	77,207	1%	19,204
5%	3,148	5%	767
10%	805	10%	191
20%	210	20%	47
30%	97	30%	20
40%	56	40%	11
50%	37	50%	7
To be 99% confident that s^2 is within	of the value of σ^2, the sample size n should be at least	To be 99% confident that s is within	of the value of σ, the sample size n should be at least
1%	133,448	1%	33,218
5%	5,457	5%	1,335
10%	1,401	10%	335
20%	368	20%	84
30%	171	30%	37
40%	100	40%	21
50%	67	50%	13

Példa:

Szeretnénk σ értékét meghatározni a testhőmérsékletekre. 95% biztonsággal szeretnénk tudni, legfeljebb 10% hibával a σ igazi értékét. Mekkorának kell lennie a mintának. Tegyük fel, hogy a populáció normális eloszlású.

A 7-2. táblázat szerint, 95% konfidenciával 10% hiba 191-es mintához tartozik.

Összefoglalás

Ebben a fejezetben megvitattuk:

- **A khi-négyzet eloszlást.**
- **A táblázatát.**
- **A szórás és a variancia konfidencia intervallumait.**
- **A minta elemszám meghatározását.**

8. előadás
Hipotézis tesztelés

8-1 Áttekintés

8-2 A hipotézis tesztelés alapjai

8-3 A populáció arányra vonatkozó feltevés tesztje

8-4 Az átlagra vonatkozó feltevés tesztje: σ ismert

8-5 Az átlagra vonatkozó feltevés tesztje: σ ismeretlen

8-6 A szórásra és a varianciára vonatkozó tesztek

8-1. fejezet
Áttekintés

Definíciók

A statisztikában, a hipotézis egy a populáció valamilyen tulajdonságára vonatkozó állítás/kijelentés.

A hipotézis teszt (vagy szignifikancia teszt) egy szabványos/bevett (standard) módszer arra, hogy próbának (tesztnek) vessük alá a populáció valamilyen tulajdonságára vonatkozó állítást (hipotézist).

A ritka esemény szabály a statisztikában

Ha, adott feltevések mellett egy bizonyos esemény valószínűsége kicsi, de mi mégis megfigyeljük egy ilyen esemény bekövetkezését, akkor arra a konklózióra jutunk, hogy a feltevés nem igaz.

kifejlesztette a "Gender Choice" nevű terméket, ami a cég hirdetései szerint a pároknak "85%-kkal növeli a fiú és 80%-kkal a lány születésének esélyét." A Gender Choice-nak kék és rózsaszín csomagolása volt, attól függően, hogy a vásárlói fiú vagy lány gyermeket szerettek volna. Tegyük fel, hogy kísérletet végzünk 100 párral, akik lány gyermeket akarnak és a rózsaszín Gender Choice "easy-to-use in-home system" terméket használják.

A tesztelés kedvéért mi azt állítjuk, hogy a Gender Choice hatástalan.

Pusztán a józan eszünkre hagyatkozva milyen konklúzióra jutnánk a saját, fenti állításunkról, ha a

Példa: ProCare Industries, Ltd.: a) rész

a) Általában kb. 50 lányt várunk 100 születésből. Az 52 közel van az 50-hez, így nem gondoljuk, hogy a „Gender Choice” hatásos. Ha a 100 pár nem használt volna semmilyen speciális módszert az 52 lány könnyen előfordulhatott volna véletlenül is. Az a feltevésünk, hogy a „Gender Choice” hatástalan korrektnek tűnik. Nincs elég bizonyíték arra, hogy a „Gender Choice” hatásos.

Példa: ProCare Industries, Ltd.: b) rész

b) A 97 lány 100 születésből nagyon ritkán történik meg véletlenül. Két magyarázatot adhatunk a 97 lány születésére: Vagy egy extrém ritka esemény következett be véletlenül, vagy a „Gender Choice” hatásos. A 97 lány születésének rendkívül kicsi valószínűsége egy erős bizonyíték azon feltevésünk ellen, hogy a „Gender Choice”hatástalan. Úgy tűnik, hogy hatásos.

8-2. fejezet
A hipotézis tesztelés alapjai

Kulcsfogalmak

Ebben a fejezetben a hipotézis vizsgálat elemi összetevőit mutatjuk be, amelyeket majd a további fejezetekben használunk fel. A következő fogalmakat kell megértenünk:

- **null hipotézis**
- **alternatív hipotézis**
- **teszt statisztika**
- **kritikus tartomány**
- **szignifikancia szint**
- **krtitikus érték**
- ***P*-érték**
- **Első és másodfajú hiba (Type I and II error)**

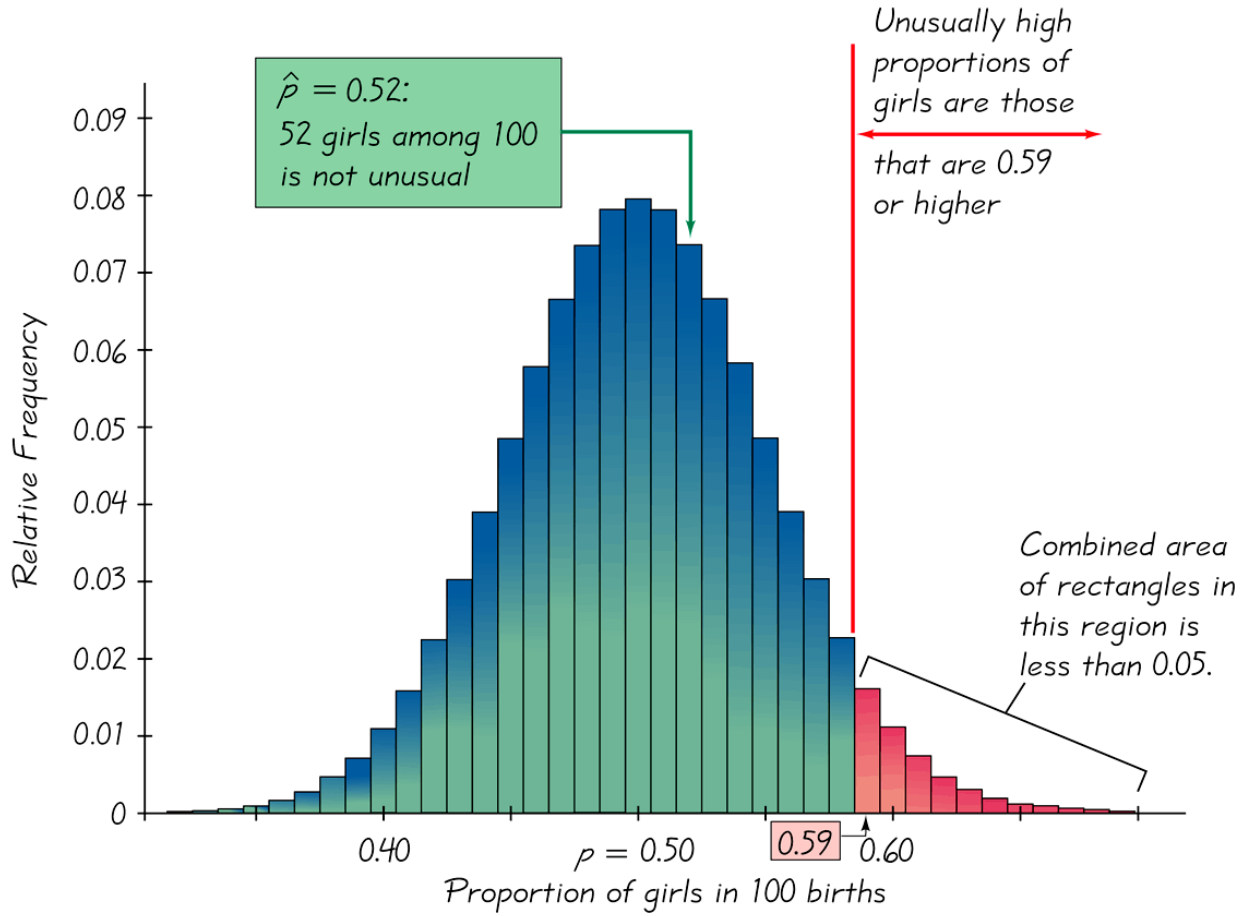
A 8-2. fejezet célkitűzései

- Adott állításhoz azonosítsuk a null hipotézist és az alternatív hipotézist és adjuk meg mindkettőt szimbolikus formában.
- Adott állításhoz és minta adatokhoz számítsuk ki a teszt statisztika értékét.
- Adott szignifikancia szint mellet határozzuk meg a kritikus értékeket.
- A teszt statisztika értékének ismeretében határozzuk meg a *P*-értéket.
- Adjuk meg a hipotézis teszt eredményét közérthető, nem technikai nyelven.

Példa: **Vegyük az előző példában az 52 lány születésének esetét.**

Normál körülmények között tekintsük 0.5-nek a lányok születésének valószínűségét. A „Gender Choice” hatásosságát ebben az esetben a $p > 0.5$ módon fejezhetjük ki (a populációbeli arány több mint 0.5).

A binomiális eloszlást normálissal közelítve kiszámíthatjuk, hogy
P**(52 vagy több lány 100 születésből) = 0.3821.**



8-1. ábra

Nem utasítjuk el a véletlent, mint elfogadható magyarázatot. Azt a konklúziót vonjuk le, hogy a „Gender Choice” által elért hatás nem szignifikánsan nagyobb, mint amit véletlenül is kaphatnánk.

Megfigyelések

- Állítás/feltételezés: A „Gender Choice”-t használóknál a lányok aránya $p > 0.5$.
- Munkafeltevés: A lányok aránya $p = 0.5$ (a „Gender Choice”).
- A minta eredmény 52 lány 100 születésből, a minta arány $\hat{p} = 52/100 = 0.52$.
- Feltéve, hogy $p = 0.5$, a normális eloszlást felhasználva azt kapjuk, hogy P (legalább 52 lány 100 születésből) = 0.3821.
- Két magyarázat van arra, hogy 52 lányt kapunk 100 születésből: Vagy véletlen esemény történt (0.3821 valószínűséggel), vagy a lányok születésének valószínűsége nagyobb mint 0.5 a „Gender Choice” hatására.
- Nincs elég bizonyíték a „Gender Choice” hatásosságának feltételezéséhez.

A formális hipotézis teszt összetevői

Null hipotézis: H_0

- A **null hipotézis** (jelölés: H_0) egy állítás a populáció valamilyen paraméter értékéről (mint arány, átlag vagy szórás) miszerint az **egyenlő** valamilyen feltételezett (hipotetikus) értékkel.
- A null hipotézist közvetlenül tesztelhetjük.
- Vagy elutasítjuk a H_0 hipotézist vagy nem tudjuk elutasítani a H_0 hipotézist.

Megjegyzés a mi saját feltevésünk (hipotézisünk) kialakításával kapcsolatban

Ha valamilyen vizsgálatot végzünk és a hipotézis tesztelést akarjuk használni a saját feltevésünk alátámasztására, akkor azt úgy kell megfogalmazni, hogy a saját feltevésünk legyen az alternatív hipotézis.

Alternatív hipotézis: H_1

- Az **alternatív hipotézis** (jelölés H_1 vagy H_a vagy H_A) egy állítás, ami szerint a paraméter értéke valamilyen módon különbözik a nulla hipotézistől.
- Az alternatív hipotézis szimbolikus kifejezése az alábbi szimbólumokat kell, hogy tartalmazza: ≠, <, >.

Megjegyzés H_0 és H_1 megválasztásának lépéseiről

1. Azonosítsd a tesztelendő hipotézist és írd le szimbolikusan.
2. Add meg azt a szimbolikus alakot, aminek akkor kell igaznak lennie, ha az eredeti hipotézis hamis.
3. A fenti kettő közül az legyen a null hipotézis, amiben = szerepel. Legyen az alternatív hipotézis, amiben $<$, $>$ vagy $\neq$ szerepel.

Példa: Azonosítsd a null és az alternatív hipotézist az előbbiek alapján!

a) Azon vezetők aránya, akik bevallják, hogy néha piros lámpán is áthaladnak nagyobb mint 0.5.

b) A profi kosarasok átlag magassága legfeljebb 213cm.

c) A színészek IQ értékeinek szórása 15.

Példa: folyt.

a) Azon vezetők aránya, akik bevallják, hogy átmennek a piroson nagyobb mint 0.5.

Az 1. lépésben: kifejezzük a feltevést szimbolikusan $p > 0.5$.

A 2. lépésben: látjuk, ha $p > 0.5$ hamis, akkor $p \leq 0.5$ lesz igaz.

A 3. lépésben: látjuk, hogy a $p > 0.5$ kifejezés nem tartalmaz egyenlőség jelet, ezért legyen az alternatív hipotézis (H_1) $p > 0.5$, és a null hipotézis (H_0) legyen $p = 0.5$.

Példa: folyt.

b) A profi kosarasok átlag magassága legfeljebb, 213cm.

Az 1. lépésben: szimbolikusan kifejezzük $\mu \leq 213$.

A 2. lépésben: látjuk, ha $\mu \leq 213$ hamis, akkor $\mu > 213$ igaz.

A 3. lépésben: látjuk, hogy $\mu > 213$ nem tartalmaz egyenlőség jelet, ezért ez lesz az alternatív hipotézis (H_1: $\mu > 213$), és H_0 lesz $\mu = 213$.

Példa: folyt.

c) A színészek IQ értékeinek szórása 15.
Az 1. lépésben: kifejezzük az állítást szimbolikusan
σ= 15.
A 2. lépésben: látjuk, hogy ha σ = 15 hamis, akkor
$\sigma \neq$ 15 igaz lesz.
A 3. lépésben: az alternatív hipotézis H_1 lesz $\sigma \neq$ 15,
és H_0 lesz σ = 15.

Teszt statisztika

A teszt statisztika egy olyan számérték, aminek segítségével döntést tudunk hozni a null hipotézisről. A minta statisztika értékéből képezzük annak a feltevésével, hogy a null hipotézis igaz.

Teszt statisztika - képletek

$$z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}}$$

Teszt statisztika az arányra

$$z = \frac{\bar{x} - \mu_{\bar{x}}}{\frac{\sigma}{\sqrt{n}}}$$

Teszt statisztika az átlagra

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

Teszt statisztika a varianciára

Feladat: Egy $n = 880$ véletlenül kiválasztott vezetőt megkérdezve 56%-uk (vagyis $p = 0.56$) mondta, hogy néha áthajt a piros jelzésen. Keressük meg a teszt statisztika értékét ahhoz a feltevéshez (hipotézishez), miszerint a vezetők többsége elismeri, hogy néha átmegy a prioson. (Majd a 8-3. fejezetben lesz néhány feltétel, aminek teljesülnie kell, most tegyük fel, hogy ezek rendben vannak.)

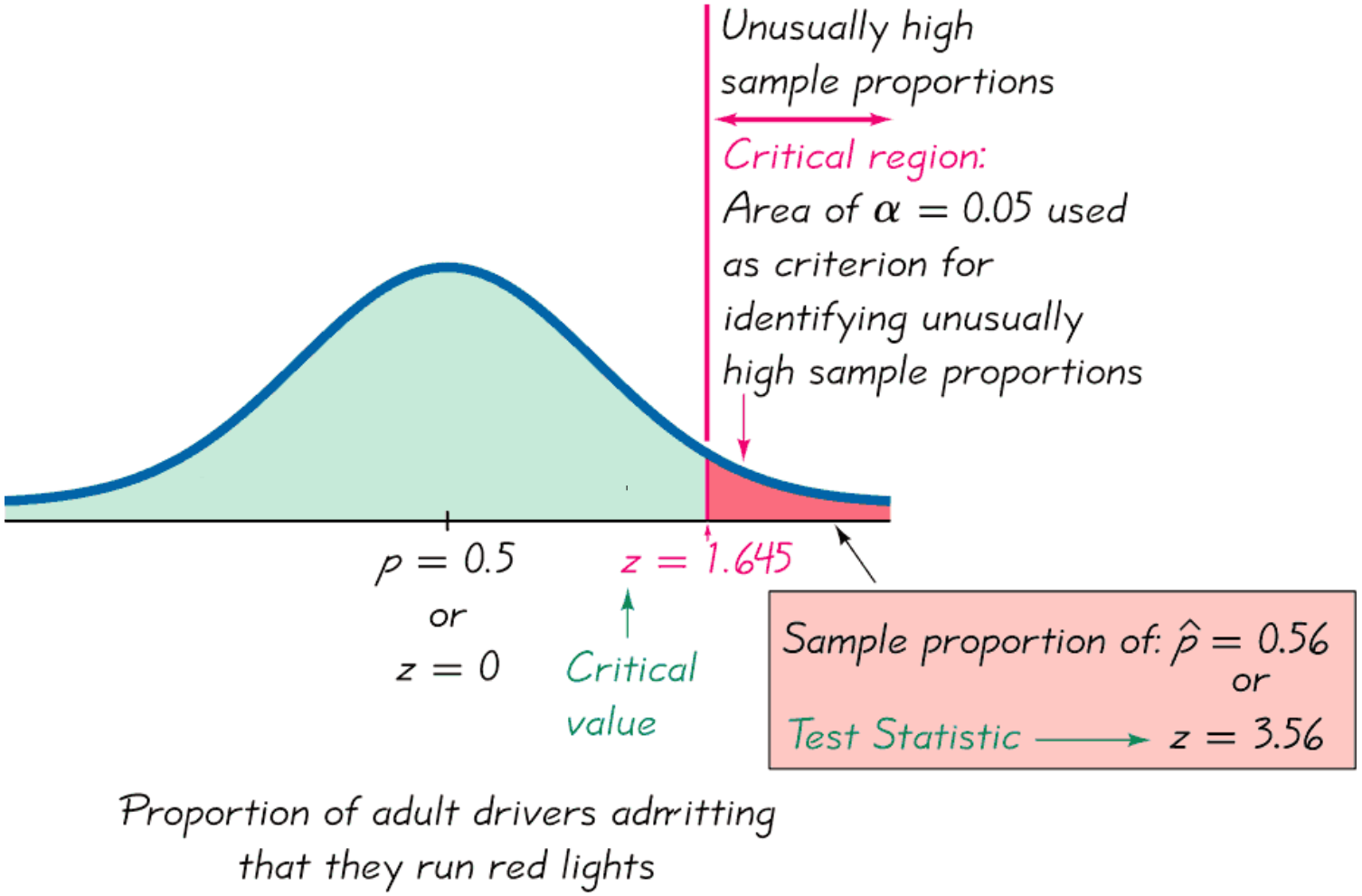
Megoldás: Az előző példában beláttuk, hogy ennek a feltevésnek az ellenőrzéséhez ah H_0: $p = 0.5$ null hipotézis és a H_1: $p > 0.5$ alternatív hipotézis tartozik. Mivel azzal a feltevéssel dolgozunk, hogy a null hipotézis igaz a $p = 0.5$ értékkel, a következő teszt statisztikát kapjuk:

$$z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}} = \frac{0.56 - 0.5}{\sqrt{\frac{(0.5)(0.5)}{880}}} = 3.56$$

Interpretáció: Tudjuk az előző fejezetekből, hogy a z=3.56 érték kivételesen nagy. Úgy tűnik, hogy azon túl, hogy az érték "több mint fél", a minta erdmény (56%) **szignifikánsan** több mint 50%.

Ld. a következő ábrát.

Kritikus tartomány, kritikus érték, teszt statisztika



Kritikus tartomány

A kritikus tartomány (vagy elutasítási tartomány) a teszt statisztika értékeinek az a tartománya, ami arra vezet, hogy a null hipotézist elutasítsuk. Példa rá az előző ábrán a pirosra színezett rész.

Szignifikancia szint

A szignifikancia szint (jelölés: α) az a valószínűség, amivel a teszt statisztika a kritikus tartományba esik, amikor a null hipotézis valójában igaz. Ez ugyanaz az α amit a 7-2. fejezetben vezettünk be. A szokásos választások α-ra: 0.05, 0.01 és 0.10.

Kritikus értékek

A kritikus értékek amik elválasztja a kritikus tartományt (ahol elutasítjuk a null hipotézist) azoktól az értékektől, ahol nem utasítjuk el. A kritikus értékek függnek a null hipotézis fajtájától, a minta eloszlástól és a szignifikancia szinttől. Ld. az előző ábrát, ahol a kritikus érték $z = 1.645$ az $\alpha = 0.05$ konfidencia szinthez tartozik.

Kétoldali, jobboldali és baloldali tesztek

Az eloszlás farkai/szélső tartományai az extrém tartományok, melyeket a kritikus értékek határolnak.

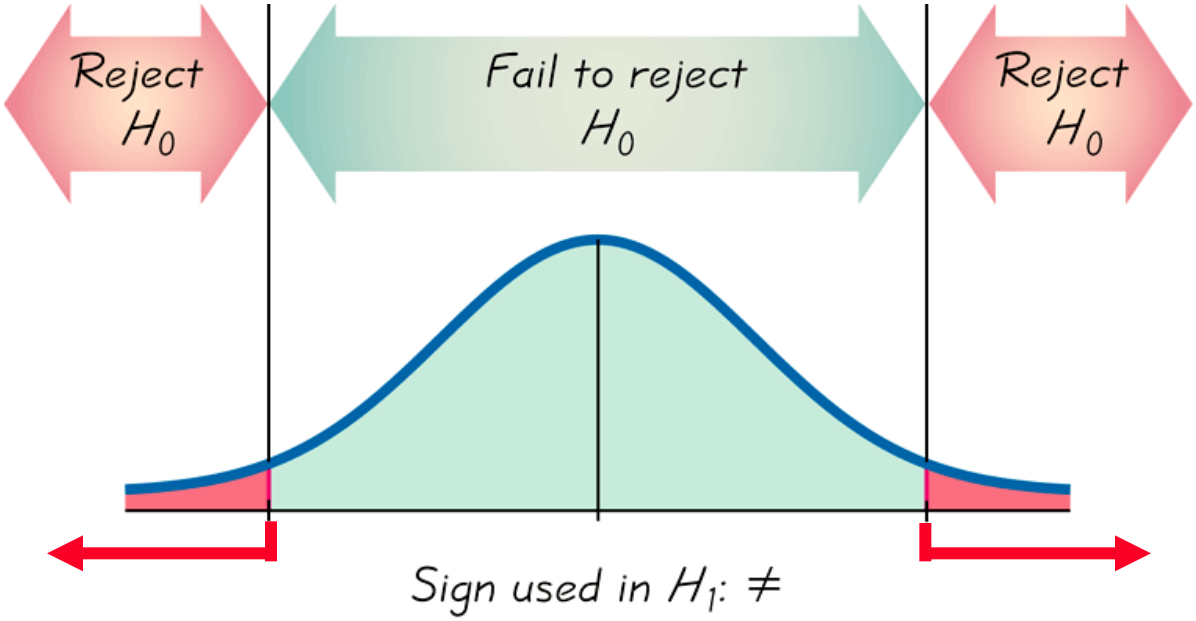
Kétoldali tesztek

H_0: =

α egyenlően van szétosztva a két farok között

H_1: ≠

Azt jelenti, kevesebb vagy több mint

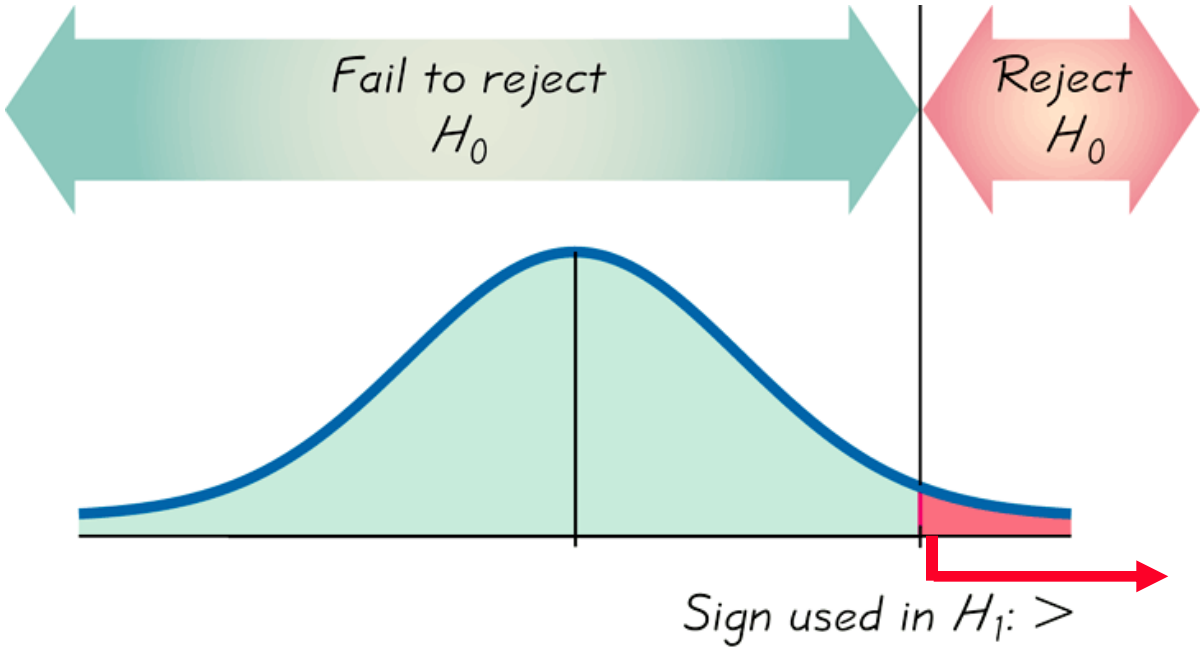


Jobboldali tesztek

H_0: =

H_1: >

Pontok jobbra vmitől

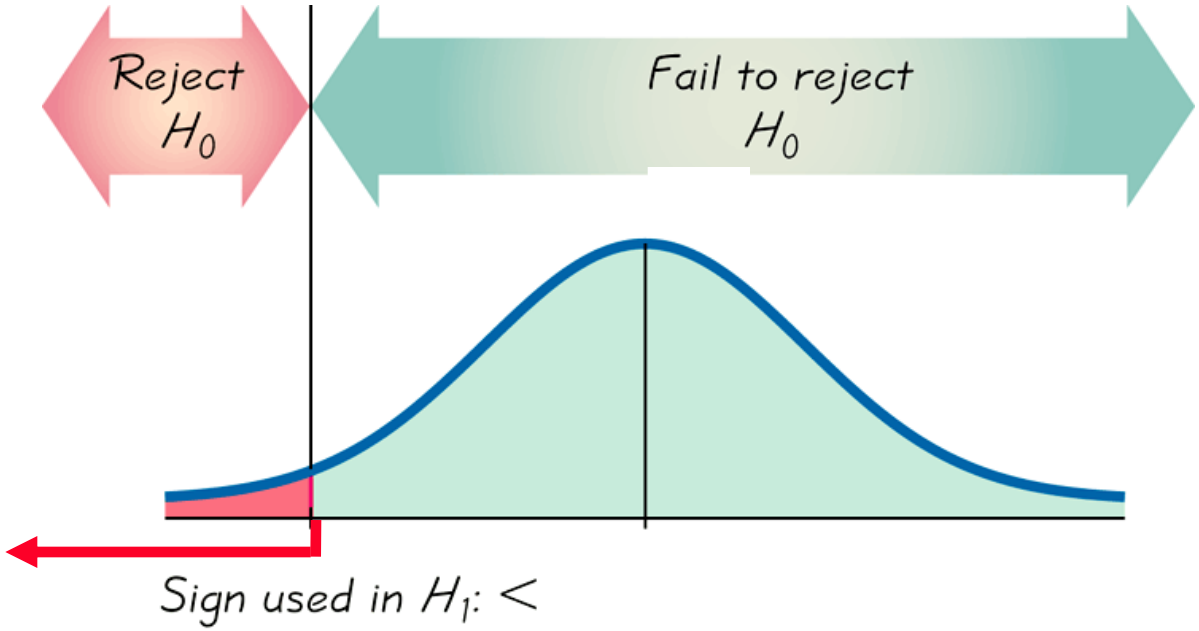


Baloldali tesztek

H_0: =

H_1: <

Pontok balra vmitől



P-érték

A *P*-érték (vagy *p*-érték vagy valószínűség érték) annak a valószínűsége, hogy a teszt statisztika olyan értéket adjon ami legalább annyira szélsőséges (extrém) mint az az érték amit a mintánkból kaptunk, azzal a feltevéssel, hogy a null hipotézis igaz. A null hipotézist elvetjük, ha a *P*-érték nagyon kicsi, mint pl. 0.05 vagy kevesebb.

A hipotézis tesztelés eredménye

Mindig a null hipotézist teszteljük. A kezdeti konklúzió mindig az alábbiak valamelyike:

1. Elvetjük a null hipotézist.

2. Nem tudjuk elvetni a null hipotézist.

Döntési kritériumok

Tradicionális módszer:

Elvetjük H_0-t, ha a teszt statisztika a kritikus tartományba esik.

Nem tudjuk elvetni H_0-t, ha a teszt statisztika nem esik a kritikus tartományba.

Döntési kritériumok - folyt

***P*-érték módszer**:

Elvetjük H_0-t ha a *P*-érték $\leq \alpha$ (ahol α a szignifikancia szint, mint pl. 0.05).

Nem tudjuk elvetni H_0-t, ha a *P*-érték $> \alpha$.

Döntési kritériumok - folyt

Egy másik lehetőség:

A szignifikancia szint megadása helyett, egyszerűen megkeressük a *P*-értéket, és a döntést az olvasóra hagyjuk.

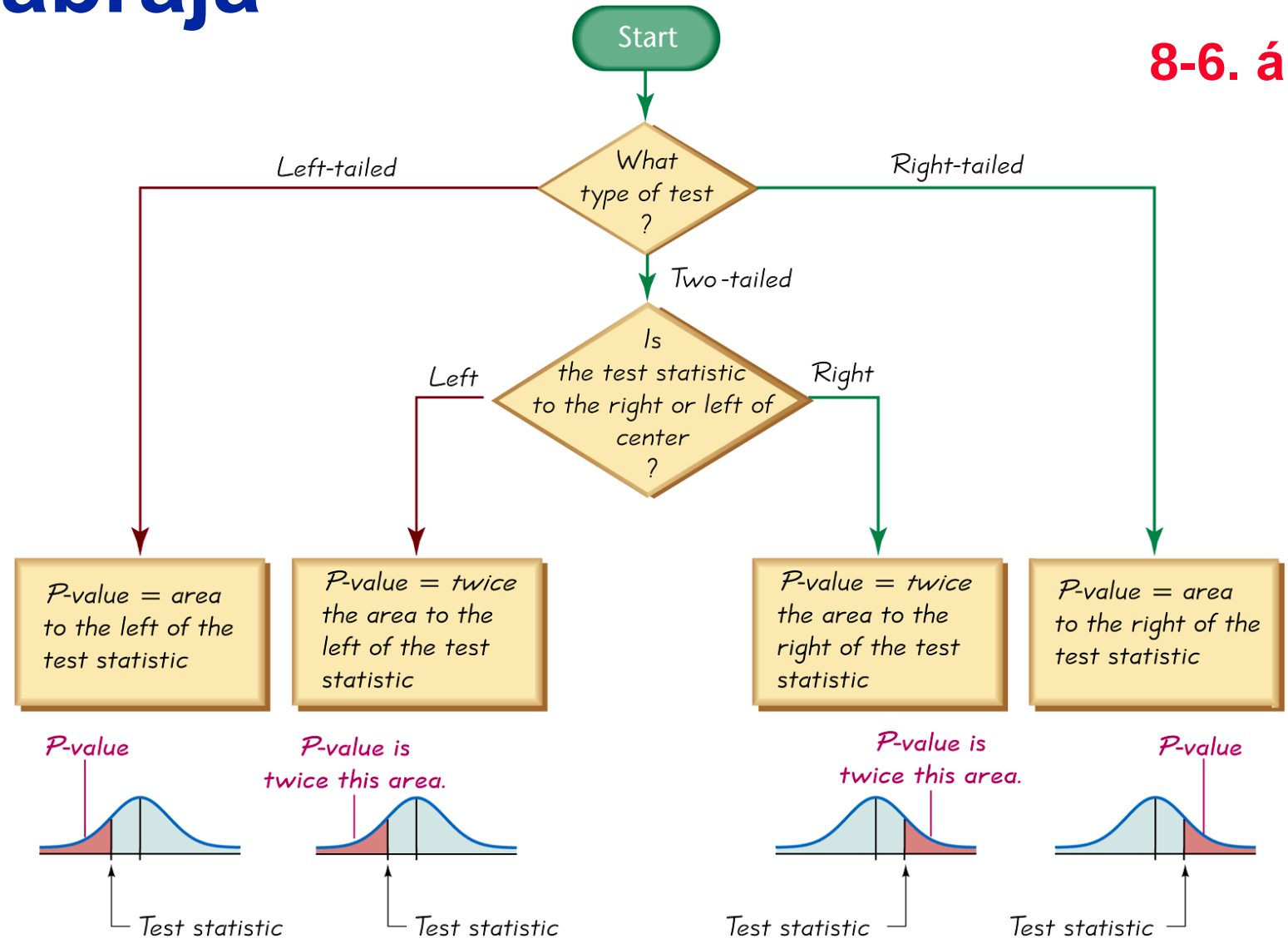
Döntési kritérium – folyt.

Konfidencia intervallum:

Mivel a konfidencia intervallum becslés tartalmazza a paraméter populációbeli értékét, utasítsuk el azokat a feltevéseket, melyek szerint a populáció paramétere a konfidencia intervallumon kívül esik.

A *P*-értékek megtalálásának ábrája

8-6. ábra



Példa: ***P*-érték kiszámítása. Először határozzuk meg, hogy az adott esetben jobboldali, baloldali vagy kétoldali tesztet végzünk-e, azután keresd meg a *P*-értéket és add meg a null hipotézissel kapcsolatos konklúziót.**

a) Az $\alpha = 0.05$ szignifikancia szintet használjuk annak a feltételezésnek a tesztelésére, hogy $p > 0.25$, és a minta adatok egy $z = 1.18$ értékű teszt statisztikát adnak.

b) Az $\alpha = 0.05$ szignifikancia szintet használjuk annak a feltételezésnek a tesztelésére, hogy $p \neq 0.25$, és a minta adatok egy $z = 2.34$ értékű teszt statisztikát adnak.

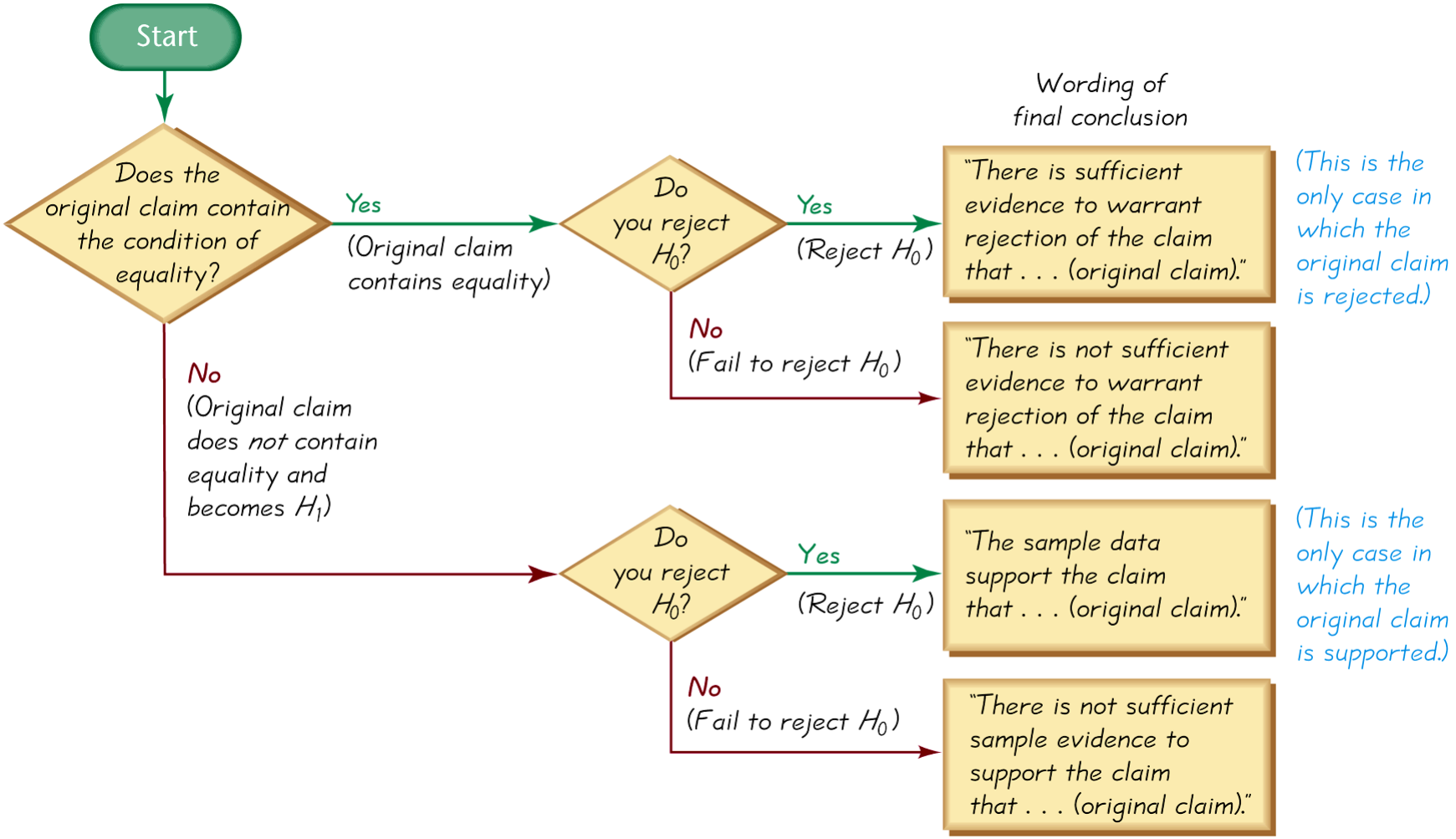
Példa: folyt.

a) A $p > 0.25$ feltevés esetén a teszt jobboldali. Mivel a teszt jobboldali, a *P*-érték a $z = 1.18$-től jobbra eső görbe alatti terület. Kikeresve a táblázatból ez 0.1190. A *P*-érték (0.1190) nagyobb mint a szignifikancia szint $\alpha = 0.05$, így nem tudjuk elvetni a null hipotézist. A *P*=0.1190 elég nagy, ami azt jelenti, hogy a minta érték könnyen megtörténhet véletlenül.

Példa: folyt.

b) A $p \neq 0.25$ feltevés esetén a teszt kétoldali. Mivel a teszt kétoldali és mivel a teszt statisztika $z = 2.34$ a középtől jobbra esik, a *P*-érték **kétszerese** a $z = 2.34$-től jobbra eső területnek. A táblázatból a $z = 2.34$-től jobbra eső terület 0.0096, így a *P*-érték = 2 x 0.0096 = 0.0192. Mivel a P=0.0192 kisebb vagy egyenlő mint a szignifikancia szintünk, el kell vetnünk a null hipotézist. A kicsiny *P*-érték (0.0192) azt mutatja, hogy a minta eredmény valószínűleg nem a véletlen eredménye.

A végső konklúziók megfogalmazása



8-7. ábra

Elfogadni vagy nem tudni elutasítani?

- Bizonyos könyvekben azt mondják “elfogadjuk a null hipotézist.”
- Nem tudjuk bizonyítani a null hipotézist.
- A minta bizonyítékok nem elég erősek ahhoz, hogy elutasítsuk (olyan mint amikor nincs elég bizonyíték, hogy elítéljék a gyanúsítottat).

I. fajú hiba

- Egy **I. fajú hiba** az, amikor hibás módon elutasítjuk a null hipotézist, amikor az igaz.
- Az α (alfa) szimbólummal jelöljük az I. fajú hiba valószínűségét.

II. fajú hiba

- Egy **II. fajú hiba** az, amikor nem utasítjuk el a null hipotézist akkor, amikor az nem igaz.
- A β (béta) szimbólummal jelöljük a II. fajú hiba valószínűségét.

Példa: Tegyük fel, hogy hipotézis tesztelést végzünk a $p > 0.5$ feltevéssel kapcsolatban. A null és az alternatív hipotézis a következő: H_0: $p = 0.5$, és H_1: $p > 0.5$.

a) Azonosítsuk az I. fajú hibát.
b) Azonosítsuk a II. fajú hibát.

Példa: folyt.

a) Az I. fajú hiba az, amikor elvetjük az igaz null hipotézist: Ha úgy látjuk, hogy elég evidencia támogatja $p > 0.5$-t, miközben a valóságban $p = 0.5$.

Példa: folyt.

b) A II. fajú hiba az, amikor nem vetjük el a null hipotézist, miközben az nem igaz: Nem utasítjuk el $p = 0.5$-öt (és ezért nem támogatjuk a $p > 0.5$-öt), miközben a valóságban $p > 0.5$.

Első és másodfajú hibák

Table 8-1 Type I and Type II Errors

		True State of Nature: The null hypothesis is true	True State of Nature: The null hypothesis is false
Decision	We decide to reject the null hypothesis	**Type I error** (rejecting a true null hypothesis) α	Correct decision
	We fall to reject the null hypothesis	Correct decision	**Type II error** (failing to reject a false null hypothesis) β

Az I. és II. fajú hibák kontrollálása

- Minden adott α esetén, a minta elemszám n növelése a β csökkenését okozza.
- Minden fix minta elemszám n esetén α csökkenése β növekedését okozza. Fordítva, α növelése β csökkenésére vezet.
- Ha α és β együttes csökkenését akarjuk elérni, akkor a minta elemszámot kell növelnünk.

Definíció

A hipotézis teszt erőssége az (1 - β) valószínűség érték, ami a helytelen null hipotézis elutasításának valószínűsége. Egy adott α szignifikancia szint és adott olyan másik populáció paraméter esetén számíthatjuk ki, ami a null hipotézisbeli érték alternatívája. Azaz a hipotézis teszt erőssége egy igaz alternatív hipotézis támogatásának valószínűsége.

Összefoglalás

Ebben a fejezetben megvitattuk:

- **Null és alternatív hipotézis.**
- **Teszt statisztika.**
- **Szignifikancia szintek.**
- ***P*-értékek.**
- **Düntési kritériumok.**
- **Első és másodfajú hibák.**
- **A hipotézis teszt ereje.**

8-3. fejezet
Az arányra vonatkozó feltevés tesztelése

Kulcsfogalmak

Ebben a fejezetben a populáció arány tesztelésének teljes folyamatát ismertetjük. Felhasználjuk az előző fejezetben bevezetett fogalmakat.

Feltevések

1) **Véletlen egyszerű mintavétel.**

2) **A binomiális elsoszlás feltételei fennállnak (5-3 fejezet).**

3) **Az $np \geq 5$ és $nq \geq 5$ feltételek fennállnak, így a binomiális eloszlást egy olyan normálissal közelíthetjük, aminek a paraméterei $\mu = np$ $\sigma = \sqrt{npq}$.**

Jelölések

n = a kísérletek száma

$\hat{p} = \frac{x}{n}$ (minta arány)

p = populáció arány (amit a null hipotézisben használunk)

$q = 1 - p$

Az arányra vonatkozó teszt statisztika

$$z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}}$$

P-érték módszer

Ugyanúgy, mint a 8-2 fejezetben …

Tradicionális módszer

Ugyanaz, mint a 8-2-ben …

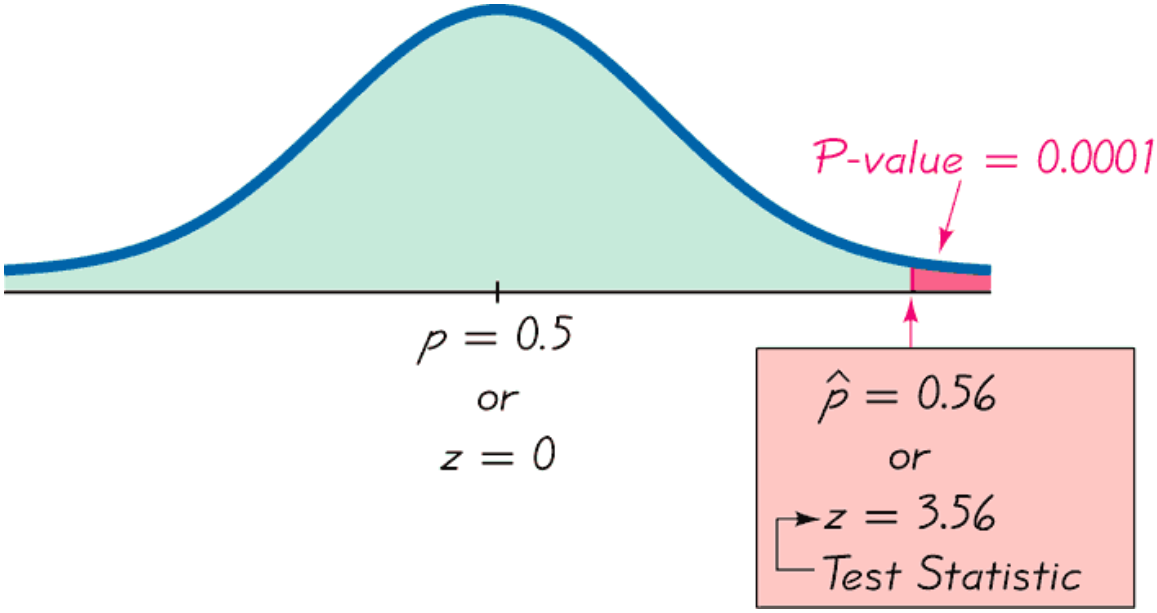
Konfidencia intervallum módszer

Ugyanaz, mint a 8-2-ben …

Példa: 880 véletlenül választott autóvezető 56%-a elismeri, hogy néha átmegy a piroson. Az a feltételezésünk, hogy a vezetők többsége néha átmegy a piroson, azaz $p > 0.5$. A minta adatok $n = 880$, $\hat{p} = 0.56$.

$$np = (880)(0.5) = 440 \geq 5$$
$$nq = (880)(0.5) = 440 \geq 5$$

Példa: folyt.



H_0: $p = 0.5$
H_1: $p > 0.5$
$\alpha = 0.05$

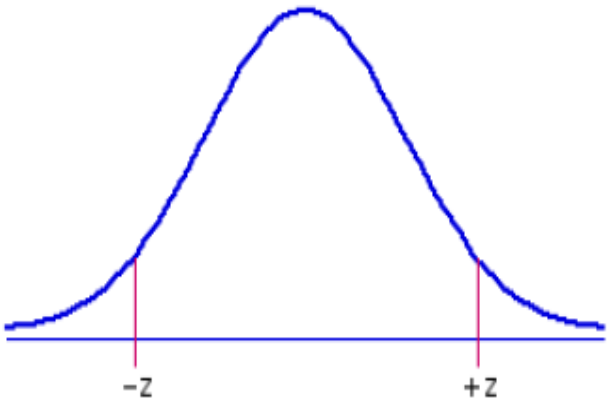
$$z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}} = \frac{0.56 - 0.5}{\sqrt{\frac{(0.5)(0.5)}{880}}} = 3.56$$

P=0.0001
Mivel P<0.05=α, ezért elutasítjuk a null hipotézist.
Mivel $z > z_\alpha = 1.645$, ezért elutasítjuk a null hipotézist.
Elegendő bizonyítékunk van a feltételezésünk elfogadására.

Példa: **Amikor Gregor Mendel borsó hibridizációs kísérletét végezte, az egyik kísérletben 428 zöld borsószem és 152 sárga borsószem termett. Mendel elmélete szerint a borsók ¼-e volt sárgának várható. Használjunk 0.05 szignifikancia szintű tesztet és a *P*-érték módszert, hogy teszteljük, vajon a sárga szemek aránya ¼-e vagy sem.**

Észrevétel: $n = 428 + 152 = 580$,
így $\hat{p} = 0.262$, **és** $p = 0.25$.

Segítség: http://faculty.vassar.edu/lowry/tabs.html



Reset Calculate z = 0.67

one-tailed for –z	0.2514
one-tailed for +z	0.2514
two-tailed for ±z	0.5029
area between ±z	0.4971

Text will appear in this box only if the value of |z| is greater than 3.75.

Return to Top

H_0: $p = 0.25$
H_1: $p \neq 0.25$
$n = 580$
$\alpha = 0.05$
$\hat{p} = 0.262$

$$z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}} = \frac{0.262 - 0.25}{\sqrt{\frac{(0.25)(0.75)}{580}}} = 0.67$$

Mivel ez egy kétoldali teszt, a *P*-érték a kétszerese a statisztika értékétől jobbra eső területnek. P=0.502. Nincs elég bizonyítékunk, hogy elutasítsuk a null hipotézist, azaz azt, hogy a borók ¼-e sárga.

Összefoglalás

Ebben a fejezetben megvitattuk:

- **Az arányokra vonatkozó teszt statisztikát.**
- ***P*-érték módszer.**

^

8-4. fejezet
Az átlagra vonatkozó feltételezés tesztelése: σ ismert

Kulcsfogalmak

Ilyet is lehet csinálni pedagógiai okokból, de gyakorlati jelentősége nincs. A következő fejezet eredményei igazak, csak s=σ-t kell feltételeznünk.

8-5. fejezet
A populáció átlagra vonatkozó feltételezés tesztelése: σ nem ismert

Kulcsfogalmak

Ebben a fejezetben a populáció átlagára vonatkozó hipotézisek vizsgálatáról lesz szó, abban az esetben, amikor σ nem adott. Ebben a fejezetben a Student t eloszlást használjuk.

Feltételek

1) A minta véletlen egyszerű.

2) Valamelyik, vagy mindkét feltétel igaz: A populáció normális eloszlású, vagy $n > 30$.

Teszt statisztika

$$t = \frac{\bar{x} - \mu_{\bar{x}}}{\frac{s}{\sqrt{n}}}$$

P-értékek és kritikus értékek

- Táblázat, online kalkulátor, program stb.
- Szabadsági fokok száma $n - 1$

Példa: 13 piros M&M csoki golyót véletlenül választuk egy zacskóból, amiben 465 M&M csoki golyó van. A tömegük (grammokban) átlagosan $x = 0.8635$ és a szórás $s = 0.0576$ g.

A zacskó szerint a nettó tömeg 396.9 g, azaz az M&M csoki golyók tömege elvben 396.9/465 = 0.8535 g. Használjuk a minta adatokat és a 0.05-ös szignifikancia szintet, hogy teszteljük a gyártósor vezetőjének azon kijelentését, mi szerint a csoki golyók tömege valójában nagyobb mint 0.8535 g.

A minta $n = 13$ elemű és a normál q-q plot szerint normálissal közelíthető.

H_0: $\mu = 0.8535$
H_1: $\mu > 0.8535$
$\alpha = 0.05$
$\bar{x} = 0.8635$
$s = 0.0576$
$n = 13$

$$t = \frac{\bar{x} - \mu_{\bar{x}}}{\frac{s}{\sqrt{n}}} = \frac{0.8635 - 0.8535}{\frac{0.0576}{\sqrt{13}}} = 0.626$$

A kritikus érték $t_\alpha = 1.782$

Mivel a teszt statisztika értéke $t = 0.626$ nem esik a kritikus tartományba, nem tudjuk elvetni a null hipotézist H_0. Nincs elegendő bizonyíték annak a feltételezésnek a támogatására, hogy az M&M csoki golyók tömege több mint 0.8535 g.

Összefoglalás

Ebben a fejezetben megvitattuk:

- **Feltételek**
- **Student *t* eloszlás.**
- ***P*-érték módszer.**

8-6. fejezet
A szórásra és a varianciára vonatkozó feltevések becslése

Kulcsfogalmak

Ebben a fejezetben a populáció szórására σ vagy varianciájára σ^2 vonatkozó feltételezés tesztelésével foglalkozunk. A módszerek támaszkodni fognak a 7-5. fejezetben bevezetett khí-négyzet eloszlásra.

Feltételek

1. Véletlen egyszerű minta.

2. A populáció normális eloszlású. (Ez egy sokkal erősebb feltétel, mint amit az átlag tesztelésekor használunk!)

Khí-négyzet eloszlás

Teszt statisztika

$$\chi^2 = \frac{(n-1)\,S^2}{\sigma^2}$$

n = minta elemszám

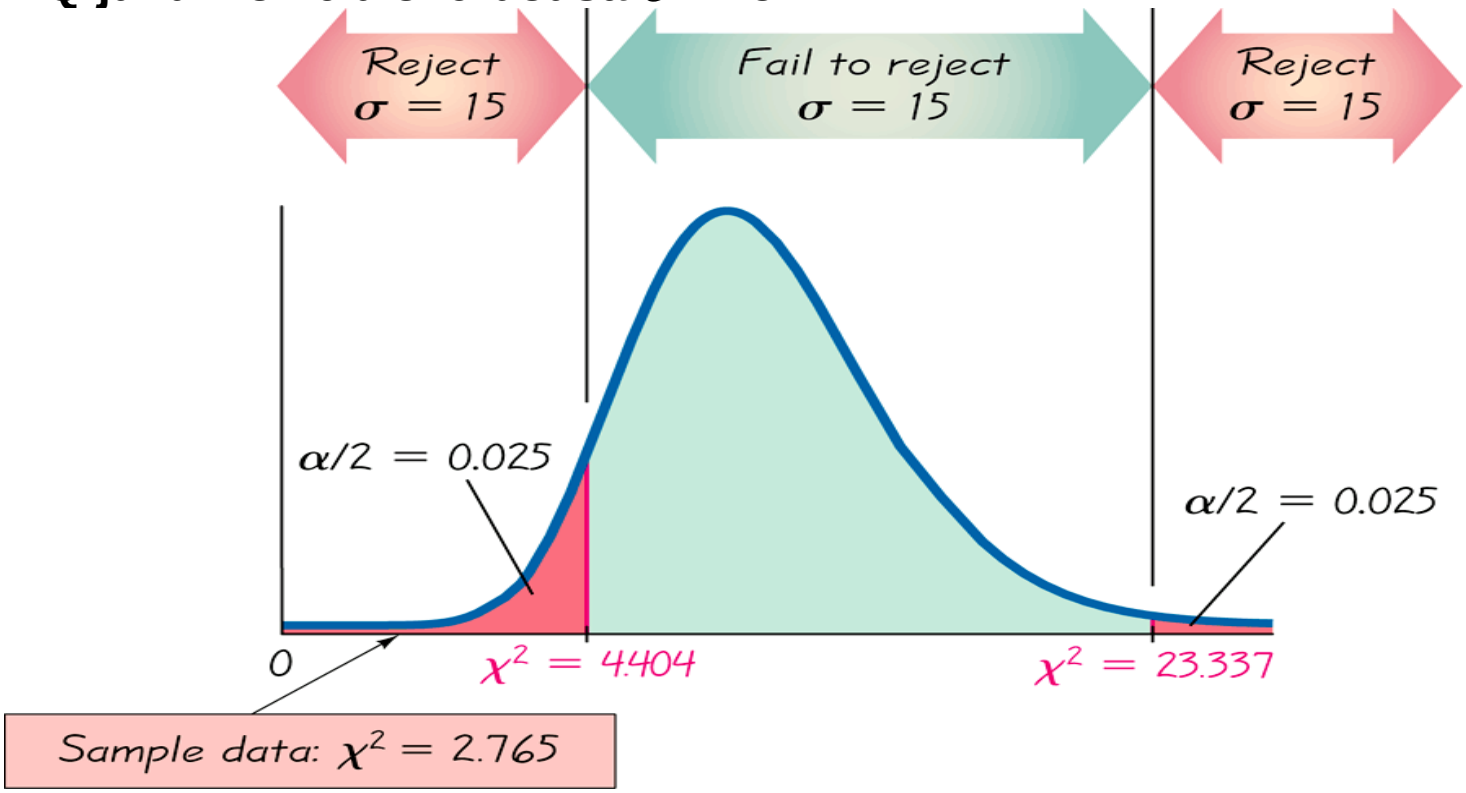
s^2 = minta variancia

σ^2 = populáció variancia

(a null hipotézisben van megadva!)

Példa: **A felnőttek egy egyszerű véletlen mintájában az IQ értékek normálisan oszlanak el 100 átlaggal és 15 szórással.**
Egy egyszerű véletlen 13 fizika professzorból álló mintának a szórása s = 7.2. Tegyük fel, hogy a fizika professzorok IQ-ja is normális eloszlású. Teszteljük 0.05 szignifikancia szinten azt a feltételezést, hogy a fizika professzorok IQ-jának is 15 a szórásasa σ = 15.

H_0: $\sigma = 15$
H_1: $\sigma \neq 15$
$\alpha = 0.05$
$n = 13$
$s = 7.2$



$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} = \frac{(13-1)(7.2)^2}{15^2} = 2.765$$

Példa: folyt.

H_0: $\sigma = 15$
H_1: $\sigma \neq 15$
$\alpha = 0.05$
$n = 13$
$s = 7.2$

$$\chi^2 = 2.765$$

A kritikus értékek 4.404 és 23.337 (szabadsági fokok száma (df) = n – 1=12) és a táblázatban a 0.025 és 0.975 értékekhez tartoznak. Mivel a statisztika értéke a kritikus tartományba esik, el kell utasítanunk a null hipotézist. Elegendő bizonyítékunk van arra, hogy elutasítsuk azt a feltételezést, miszerint a fizika professzorok IQ-jának szórása éppen 15.

Összefoglalás

Ebben a fejezetben megvitattuk:

- **A szórásra és a varianciára vonatkozó teszteket.**
- **A teszt statisztikát.**
- **A kritikus értékeket.**

10. Előadás
Korreláció és regresszió

10-1 Áttekintés

10-2 Korreláció

10-3 Regresszió

10-4 Konfidencia és predikciós sávok

10-5 Többszörös regresszió

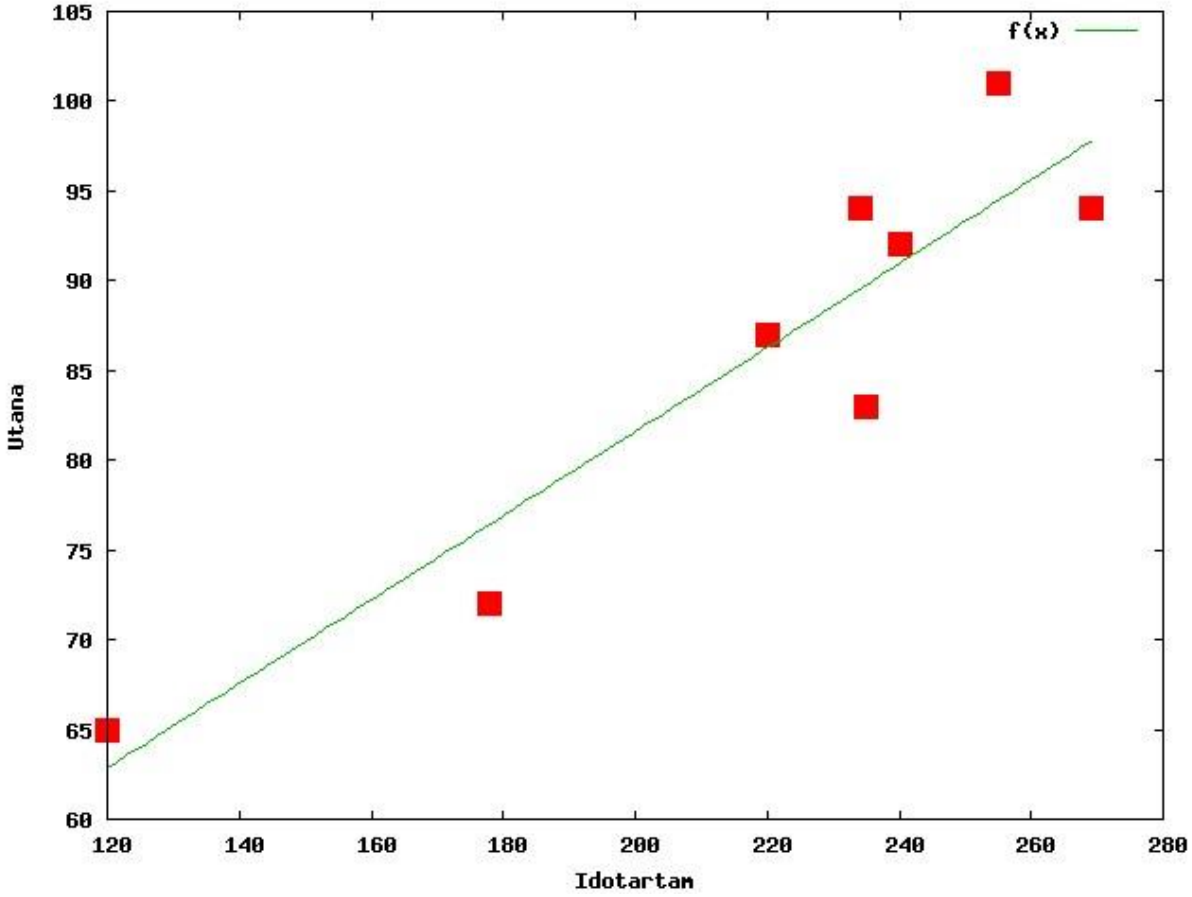
10-6 Modellezés

10-1. fejezet
Áttekintés

Old Faithful Geyser (Yellowstone)

- http://www.nps.gov/yell/tours/livecams/oldfaithful/OFVChours.htm
- A kitörések néhány adata percekben ill. méterekben
- **10-1. Táblázat:**

Időtartam	240	120	178	234	235	269	255	220
Előző	98	90	92	98	93	105	81	108
Következő	92	65	72	94	83	94	101	87
Magasság	42	33	38	36	42	36	38	45


f(x)
Utana
Idotartam
105
100
95
90
85
80
75
70
65
60
120
140
160
180
200
220
240
260
280

Áttekintés

Ebben a fejezetben bevezetjük a korreláció fogalmát, amelynek segítségével összefüggést lehet keresni két valószínűségi változó között, és bizonyos esetekben az egyik változó értékének ismeretében a másik értékére lehet következtetni.

Olyan mintákkal fogunk foglalkozni, ahol a minta adatok párokba vannak rendezve.

10-2. fejezet
Korreláció

Kulcsfogalmak

Ebben a fejezetben bevezetjük a lineáris korrelációs együttható *r* fogalmát, ami két véletlen változó közti kapcsolat erősségét számszerűen méri.

Mivel a korrelációs együttható könnyen kiszámítható, ezért itt főleg a fogalom megértésére koncentrálunk.

Definíció

Két változó között korreláció lép fel, ha az egyik a másikkal valamilyen módon kapcsolatban van.

Definíció

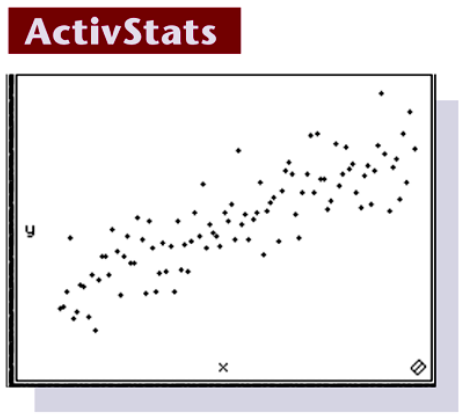
A lineáris korrelációs együttható r méri a lineáris kapcsolat erősségét egy x és y párokból álló minta értékei között.

Az adatok feltárása

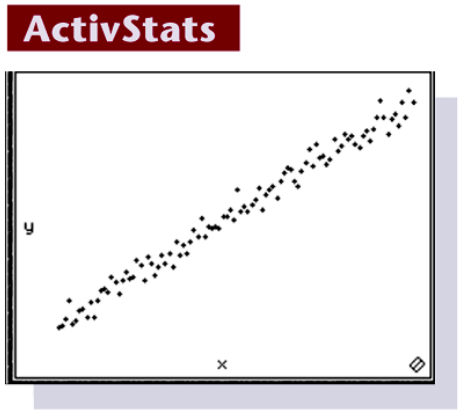
Gyakran felfedezhetünk kapcsolatot két változó között a szórásdiagram segítségével.

A következő 10-2. ábra néhány különböző tulajdonságokkal rendelkező szórásdiagramot mutat be.

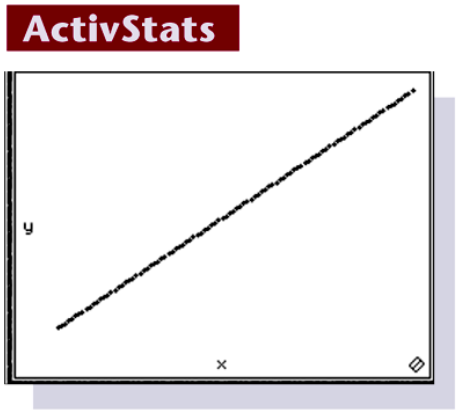
Szórásdiagramok párosított adatokra



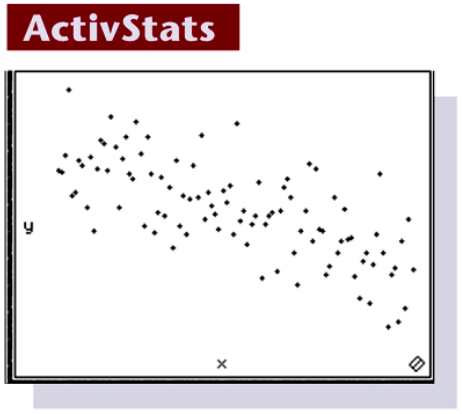
(a) Positive correlation:
$r = 0.851$



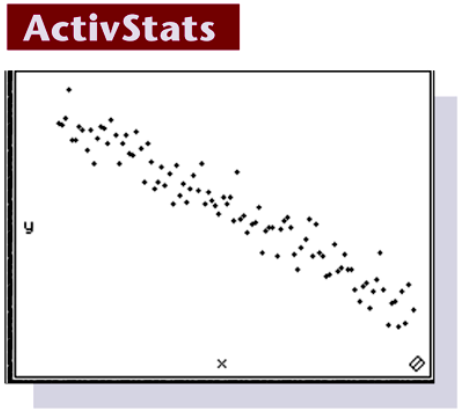
(b) Positive correlation:
$r = 0.991$



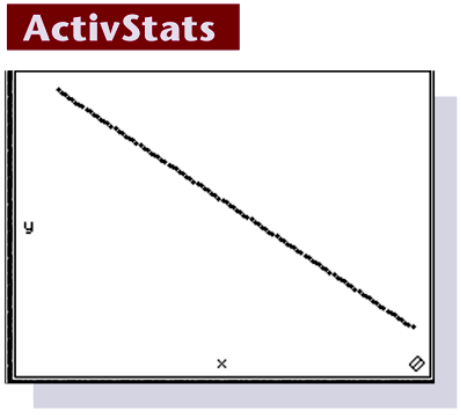
(c) Perfect positive correlation:
$r = 1$



(d) Negative correlation:
$r = -0.702$



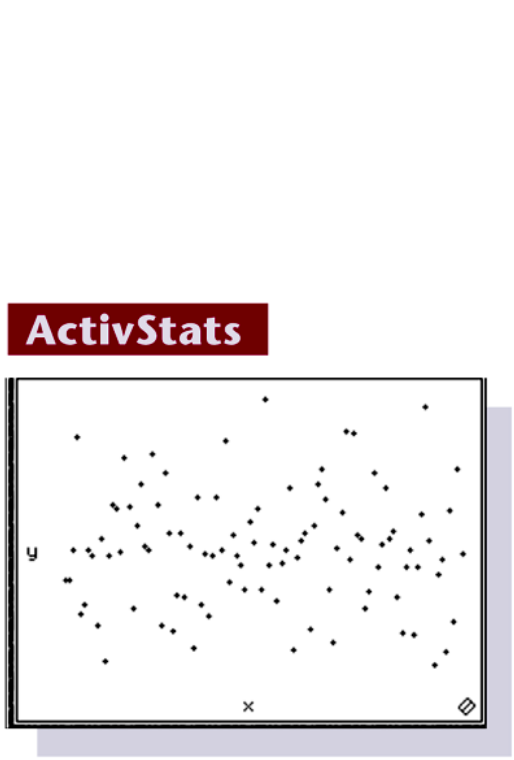
(e) Negative correlation:
$r = -0.965$



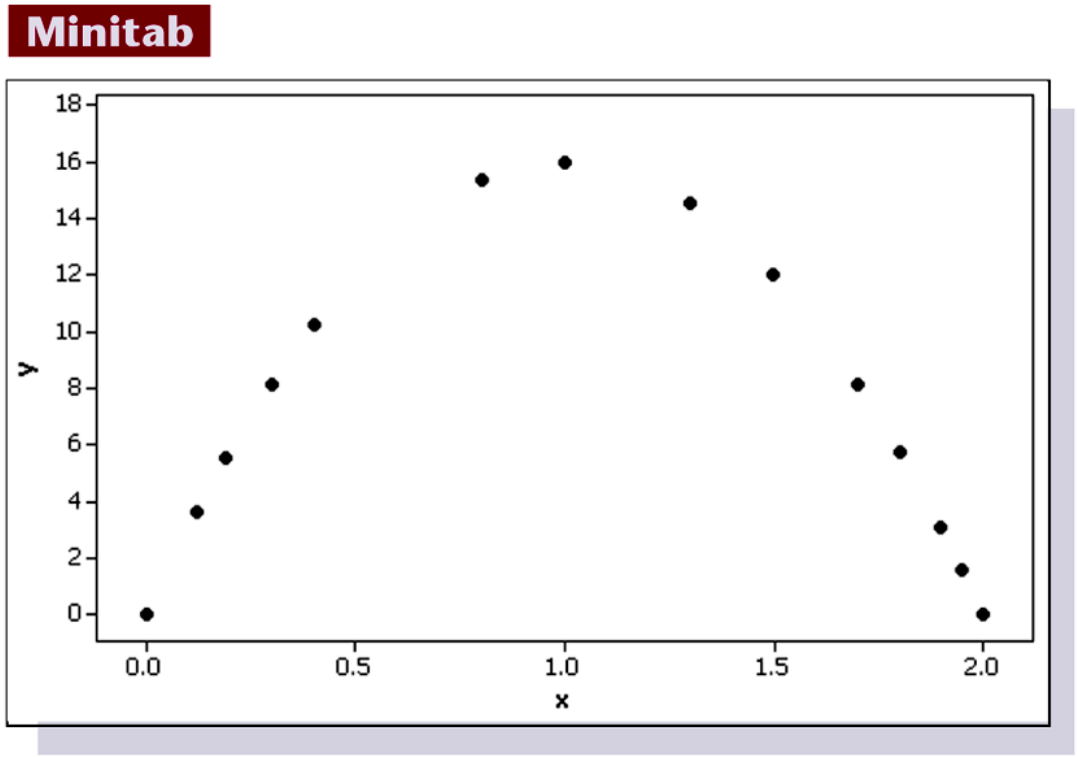
(f) Perfect negative correlation:
$r = -1$

10-2. ábra

Szórásdiagramok párosított adatokra



(g) No correlation: $r = 0$



(h) Nonlinear relationship: $r = -0.087$

10-2. ábra

Követelmények

1. Az (x, y) párokból álló adatok véletlen független minta adatok.
2. Vizuálisan meg kell győződnünk arról, hogy az adatok nagyjából egyenest alkotnak. (Nem determinisztikusak vagy más bonyolultabb alakjuk van.)
3. Az outliereket el kell távolítani, amennyiben meggyőződtünk arról, hogy hibásak voltak. Az *r* értékét ki kell számítani az outlierekkel együtt és azok nélkül. Meg kell nézni, mekkora a hatásuk.

Jelölések

n	az adatpárok száma
Σ	az adott értékek összegzése
Σx	az x értékek összege
Σx^2	minden x értéket négyzetre kell emelni és utána összeadni
$(\Sigma x)^2$	először össze kell adni az x értékeket, majd az eredményt négyzetre kell emelni
Σxy	minden x értéket meg kell szorozni a párjának y értékével, majd a szorzatokat összeadni
r	a minta lineáris korrelációs együtthatója.
ρ	a populáció lineáris korrelációs együtthatója.

Képletek

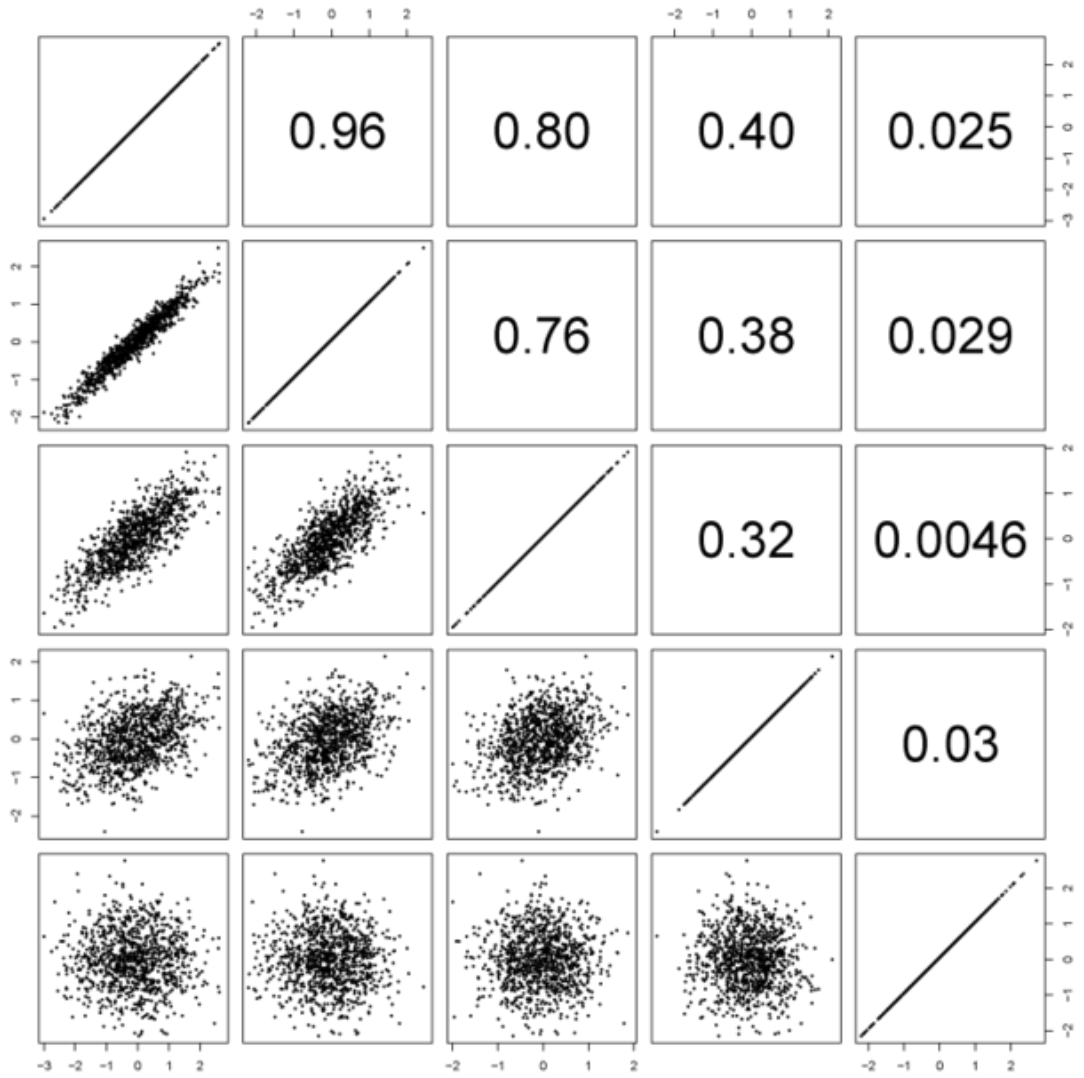
Az *r* lineáris korrelációs együttható méri a lineáris kapcsolat erősségét a minta adatpárok tagjai között (x és y között!).

$$r = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n(\Sigma x^2) - (\Sigma x)^2}\sqrt{n(\Sigma y^2) - (\Sigma y)^2}}$$

10-1. képlet

$$r_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$

1000 normális eloszlású adatpár különböző r értékekkel



r interpretálása

Táblázat: Ha az *r* abszolút értéke nagyobb, mint a következő táblázatban, akkor arra következtetünk, hogy van lineáris korreláció.

Critical Values for the Correlation Coefficient		
Number of Points	95% Confidence	99% Confidence
3	0.997	1.000
4	0.950	0.990
5	0.878	0.959
6	0.811	0.917
7	0.754	0.875
8	0.707	0.834
9	0.666	0.798
10	0.632	0.765
11	0.602	0.735
12	0.576	0.708

Példa: r kiszámítása

Az alábbi egyszerű véletlen mintaadatokat használva számítsuk ki r értékét.

Adatok:

x	3	1	3	5
y	5	8	6	4

Table 10-2 Finding Statistics Used to Calculate r

	x	y	$x \cdot y$	x^2	y^2
	3	5	15	9	25
	1	8	8	1	64
	3	6	18	9	36
	5	4	20	25	16
Total	**12**	**23**	**61**	**44**	**141**
	↑	↑	↑	↑	↑
	Σx	Σy	Σxy	Σx^2	Σy^2

483. oldal

Példa: folyt.

Adatok:

x	3	1	3	5
y	5	8	6	4

$$r = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n(\Sigma x^2) - (\Sigma x)^2}\sqrt{n(\Sigma y^2) - (\Sigma y)^2}}$$

$$r = \frac{4(61) - (12)(23)}{\sqrt{4(44) - (12)^2}\ \sqrt{4(141) - (23)^2}}$$

$$r = \frac{-32}{33.466} = -0.956$$

Példa: folyt.

Adott r = - 0.956, ha 0.05-ös szignifikancia szintet használunk, akkor arra jutunk, hogy van lineáris kapcsolat x és y között, mivel r abszolút értéke meghaladja a 0.950-ös kritikus értéket. Azonban, ha a 0.01-es szignifikancia szintet használjuk, akkor nem jutunk arra, hogy lineáris kapcsolat van, mert r abszolút értéke nem haladja meg a 0.990-es kritikus értéket.

Példa: Old Faithful

A 10-1. táblázat adatait használva, keressük meg a lineáris korrelációs együttható értékét *r*, majd ellenőrizzük, hogy van-e szignifikáns lineáris kapcsolat a változók között.

Ugyanúgy számolva, mint előbb $r = 0.926$ adódik.

A táblázatban az $n = 8$ adatpont esetét keressük ki. Az $\alpha = 0.05$-höz tartozó értéket leolvasva, 0.707 kritikus értéket kapunk. Mivel $r = 0.926$, abszolút értéke több mint 0.707, úgy döntünk, hogy van lineáris kapcsolat a kitörések hossza és az utánuk következő várakozási idők között.

A lineáris korrelációs együttható tulajdonságai

1. $-1 \leq r \leq 1$
2. Az r értéke nem változik, ha bármelyik változónak megváltoztatjuk a mértékegységét.
3. Az r értékét nem befolyásolja az x és y felcserélése.
4. r méri a lineáris kapcsolat erősségét.

Interpretáció: Megmagyarázott variabilitás

Az r^2 érték mondja meg, hogy y variabilitásának hányad részét magyarázza az x és y közti lineáris kapcsolat.

Példa: Old Faithful

A kitörés után eltelő idő ingadozásának mekkora részét magyarázza meg a kitörés időtartamának ingadozása?

$r = 0.926$, akkor $r^2 = 0.857$.

Azt mondhatjuk, hogy 0.857-ed részét (vagy 86%-át) magyarázza meg a kitörések után eltelő idő ingadozásának a kitörés hosszával való lineáris kapcsolata. Ez azt is jelenti, hogy a kitörések után eltelő idő hosszának 14%-ára nem ad magyarázatot a kitörések hossza.

Szokásos hibák a korrelációval kapcsolatban

1. **Oksági összefüggés**: Hibás azt állítani, hogy a korreláció oksági kapcsolatot jelent.

2. **Átlagolás**: Az átlagolás elnyomja az az eredeti adatokban meglévő ingadozásokat, ami csökkenti a korrelációs együtthatót.

3. **Linearitás**: Lehatséges, hogy van valamilyen kapcsolat x és y között, még akkor is, ha nincs köztük lineáris korreláció.

Formális hipotézis tesztelés

- Szeretnénk meghatározni, hogy van-e szignifikáns lineáris kapcsolat két változó között.
- Legyen a null és alternatív hipotézis:

H_0: $\rho = 0$ (nincs szignifikáns lin. korreláció)

H_1: $\rho \neq 0$ (szignifikáns lin. korreláció)

Teszt statisztika

Teszt statisztika:

$$t = \frac{r}{\sqrt{\frac{1 - r^2}{n - 2}}}$$

A transzformáció után t statisztika!

Kritikus értékek:

Megegyezik az n-2 szabadsági fokú Student *t* statisztikával!

Összefoglalás

Ebben a fejezetben megvitattuk a:

- **Korrelációt.**
- **A lineáris korrelációs együtthatót.**
- **A feltételeket .**
- **Az interpretációt.**
- **Formális hipotézis tesztelést.**

10-3. fejezet
Regresszió

Kulcsfogalmak

A legfontosabb ebben a fejezetben, hogy meghatározzuk azt az egyenest, és azt az egyenletet, ami legjobban reprezentálja a változók közti kapcsolatot.

Az egyenest regressziós egyenesnek nevezik és az egyenletet regressziós egyenletnek.

Regresszió

A regressziós egyenlet az x változó (független változó, prediktor változó vagy magyarázó változó) , és az y változó (függő változó vagy válasz változó vagy magyarázott változó) közötti kapcsolatot adja meg.

A tipikus lineáris kapcsolatot $\hat{y} = mx + b$, *vagy* az $\hat{y} = b_0 + b_1 x$, formában fejezzük ki, ahol b_0 az y-tengelymetszet és b_1 a meredekség.

Feltételek

1. Az adatpárok (x, y) véletlen minta adatok.
2. Vizuális vizsgálattal arra jutunk, hogy a szórásdiagram egy egyeneshez hasonló.
3. Ki kell hagyni azokat az outliereket, amik hibák miatt vannak jelen.

Definíciók

❖ Regressziós egyenlet

Az adatpárok egy halmaza esetén a regressziós egyenlet:

$$\hat{y} = b_0 + b_1 x$$

algebrailag leírja a kapcsolatot a két változó között.

❖ Regressziós egyenes

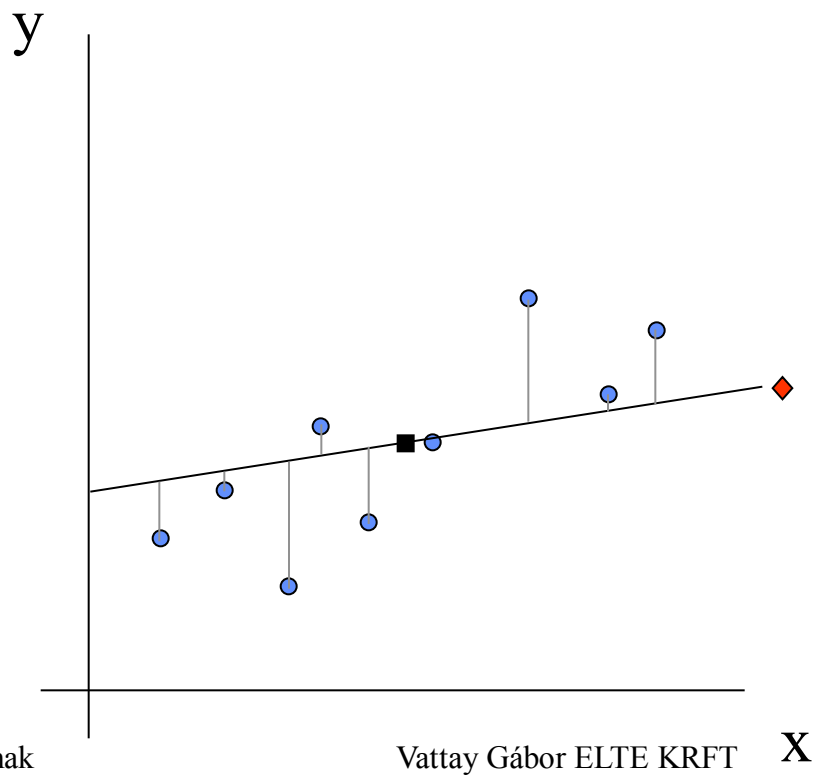
A regressziós egyenes (vagy legjobban illő egyenes, vagy a négyzetesen legjobb egyenes) a regressziós egyenlet gráfja.

Jelölések

	Populáció paraméter	Minta becslés
y-tengelymetszet	β_0	b_0
Meredekség	β_1	b_1
Egyenlet	$y = \beta_0 + \beta_1 x$	$\hat{y} = b_0 + b_1 x$

Speciális tulajdonság

A regressziós egyenes illik legjobban az adatokhoz.



A legkisebb négyzetek módszere

Keressük azt az egyenest, aminél a reziduumok négyzetének összege a lehető legkisebb:

$$F(b_0, b_1) = \sum_i (y_i - (b_0 + b_1 x_i))^2$$

Megkeressük azokat a paramétereket, amelyek mellett a fenti összeg a legkisebb:

$$\frac{\partial F}{\partial b_0} = 0 \qquad \frac{\partial F}{\partial b_1} = 0$$

folyt.

Bontsuk fel a négyzetet:

$$F(b_0, b_1) = \sum y_i^2 + b_1^2 \sum x_i^2 + n b_0^2 - 2 b_1 \sum x_i y_i - 2 b_0 \sum y_i + 2 b_1 b_0 \sum x_i$$

Végezzük el az egyik deriválást:

$$0 = \frac{\partial F(b_0, b_1)}{\partial b_0} = 2(n b_0 - \sum y_i + b_1 \sum x_i)$$

Fejezzük ki az egyik paramétert:

$$b_0 = \frac{1}{n} \sum y_i - b_1 \frac{1}{n} \sum x_i = \bar{y} - b_1 \bar{x}$$

folyt.

Végezzük el a másik deriválást is:

$$0 = \frac{\partial}{\partial \beta_1}(\ldots) = 2\left(\beta_1 \sum x_i^2 + \beta_0 \sum x_i - \sum x_i y_i\right)$$

Oldjuk meg:

$$\sum x_i y_i = \beta_0 \sum x_i + \beta_1 \sum x_i^2$$

$$= \sum x_i (\beta_0 + \beta_1 x_i) + \sum \varepsilon_i (\beta_0 + \beta_1 x_i) + \sum \varepsilon_i^2$$

$$\rightarrow \quad \beta_1 = \frac{\sum x_i y_i - \frac{1}{n}\left(\sum x_i\right)\left(\sum y_i\right)}{\sum x_i^2 - \frac{1}{n}\left(\sum x_i\right)^2}$$

A b_0 és b_1 képletei

10-2. képlet $$b_1 = \frac{n(\Sigma xy) - (\Sigma x)\,(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2}$$ **(meredekség)**

10-3. képlet $$b_0 = \bar{y} - b_1 \bar{x}$$ **(y tengelymetszet)**

A regressziós egyenes kiszámítása

Adatok:

x	3	1	3	5
y	5	8	6	4

A 10-2. fejezetben ezeket az adatokat használva kiszámítottuk a korrelációs együtthatót $r = -0.956$. Határozzuk meg a regressziós egyenest!

folyt.

Adatok:

x	3	1	3	5
y	5	8	6	4

$n = 4$
$\Sigma x = 12$
$\Sigma y = 23$
$\Sigma x^2 = 44$
$\Sigma y^2 = 141$
$\Sigma xy = 61$

$$b_1 = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2}$$

$$b_1 = \frac{4(61) - (12)(23)}{4(44) - (12)^2}$$

$$b_1 = \frac{-32}{32} = -1$$

folyt.

Adatok:

x	3	1	3	5
y	5	8	6	4

$n = 4$
$\Sigma x = 12$
$\Sigma y = 23$
$\Sigma x^2 = 44$
$\Sigma y^2 = 141$
$\Sigma xy = 61$

$$b_0 = \overline{y} - b_1 \overline{x}$$

$$5.75 - (-1)(3) = 8.75$$

folyt.

Adatok:

x	3	1	3	5
y	5	8	6	4

$n = 4$
$\Sigma x = 12$
$\Sigma y = 23$
$\Sigma x^2 = 44$
$\Sigma y^2 = 141$
$\Sigma xy = 61$

A kiszámított regressziós egyenlet:

$$\hat{y} = 8.75 - 1x$$

Példa: Old Faithful

A 10-1. táblázat alapján, számítsuk ki a regressziós egyenest.

Ugyanazokat a lépéseket végigcsinálva, mint az előbb, kapjuk b_1 = 0.234 és b_0 = 34.8. Így a regressziós egyenlet:

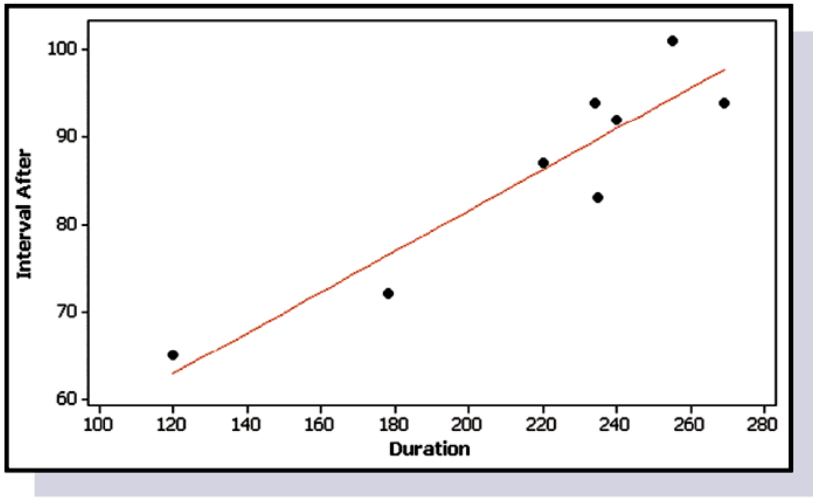
$$\hat{y} = 34.8 + 0.234x$$

Példa: Old Faithful - folyt

```
The regression equation is
Interval After = 34.8 + 0.234 Duration

Predictor      Coef  SE Coef     T      P
Constant     34.770    8.732  3.98  0.007
Duration    0.23406  0.03908  5.99  0.001

S = 4.97392   R-Sq = 85.7%   R-Sq(adj) = 83.3%
```



Predikciók

Az y értékének becslése az x adott értékére alapozva ...

1. Ha nem tudunk semmilyen kapcsolatról x és y között, akkor a legjobb predikció y értékére $\overline{y}$.
2. Ha van ismert lineáris kapcsolat, akkor a legjobb predikció, ha a regressziós egyenletbe behelyettesítjük x értékét és kiszámítjuk hozzá az y értékét.

Példa: Old Faithful

A 10-1. táblázat alapján azt találtuk, hogy a regressziós egyenlet $\hat{y} = 34.8 + 0.234x$. Feltéve, hogy az utolsó kitörés hossza $x = 180$ másodperc volt, keressük meg a legjobb becslést y-ra, azaz a következő kitörésig eltelő időre.

$$\hat{y} = 34.8 + 0.234x$$

$$34.8 + 0.234(180) = 76.9 \text{ perc}$$

Az előrejelzett idő 76.9 perc.

Definíciók

- **Marginális változás**

  A **marginális változás** az a mennyiség, amennyit a változó változik, miközben a másikat egy egységnyivel megváltoztatjuk.

- **Outlier**

  Egy **outlier** egy olyan pont, ami a többitől messze esik.

- **Torzító pont**

  Egy torzító pont erősen befolyásolja a regressziós egyenes elhelyezkedését.

Definíciók

Reziduum

A reziduum egy (x, y) adatpár esetén , az $(y - \hat{y})$ különbség a megfigyelt y minta érték és a regressziós egyenes által adott y érték között.

reziduum = megfigyelt y – prediktált $y = y - \hat{y}$

Definíciók

- **Legkisebb négyzetek tulajdonság**

  Egy egyenes rendelkezik a legkisebb négyzetek tulajdonsággal ha a reziduumok négyzeteinek összege a lehető legkisebb.

- **Reziduális diagram**

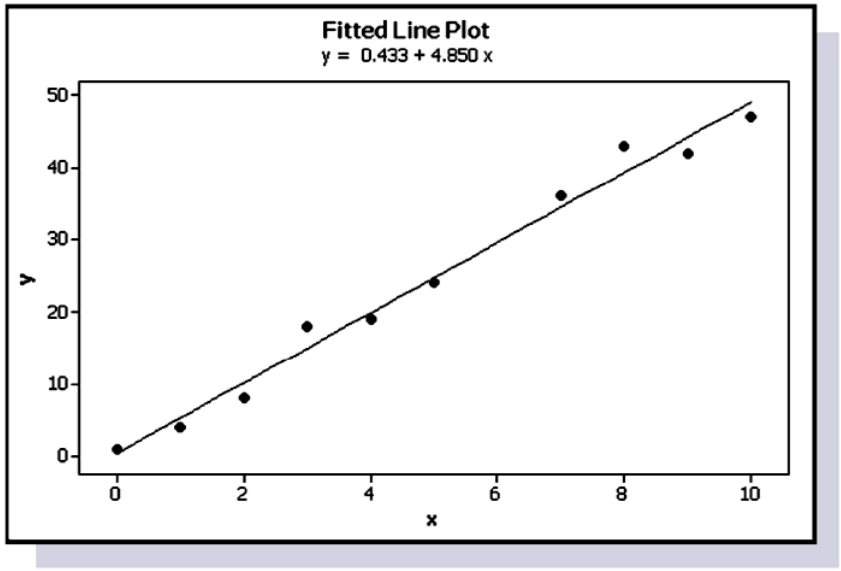
  Az (x, y) értékekből képzett szórásdiagramban az y-koordinátát az $y - \hat{y}$ reziduummal helyettesítjük. A reziduális diagram az $(x, y - \hat{y})$ pontpárokból áll.

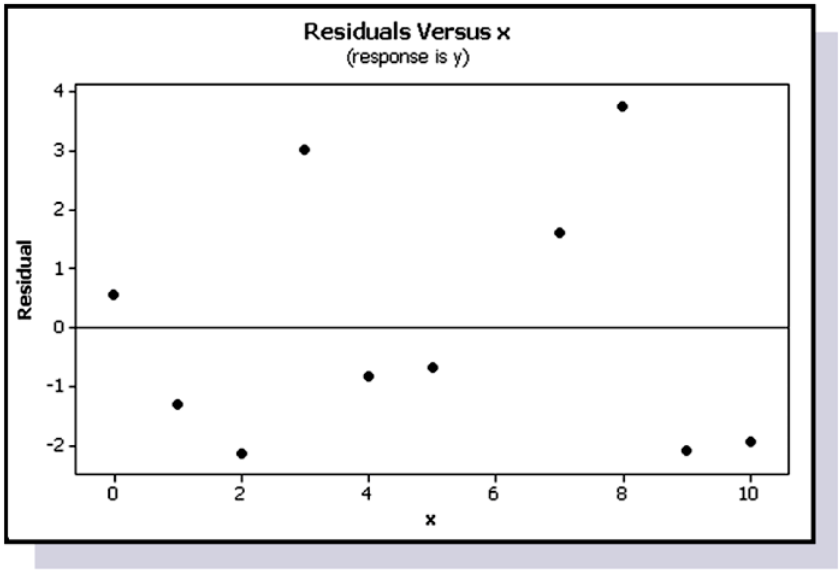
Reziduális diagram

Ha a reziduális diagram nem mutat semmilyen szabályosságot vagy alakzatot, akkor a regressziós egyenlet jól reprezentálja a két változó közti kapcsolatot.

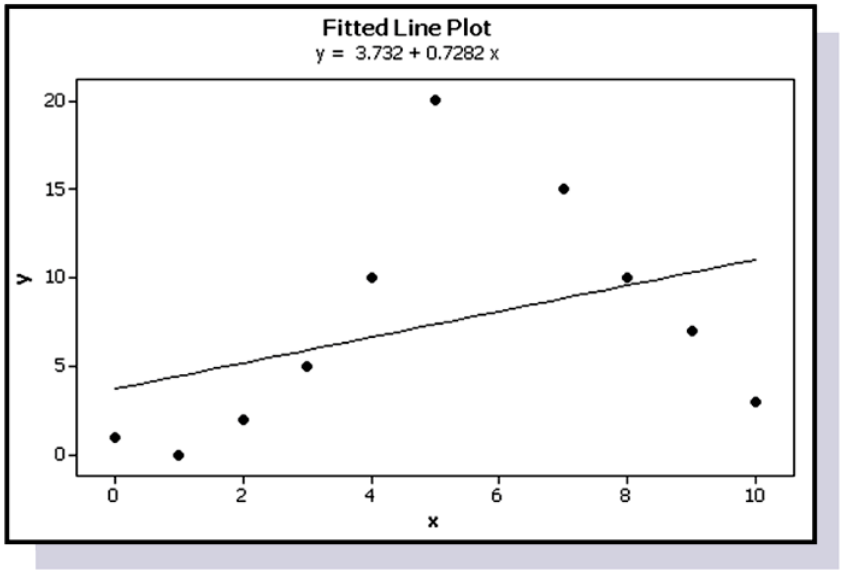
Ha a reziduális diagram valamilyen szabályos mintázatot mutat, akkor a regressziós egyenlet nem jó reprezentáció.

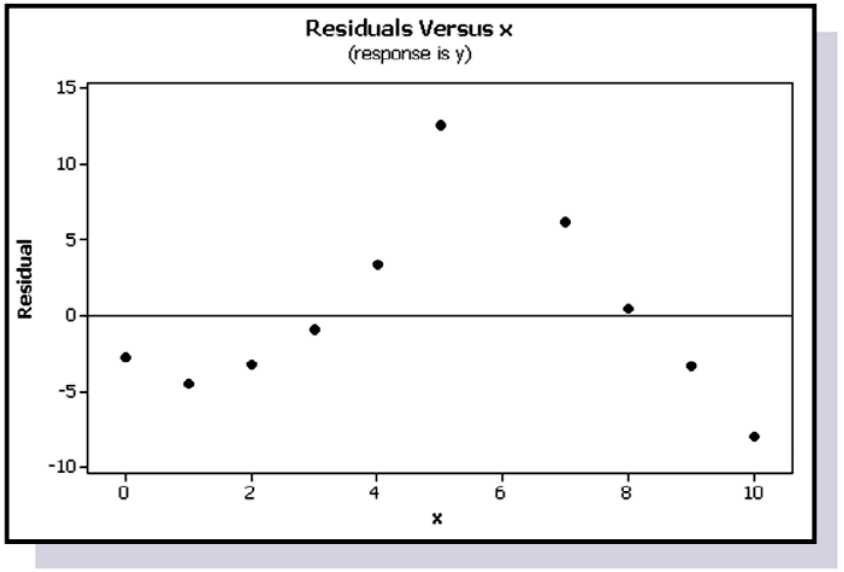
Reziduális diagram



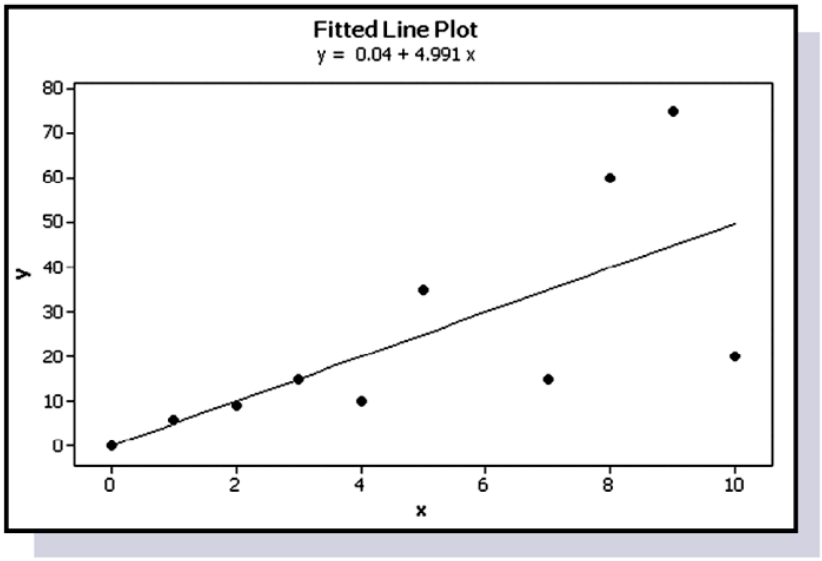


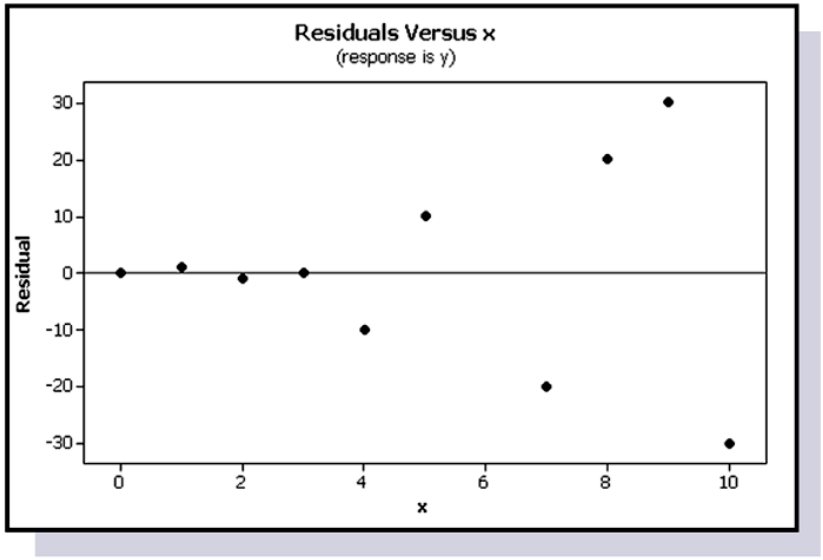
Reziduális diagram





Reziduális diagram





Összefoglalás

Ebben a fejezetben megvitattuk:

- **A regresszió alapjait.**
- **A regressziós egyenes előrejelzésre való használatát.**
- **A regressziós egyenlet interpretálását.**
- **Outlier-eket**
- **Reziduumokat és a legkisebb négyzeteket.**
- **Reziduális diagramokat.**

10-4. fejezet
Variabilitás és predikciós intervallum

Kulcsfogalmak

Ebben a fejezetben a predikciós intervallum megkonstruálásnak módszerét tekintjük át, ami az y értékének egy intervallum becslése.

Definíció

Teljes deviancia (eltérés)

A **teljes deviancia** az (x, y) pont párra vonatkozóan az a függőles $y - \bar{y}$ távolság ami az (x, y) pont és a minta átlagon $\bar{y}$ keresztül húzott vízszintes vonal között van.

Definíció

Magyarázott deviancia

A magyarázott deviancia az a függőleges távolság, ami a becsült $\hat{y}$-érték $\hat{y} - \bar{y}$ távolsága a minta átlagától.

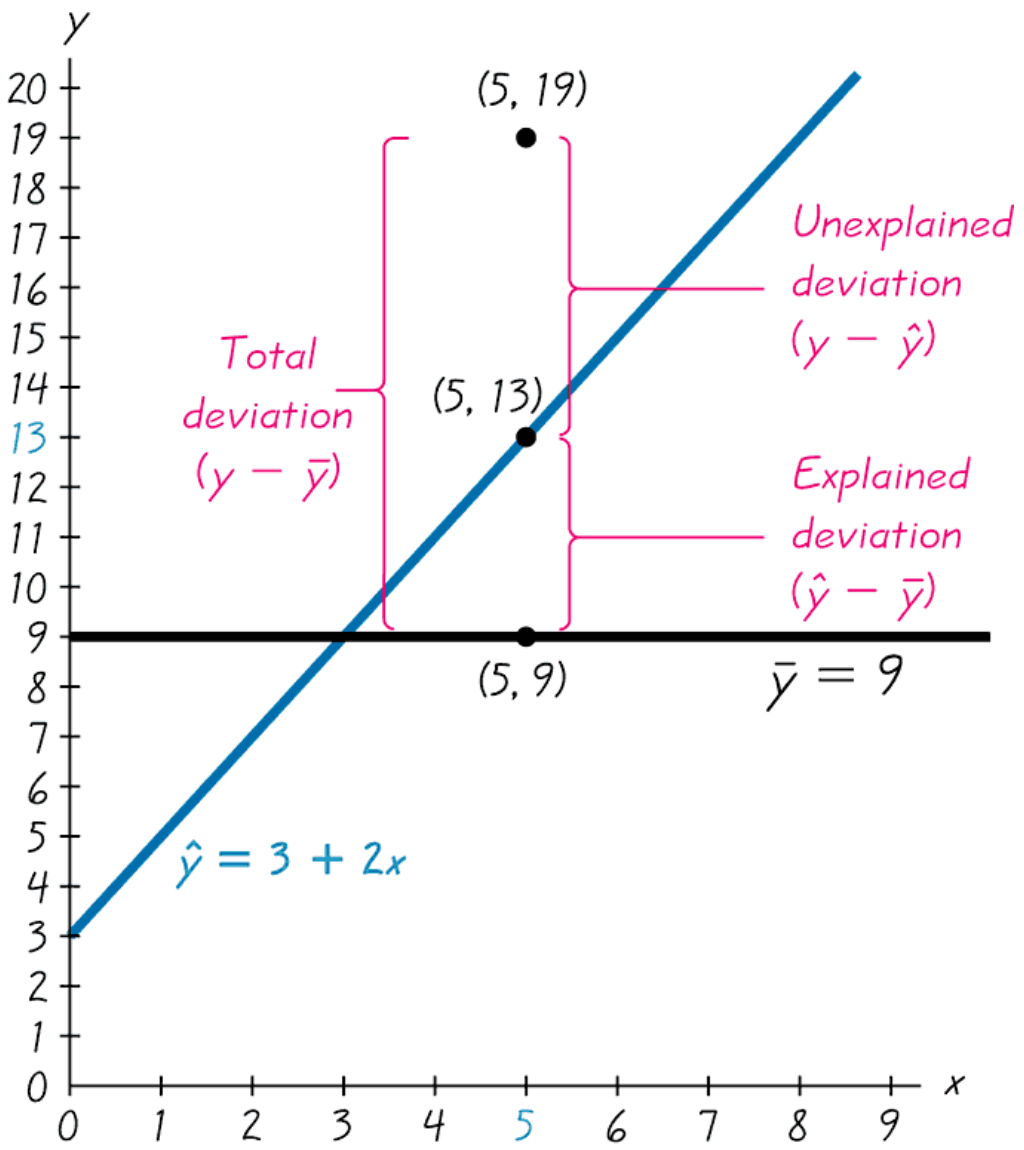
Definíció

Nem magyarázott deviancia

A nem magyarázott (reziduális) deviancia az $y - \hat{y}$ eltérés, ami a becsült és az igazi y érték különbsége. (Reziduumnak neveztük10-3.-ban.)

Nem magyarázott, magyarázott és teljes deviancia

10-9. ábra



Összefüggések

(teljes deviancia) = (magyarázott) + (nem magyarázott)

$$(y - \bar{y}) = (\hat{y} - \bar{y}) + (y - \hat{y})$$

(teljes eltérésnégyzetösszeg) = (magyarázott) + (nem magyarázott)

$$\Sigma (y - \bar{y})^2 = \Sigma (\hat{y} - \bar{y})^2 + \Sigma (y - \hat{y})^2$$

10-4. képlet

Definíció

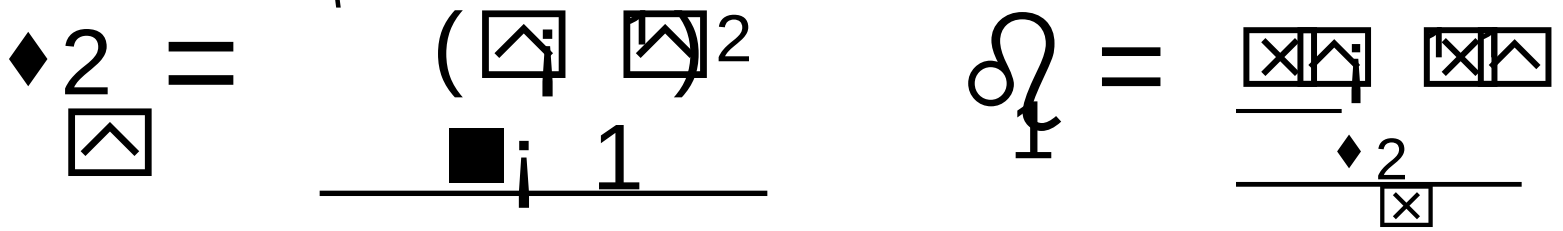
Determinációs együttható

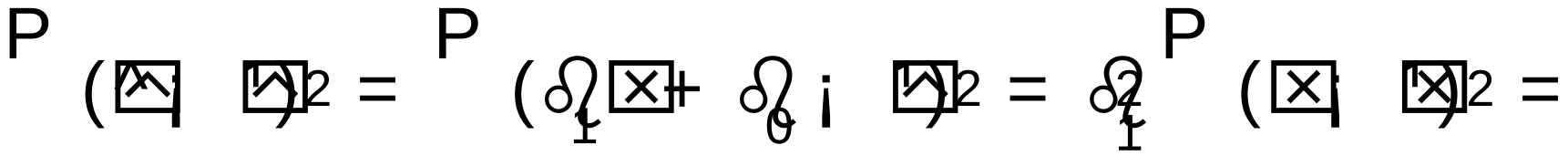
az y variabilitásának az a része, amit a regressziós egyenes megmagyaráz.

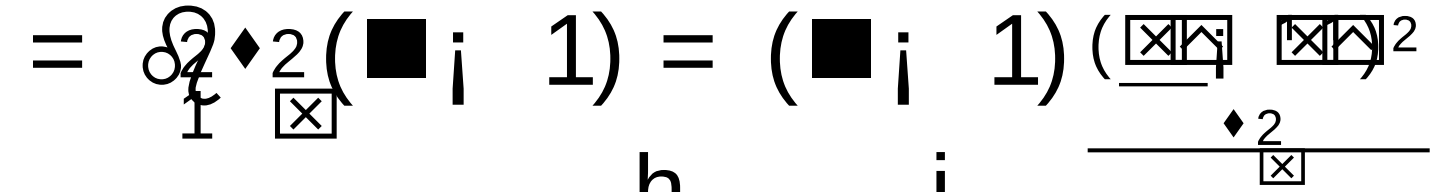
$$r^2 = \frac{\text{magyarázott eltérésnégyzetösszeg.}}{\text{teljes eltérésnégyzetösszeg}}$$

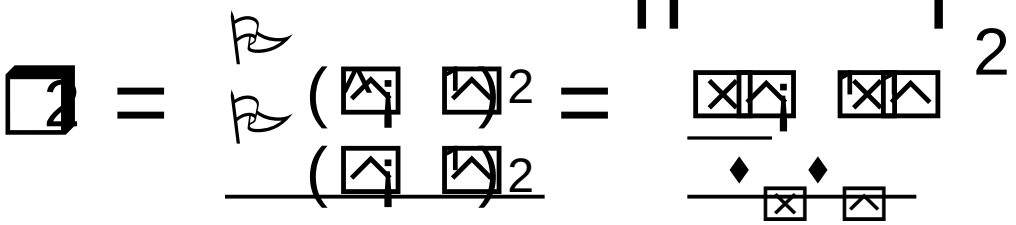
Az r^2 értéke a variblitásnak az a hányada, amit az x és y közti lineáris kapcsolat megmagyaráz

Néhány mellékszámítás:









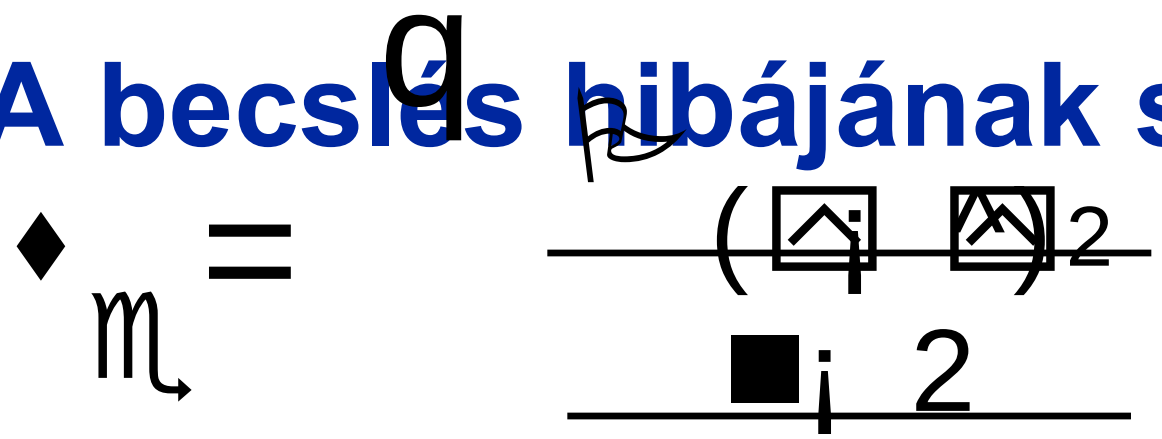
Ez ugyanaz mint a lin. korr. együttható.

Definíció

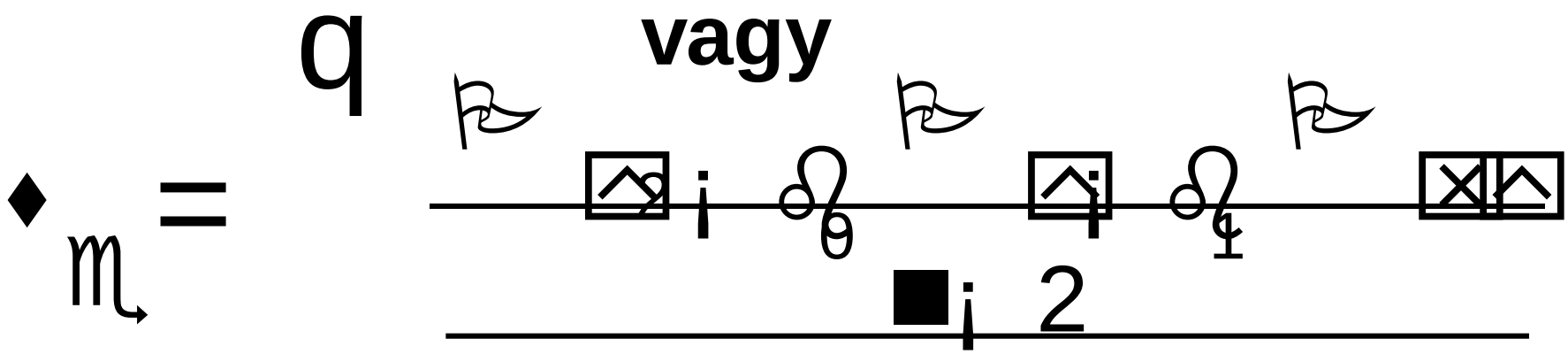
A becslés hibájának szórása

A becslés hibájának szórása, $s_{e,}$ a mérőszáma a minta megfigyelt y értékei és a regressziós egyenes eltérésének.

A becslés hibájának szórása



vagy



Példa: Old Faithful

A 10-1 táblázat adatait használva határozzuk meg a becslés hibájának szórását.

$n = 8$

$\Sigma y^2 = 60{,}204$

$\Sigma y = 688$

$\Sigma xy = 154{,}378$

$b_0 = 34.7698041$

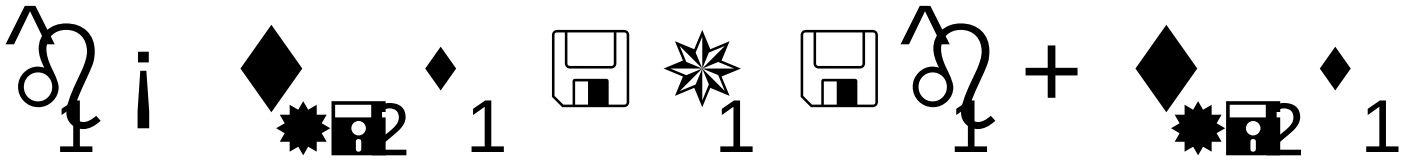
$b_1 = 0.2340614319$

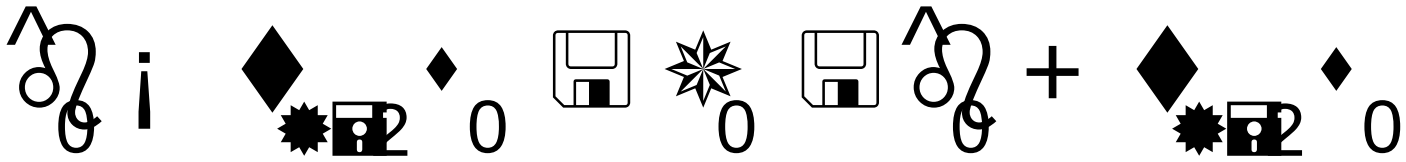
$$s_e = \sqrt{\frac{\Sigma y^2 - b_0 \Sigma y - b_1 \Sigma xy}{n - 2}}$$

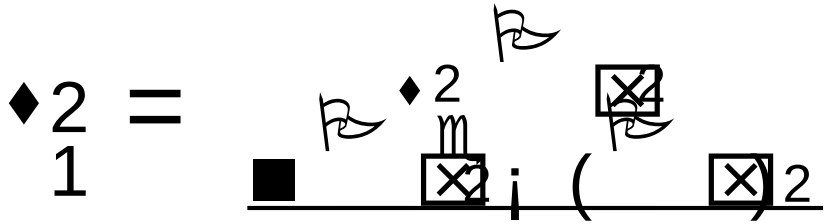
$$s_e = \sqrt{\frac{60{,}204 - (34.7698041)(688) - (0.2340614319)(154{,}378)}{8 - 2}}$$

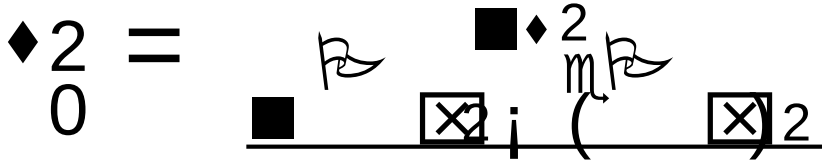
$$= 4.973916052$$

A regressziós paraméterek konfidencia intervallumai









A becslési intervallum egyes y értékekre vonatkozóan

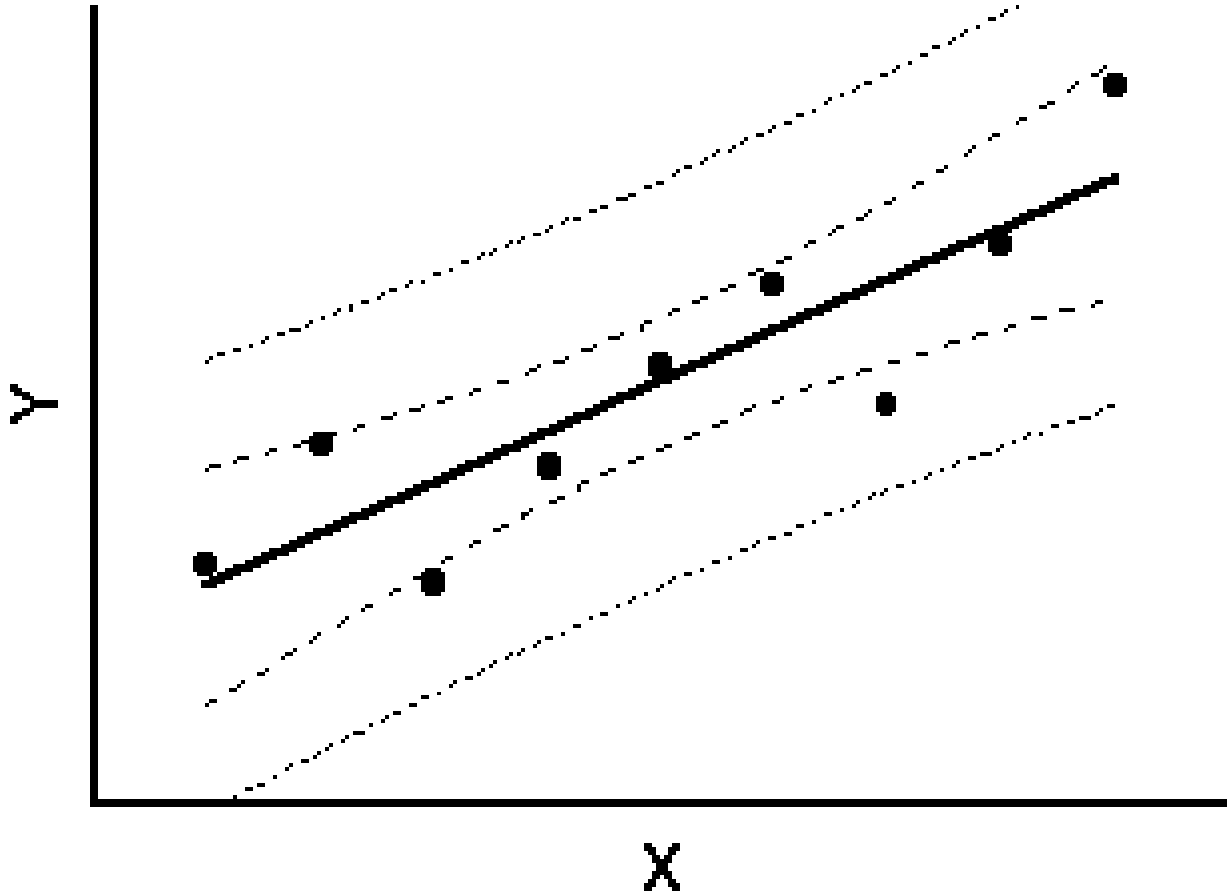
$$\hat{y} - E < y < \hat{y} + E$$

ahol

$$E = t_{\alpha/2} s_e \sqrt{1 + \frac{1}{n} + \frac{n(x_0 - \bar{x})^2}{n(\Sigma x^2) - (\Sigma x)^2}}$$

x_0 az x megadott értéke $t_{\alpha/2}$-nek $n - 2$ szabadsági foka van

Predikciós és konfidencia intervallumok



Példa: Old Faithful

Az 10-1 táblázat adataihoz illesztett egyenes alapján azt találtuk, hogy a 180 sec. hosszúságú kitörés után a legközelebbi kitörés idejére adott becslés 76.9 perc. Adjuk meg a 95%-os becslés intervallumot ehhez az értékhez!

$$E = t_{\alpha/2}\, s_e \sqrt{1 + \frac{1}{n} + \frac{n(x_0 - \overline{x})^2}{n(\Sigma x^2) - (\Sigma x)^2}}$$

$$E = (2.447)(4.973916052)\sqrt{1 + \frac{1}{8} + \frac{8(180 - 218.875)^2}{8(399{,}451) - (1751)^2}}$$

$$E = 13.4 \text{ (kerekítve)}$$

Példa: Old Faithful - folyt

$$\hat{y} - E < y < \hat{y} + E$$

$$76.9 - 13.4 < y < 76.9 + 13.4$$

$$63.5 < y < 90.3$$

Összefoglalás

Ebben a fejezetben foglalkoztunk:

- **Magyarázott és nem magyarázott devianciával.**
- **A determinációs együtthatóval.**
- **A hiba szórásával.**
- **A becslési intervallumokkal.**

10-5. fejezet
Többszörös regresszió

Kulcsfogalmak

Ebben a fejezetben a több mint két változó közti lineáris kapcsolatok elemzési módszerét vizsgáljuk meg.

Három kulcs elemre koncentrálunk:

1. A többszörös regressziós egyenletre.

2. Az adjusztált R^2 értékeire.

3. A P-értékre.

Definíció

Többszörös regressziós egyenlet

Lineáris kapcsolat a válasz változó y és a kettő vagy több prediktor változó között ($x_1, x_2, x_3 \ldots, x_k$)

Általános alakja:

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k.$$

Jelölés

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + \ldots + b_k x_k$$

(Az általános alakja a becsült regressziós egyenletnek)

n = minta méret

k = a prediktor változók száma

$\hat{y}$ = az y becsült értéke

$x_1, x_2, x_3 \ldots, x_k$ a prediktor változók

Jelölések- folyt

β_0 = az y tengelymetszet, azaz az y értéke, amikor minden prediktor változó 0.

b_0 = becslése β_0 –nak a minta alapján

$\beta_1, \beta_2, \beta_3 \ldots, \beta_k$ együtthatók a független változók előtt $x_1, x_2, x_3 \ldots, x_k$

$b_1, b_2, b_3 \ldots, b_k$ a mintabecslései az együtthatóknak $\beta_1, \beta_2, \beta_3 \ldots, \beta_k$

Példa: Old Faithful

A 10-1. táblázat alapján keressük meg a többszörös regressziós egyenletet, ahol a válasz változó (y) a kitörés után eltelő idő, és a prediktor változók (x) a kitörés hossza és magassága.

Az együtthatók megkeresését számítógépes csomagok (pl. Excel) végzik …

Példa: Old Faithful - folyt

The regression equation is
Interval After = 45.1 + 0.245 Duration - 0.098 Height

Predictor	Coef	SE Coef	T	P
Constant	45.10	19.41	2.32	0.068
Duration	0.24464	0.04486	5.45	0.003
Height	-0.0983	0.1623	-0.61	0.571

S = 5.25937 **R-Sq = 86.7%** **R-Sq(adj) = 81.3%**

Analysis of Variance

Source	**DF**	**SS**	**MS**	**F**	**P**
Regression	2	897.69	448.85	16.23	**0.007**
Residual Error	5	138.31	27.66		
Total	7	1036.00			

Példa: Old Faithful - folyt

Eredmény:

$$\text{Utána} = 45.1 + 0.245\ \text{időtartam} - 0.098\ \text{magasság}$$

Vagy:

$$y = 45.1 + 0.245\, x_1 - 0.098 x_2$$

Definíció

- **Többszörös determinációs együttható**

  A többszörös determinációs együttható R^2 annak a mérőszáma, hogy mennyire illik a többszörös regressziós egyenlet a mintaadatokhoz.

- **Korrigált többszörös determinációs együttható**

  A korrigált többszörös determinációs együttható az előző R^2 olyan korrekciója, amely figyelembe veszi a változók számát és a minta méretét is.

Korrigált R^2

$$\text{Korrigált } R^2 = 1 - \frac{(n-1)}{[n-(k+1)]}(1-R^2)$$

10-6. képlet

ahol n = minta elemszáma

k = a független (x) változók száma

A legjobb többszörös regressziós egyenlet megkeresése

1. Használd a józan eszedet arra, hogy kiválaszd a fontos és a nem fontos változókat.
2. Vedd figyelembe a *P*-értéket. Válassz olyan egyenletet, aminek nagy a szignifikanciája a számítógép által által adott *P*-értékek szerint.
3. Használd a nagy korrigált R^2 –tel rendelkező egyenleteket és csak kevés változót vegyél be.
   - Ha egy újabb prediktor változót veszel be és a korrigált R^2 nem növekszik lényegesen.
   - Adott számú prediktor (x) változó használata esetén használd a legnagyobb korrigált R^2-et adó változókat.
   - Hogy kidobáljuk a felesleges (x) változókat, amelyeknek nincs nagy hatásuk y-ra, segíthet a változók közti lineáris korrelációs együttható ismerete.

Összefoglalás

Ebben a fejezetben megviattuk:

- **A többszörös regresszió egyenleteit.**
- **Korrigált R^2-et.**
- **A legjobb többszörös regressziós egyenlet megkeresését.**

10-6. fejezet
Modellezés

Kulcsfogalmak

Ebben a fejezetben bemutatjuk annak a részleteit, hogyan illeszthetünk matematikai modellt az adatainkhoz.

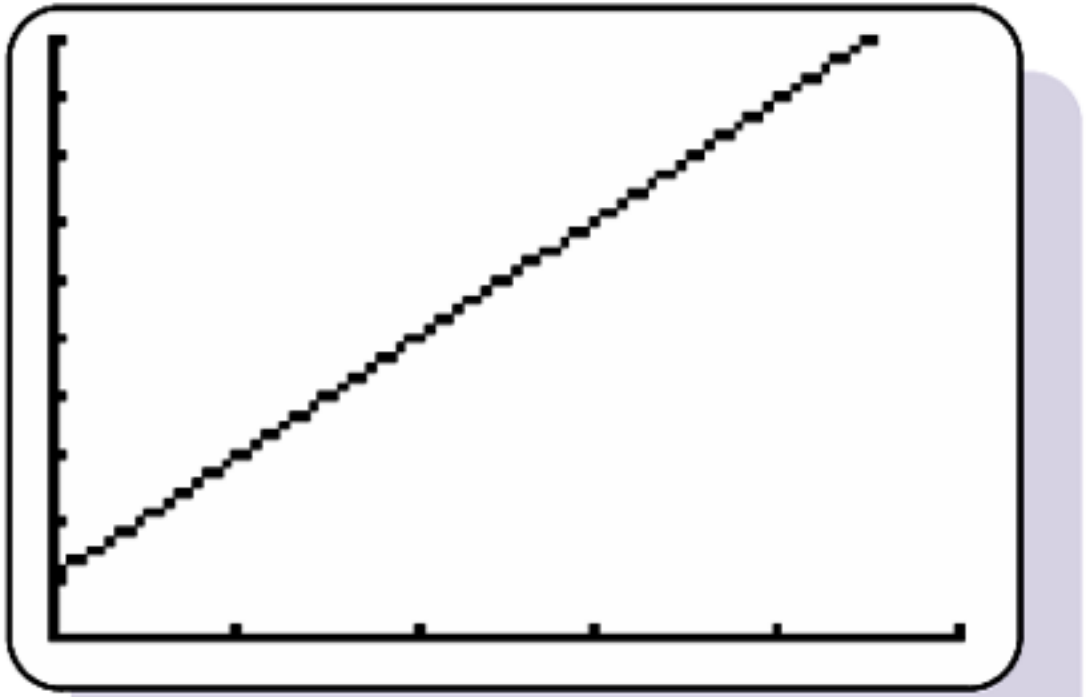
Ezt a folyamatot nemlineáris regressziónak is nevezik.

Példák

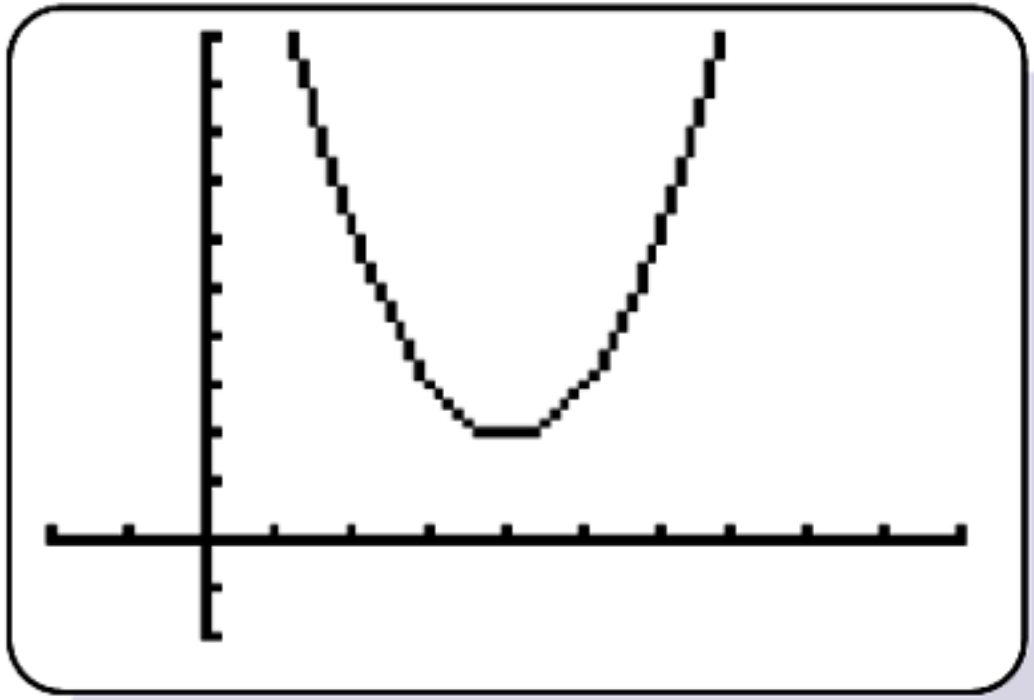
- **Lineáris:** $y = a + bx$
- **Kvadratikus:** $y = ax^2 + bx + c$
- **Logaritmikus:** $y = a + b \ln x$
- **Exponenciális:** $y = ab^x$
- **Hatvány:** $y = ax^b$

Illusztrációk:

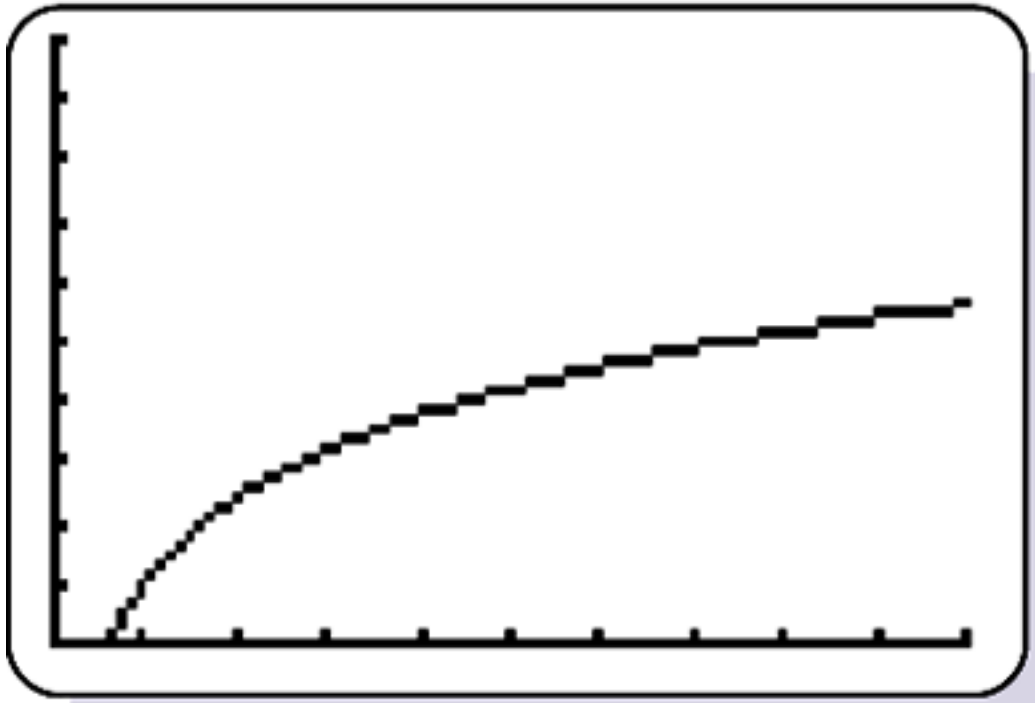
Linear: $y = 1 + 2x$



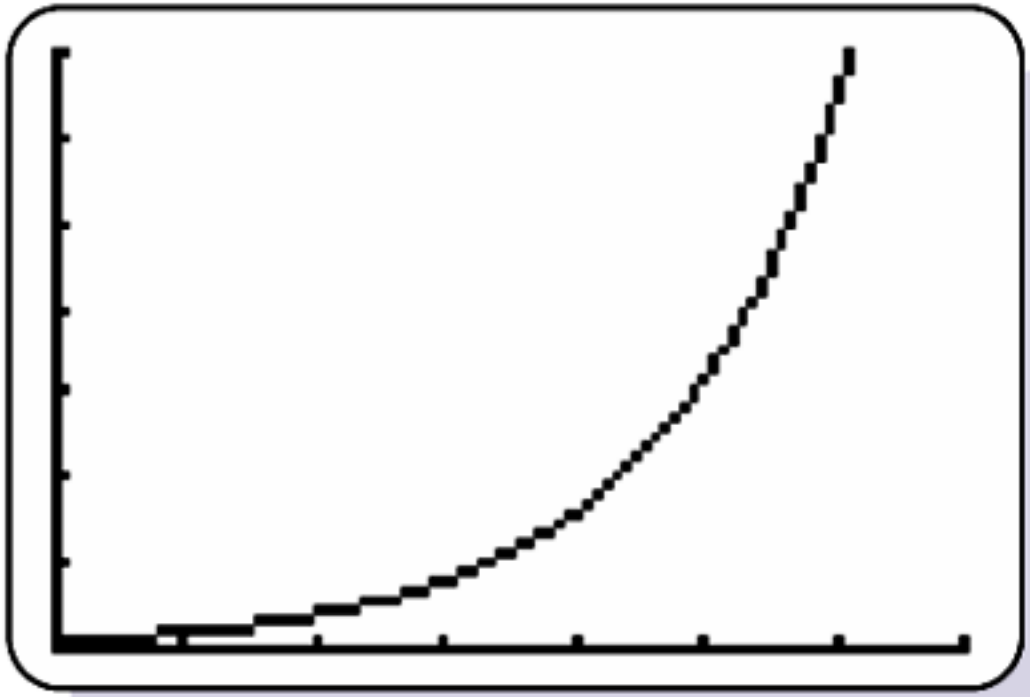
Quadratic: $y = 2x^2 - 8x + 9$



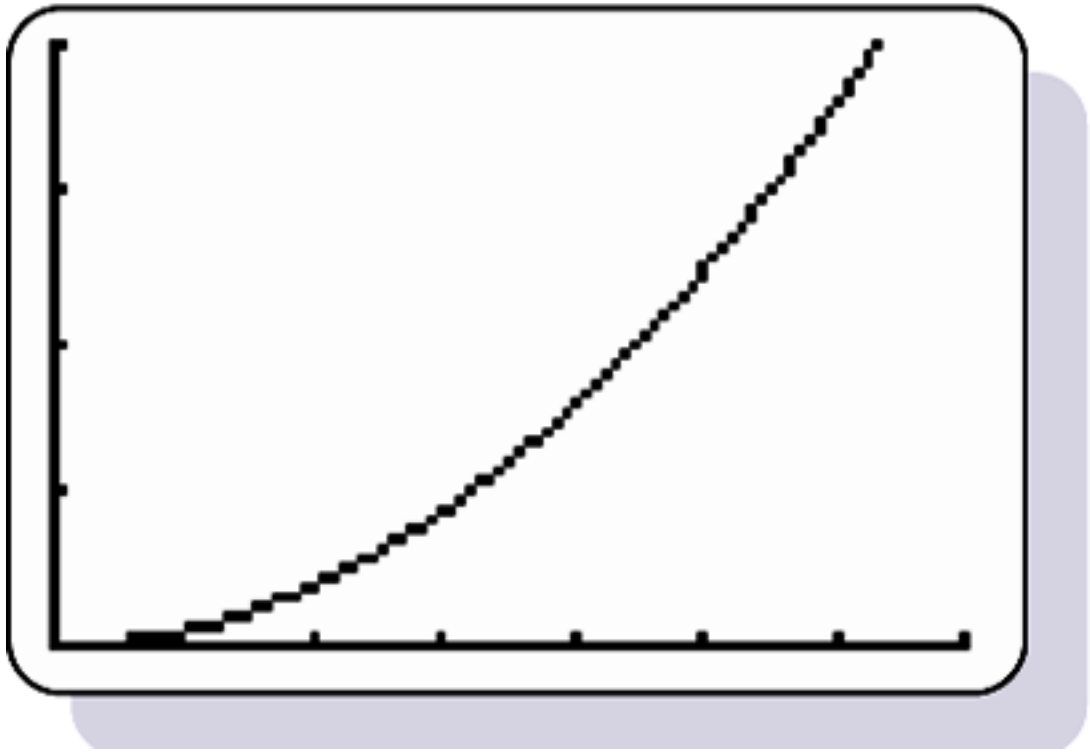
Logarithmic: $y = 1 + 2\ln x$



Exponential: $y = 2^x$

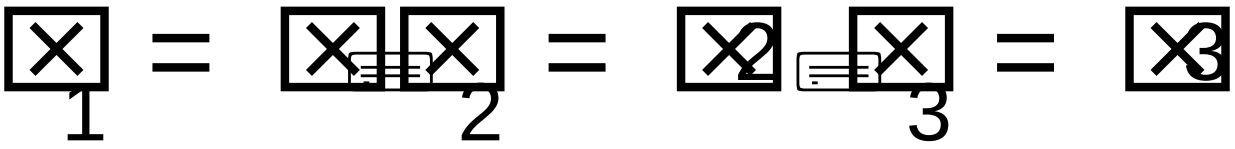


Power: $y = x^2$

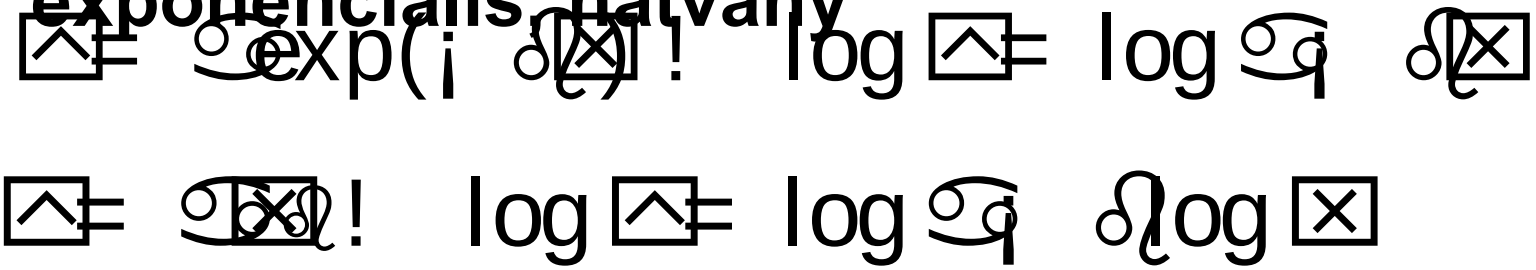


Nemlineáris visszavezetése lineárisra

- **1, Polinom illesztés: visszavezethető lineárisra**

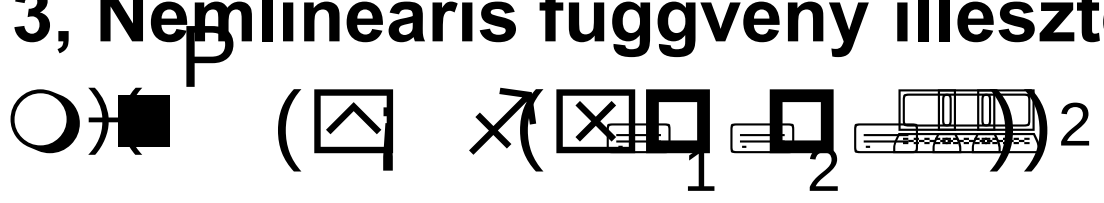


- **2, Transzformációval visszavezethetők: exponenciális, hatvány**



folyt

- **3, Nemlineáris függvény illesztése:**



A jó modell (illesztő függvény) megkeresése

- **Keresd az adathalmazban a szabályosságot:** Nézegesd az ábrát és próbáld meg kitalálni, milyen függvényt követnek az adatok.
- **Számítsd ki R^2 -et** és keress olyan függvényeket, amelyek minél nagyobb R^2 –et adnak, mivel ez azt jelenti, hogy azok jobban illenek az adatokhoz.
- **Gondolkozz:** Zárd ki a nem realisztikus modelleket, melyek hibás következtetésekre vezetnek.

Összefoglalás

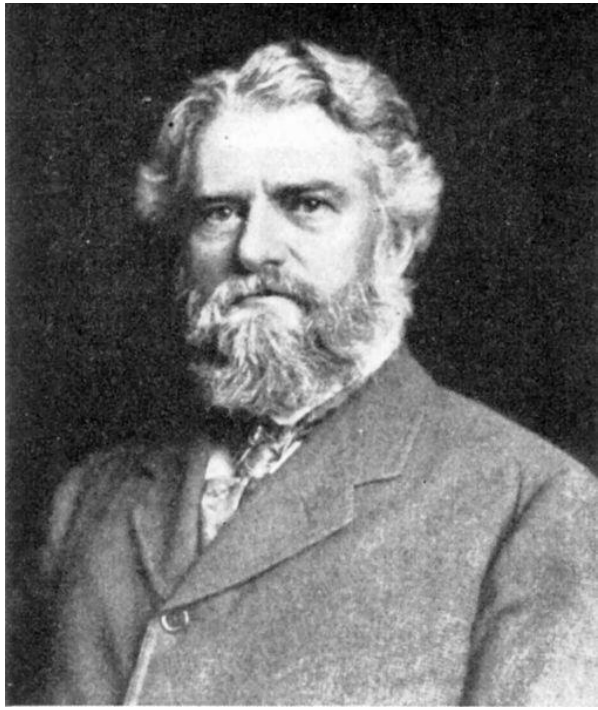
Ebben a fejezetben megvitattuk:

- **A nemlineáris regressziót.**
- **Néhány jó tanácsot.**

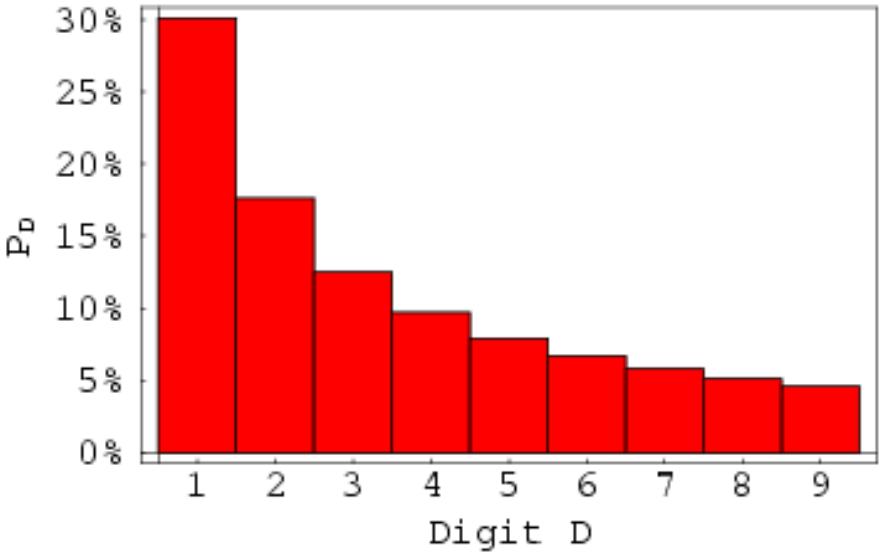
Az első számjegyek Branford törvénye

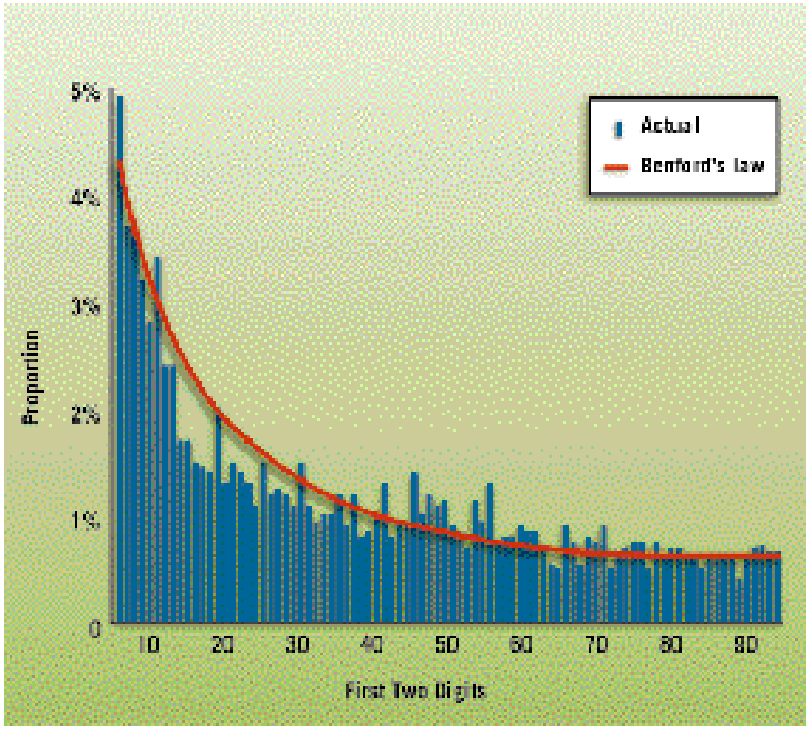
Frank Benford (1883-1948)
A General Electric fizikusa

Simon Newcomb (1835 – 1909)
asztronómus



$$P(D) = \frac{\log_{10}(1 + 1/D)}{\log_{10}}$$





A híres arizonai csekk sikkasztási eset

The table lists the checks that a manager in the office of the Arizona State Treasurer wrote to divert funds for his own use. The vendors to whom the checks were issued were fictitious.

Date of Check	Amount
October 9, 1992	$ 1,927.48
	27,902.31
October 14, 1992	86,241.90
	72,117.46
	81,321.75
	97,473.96
October 19, 1992	93,249.11
	89,658.17
	87,776.89
	92,105.83
	79,949.16
	87,602.93
	96,879.27
	91,806.47
	84,991.67
	90,831.83
	93,766.67
	88,338.72
	94,639.49
	83,709.28
	96,412.21
	88,432.86
	71,552.16
TOTAL	$ 1,878,697.58

http://www.aicpa.org/pubs/jofa/may1999/nigrini.htm

11. előadás
Multinomiális kísérletek és kontingencia táblák

11-1 Áttekintés

11-2 Multinomiális kísérletek: az illeszkedés jósága

11-3 Kontingencia táblák: Függetlenség és homogenitás

11-1 & 11-2 fejezetek Áttekintés és multinomiális kísérletek: az illeszkedés jósága

Áttekintés

- **Kategoriális adatokkal foglalkozunk, vagy olyan kvantitatív adatokkal, amelyeket különböző kategóriákba lehet sorolni (gyakran bineknek vagy celláknak hívjuk).**
- **A χ^2 (khí-négyzet) teszt statisztika.**
- **Az illeszkedés vizsgálat (goodness of fit test) egy egydimenziós gyakorisági táblázat (egy sor vagy oszlop).**
- **A kontingencia tábla egy kétdimenzós gyakorisági táblázat (kettő vagy több oszlop és sor).**

Kulcsfogalmak

Adott, kategóriákba sorolt adatok esetén azt a hipotézist teszteljük, hogy az adatok eloszlása megegyezik valamilyen általunk feltételezett eloszlással.

A hipotézis teszt a khí-négyzet eloszlást használja a megfigyelt gyakoriságok és az általunk várt gyakoriságok összehasonlítására.

Definíció

Multinomiális kísérlet

Egy olyan kísérlet, ami az alábbi feltételeknek tesz eleget:

1. **A próbálkozások/kísérletek száma előre adott.**

2. **A próbálkozások/kísérletek függetlenek.**

3. **A kísérlet minden kimenetele egyértelműen besorolható pontosan egybe a lehetséges kategóriák közül.**

4. **A kísérletek során a kategóriák valószínűsége nem változik, állandó marad.**

Példa: A tömegek utolsó számjegye

Amikor az embereket megkérdezik, hogy mekkora a tömegük, gyakran mondanak a valóságosnál kisebb értékeket. Hogyan lehet eldönteni egy adathalmazról, hogy igazi mérésből származnak, vagy az emberek megkérdezéséből nyert értékek?

Példa: A tömegek utolsó jegye

Teszteljük azt a feltevést, hogy az 11-2. táblázatban található értékek ugyanazzal a gyakorisággal lépnek fel.

11-2. táblázat összesítés 80 hallgató tömegének utolsó számjegyei

Table 11-2
Last Digits of Weights

Last Digit	Frequency
0	35
1	0
2	2
3	1
4	4
5	24
6	1
7	4
8	7
9	2

Példa: folyt.

Ellenőrizzük, hogy a multinomiális kísérlet feltételei fennállnak-e.

1. **A kísérletek száma adott, 80.**
2. **A kísérletek függetlenek, mert valaki tömegének utolsó számjegye nincs hatással valaki más tömegének utolsó számjegyére.**
3. **Minden kimenet (utolsó számjegy) pontosan egy kategóriába sorolható. A kategóriák 0, 1, … , 9.**
4. **Végül, pedig nem változik a kimenetek valószínűsége a kísérlet során.**

Definíció

Illeszkedés vizsgálat

Az illeszkedés vizsgálatot annak tesztelésére használjuk, hogy a megfigyelt gyakoriságok illeszkednek a feltételezett gyakoriság eloszláshoz.

Illeszkedés vizsgálat

Jelölések

O jelöli egy kimenetel megfigyelt gyakoriságát.

E jelöli egy kimenetel várt gyakoriságát.

k jelöli a lehetséges kimenetek/kategóriák számát.

n jelöli a kísérletek teljes számát.

Várt gyakoriságok

Ha minden gyakoriság egyenlő:

$$E = \frac{n}{k}$$

az összes megfigyelt előfordulások száma elosztva a kategóriák számával

Várt gyakoriságok

Ha nem mindegyik gyakoriság egyforma:

$$E = n\,p$$

Meg kell szorozni a kategória valószínűséget az összes esetek számával.

Illeszkedés vizsgálat (teszt) multinomiális kísérletekben

Követelmények

1. **Az adatokat véletlenül választjuk ki**
2. **A minta gyakoriság adatokból áll minden kategóriára vonatkoztatva.**
3. **Minden kategóriában legalább 5 legyen a várt megfigyelések száma! (A várt gyakoriság az, amit a feltételezésünk alapján várunk. A megfigyelt esetek számának nem kell legalább 5-nek lennie.)**

Illeszkedés vizsgálat (teszt) multinomiális kísérletekben

Teszt statisztika

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

Kritikus értékek

1. **A khí-négyzet táblázatot kell használnunk**

k – 1 szabadsági fokok számával, ahol k = a kategóriák száma.

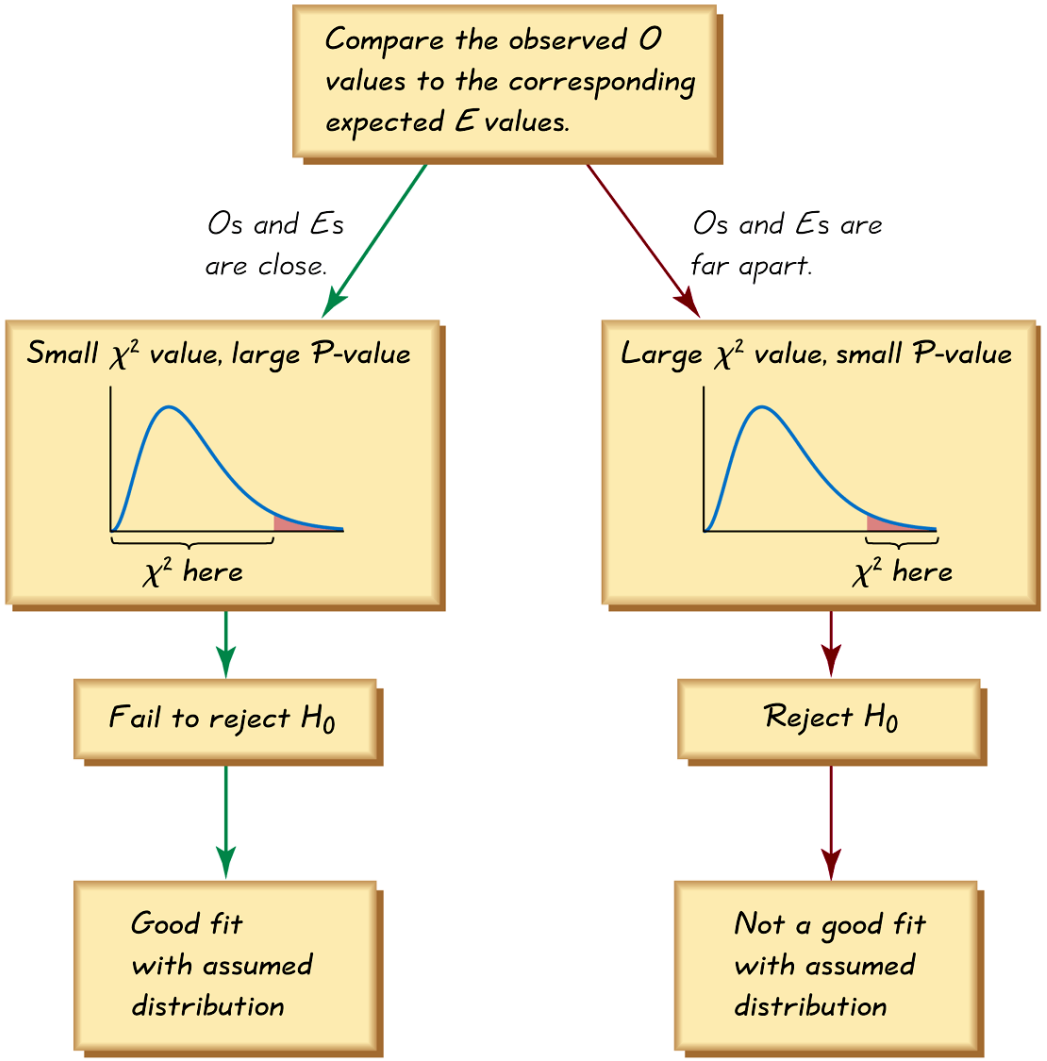
2. **Az illeszkedés vizsgálatok mindig jobboldali tesztek.**

Illeszkedés vizsgálat (teszt) multinomiális kísérletekben

- A **közeli egyezés** a megfigyelt és a várt értékek között kicsi χ^2 és nagy P-értékre vezetnek.
- A **nagy eltérés** a megfigyelt és a várt értékek között nagy χ^2 és kis P-értékre vezetnek.
- Egy **szignifikánsan nagy** χ^2 érték a null hipotézis **elutasítását** fogja okozni, amennyiben a null hipotézis szerint nincs különbség a megfigyelt és a várt gyakoriságok között.

Kapcsolat a χ^2 teszt statisztika, P-érték, és az illeszkedés vizsgálat között

11-3. ábra



Példa: utolsó számjegy elemzés

Teszteljük azt a feltevést, hogy a 11-2. táblázatban a számjegyek nem ugyanazzal a gyakorisággal fordulnak elő.

$H_0: p_0 = p_1 = \ldots = p_9$

H_1: Legalább az egyik vsz. különbözik a többitől.

$\alpha = 0.05$

$k - 1 = 9$

$\chi^2_{.05,\,9} = 16.919$

Table 11-2
Last Digits of Weights

Last Digit	Frequency
0	35
1	0
2	2
3	1
4	4
5	24
6	1
7	4
8	7
9	2

Példa: utolsó számjegy elemzés

Teszteljük azt a feltevést, hogy a 11-2. táblázatban a számjegyek nem ugyanazzal a gyakorisággal fordulnak elő.

Ha a 80 számjegy egyenletesen oszlana el a 10 kategória között, akkor minden gyakoriságra 8-at várunk.

Table 11-2
Last Digits of Weights

Last Digit	Frequency
0	35
1	0
2	2
3	1
4	4
5	24
6	1
7	4
8	7
9	2

Példa: utolsó számjegy elemzés

Teszteljük azt a feltevést, hogy a 11-2. táblázatban a számjegyek nem ugyanazzal a gyakorisággal fordulnak elő.

Table 11-3 Calculating the χ^2 Test Statistic for the Last Digits of Weights

Last Digit	Observed Frequency O	Expected Frequency E	$O - E$	$(O - E)^2$	$\frac{(O - E)^2}{E}$
0	35	8	27	729	91.1250
1	0	8	−8	64	8.0000
2	2	8	−6	36	4.500
3	1	8	−7	49	6.125
4	4	8	−4	16	2.000
5	24	8	16	256	32.000
6	1	8	−7	49	6.125
7	4	8	−4	16	2.000
8	7	8	−1	1	0.125
9	2	8	−6	36	4.500
	80	80			$\chi^2 = \sum \frac{(O - E)^2}{E} = 156.500$

(Except for rounding errors, these two totals must agree.)

Példa: utolsó számjegy elemzés

Teszteljük azt a feltevést, hogy a 11-2. táblázatban a számjegyek nem ugyanazzal a gyakorisággal fordulnak elő.

A 11-3. táblázat szerint, a teszt statisztika értéke $\chi^2 = 156.500$.

Mivel a kritikus éték 16.919, elutasítjuk a null hipotézist, amely szerint a valószínűségek megegyeznek.

Elegendő evidencia van arra, hogy támogassuk azt a feltevést, hogy az utolsó számjegyek nem mind ugyanakkora gyakorisággal fordulnak elő.

Példa: Csalás detektálás

11-1. táblázat: Az első számjegyek statisztikája és a Brenford szabály.

Table 11-1 Benford's Law: Distribution of Leading Digits

Leading Digit	1	2	3	4	5	6	7	8	9
Benford's law: frequency distribution of leading digits	30.1%	17.6%	12.5%	9.7%	7.9%	6.7%	5.8%	5.1%	4.6%
Expected frequencies of leading digits from 784 checks following Benford's law	235.984	137.984	98.000	76.048	61.936	52.528	45.472	39.984	36.064
Observed leading digits of 784 actual checks analyzed for fraud	0	15	0	76	479	183	8	23	0

Példa: Csalás detektálás

Teszteljük azt a feltevést, hogy szignifikáns eltérés van a Brenford szabály és a 784 db számla első számjegye között.

Observed Frequencies and Frequencies Expected with Benford's Law

Digit	Observed Frequency	Expected Frequency	$O - E$	$(O - E)^2$	$\frac{(O - E)^2}{E}$
1	0	235.984	−235.984	55688.4483	235.9840
2	15	137.984	−122.984	15125.0643	109.6146
3	0	98.000	−98.000	9604.0000	98.0000
4	76	76.048	−0.048	0.0023	0.0000
5	479	61.936	417.064	173942.3801	2808.4213
6	183	52.528	130.472	17022.9428	324.0737
7	8	45.472	−37.472	1404.1508	30.8795
8	23	39.984	−16.984	288.4563	7.2143
9	0	36.064	−36.064	1300.6121	36.0640
				Total: $\chi^2 = \sum \frac{(O - E)^2}{E}$	= 3650.2514

Példa: Csalás detektálás

Teszteljük azt a feltevést, hogy szignifikáns eltérés van a Brenford szabály és a 784 db számla első számjegye között.

H_0: $p_1 = 0.301$, $p_2 = 0.176$, $p_3 = 0.125$, $p_4 = 0.097$, $p_5 = 0.079$, $p_6 = 0.067$, $p_7 = 0.058$, $p_8 = 0.051$ and $p_9 = 0.046$

H_1: Legalább egy gyakoriság eltér ezektől az arányoktól.

$\alpha = 0.01$

$k - 1 = 8$

$\chi^2_{.01,8} = 20.090$

Példa: Csalás detektálás

Teszteljük azt a feltevést, hogy szignifikáns eltérés van a Brenford szabály és a 784 db számla első számjegye között.

A teszt statisztika értéke $\chi^2 = 3650{,}251$.

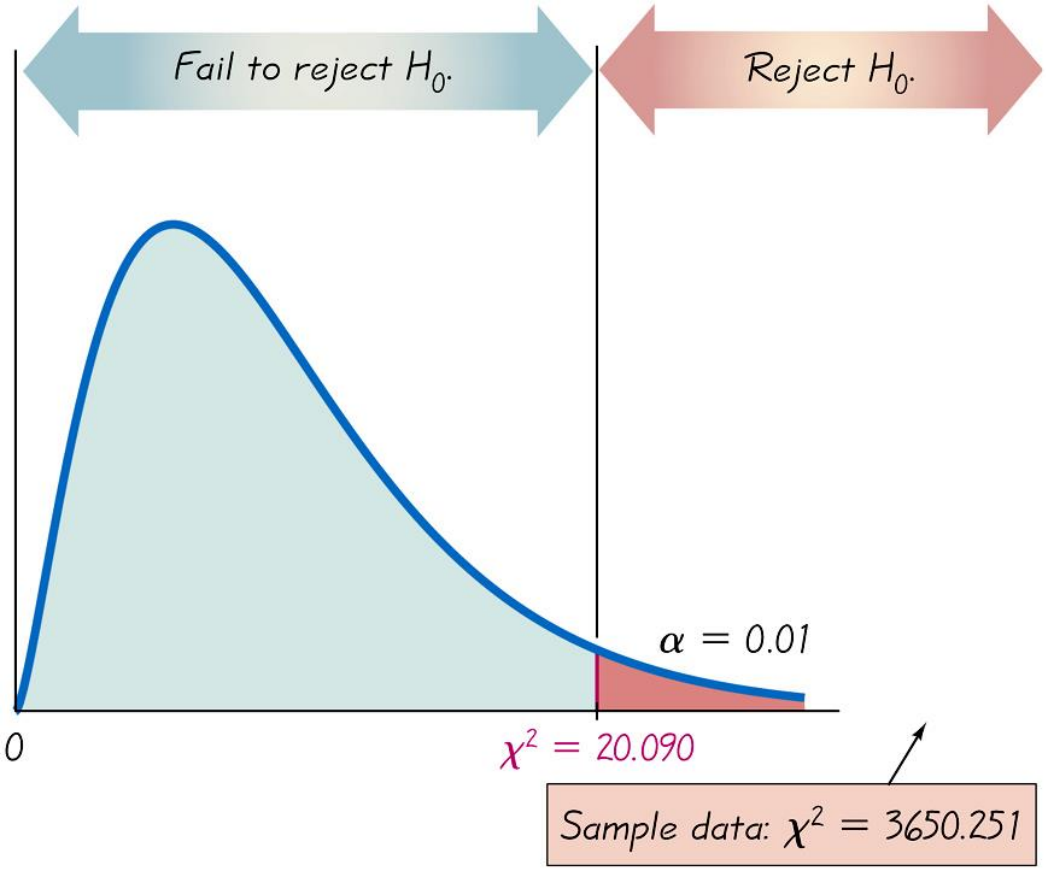
Mivel a kritikus érték 20,090 , elutasítjuk a null hipotézist.

Elég bizonyíték van a null hipotézis elutasítására - Elég bizonyíték van arra, hogy legalább az egyik arány eltér a várhatótól.

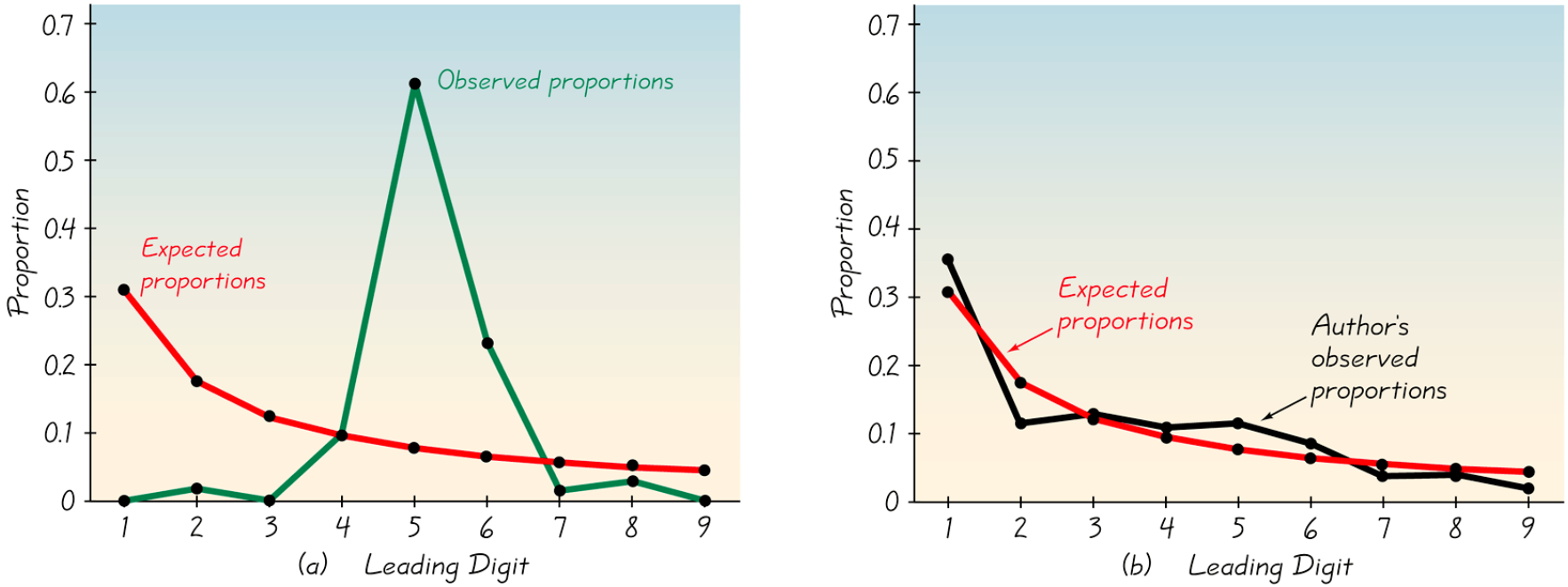
Példa: Csalás detektálás

Teszteljük azt a feltevést, hogy szignifikáns eltérés van a Brenford szabály és a 784 db számla első számjegye között.

11-5. ábra



Példa: Csalás detektálás



11-6. ábra A megfigyelt és a Brenford törvénynek megfelelő első számjegy eloszlások

Összefoglalás

Ebben a fejezetben megbeszéltük:

Multinomiális kísérletek: Illeszkedés jósága

Annak a hipotézisnek a tesztelése, hogy a megfigyelt gyakoriság eloszlás illeszkedik a feltételezett eloszláshoz.

11-3. fejezet
Kontingencia táblázatok: Függetlenség és homogenitás

Kulcsfogalmak

Ebben a fejezetben kontingencia vagy más néven két dimenziós gyakorisági táblázatokkal foglalkozunk.

Olyan eljárást mutatunk be, amivel vizsgálni lehet, hogy a sor és az oszlop változók függetlenek-e egymástól.

A homogenitás vizsgálatára ugyanezt módszert használjuk, amellyel eldönthető, hogy különböző populációkban valamilyen tulajdonság ugyanolyan megoszlásban van-e jelen.

Definíció

Kontingencia táblázat

(vagy kétdimenziós gyakorisági táblázat)

Egy kontingencia táblázat olyan táblázat, melyekben a gyakoriságok két változóhoz tartoznak.

(Az egyik változó kategorizálja az oszlopokat, a másik a sorokat.)

A kontingencia táblázatok minimum 2 X 2-esek.

Esettanulmány motorosokról

A bukósisak színe és a baleseti sérülések között van-e valamilyen kapcsolat?

	Fekete	Fehér	Sárga/Narancs	Sorösszeg
Kontroll (nem sérült)	**491**	**377**	**31**	**899**
Balesetes (sérült v. meghalt)	**213**	**112**	**8**	**333**
Oszlopösszeg	**704**	**489**	**39**	**1232**

Definíció

Függetlenség vizsgálat (teszt)

A függetlenség vizsgálat azt a null hipotézist teszteli, hogy nincs kapcsolat az oszlop és a sor változó között a kontingencia táblában. A null hipotézis az, hogy a „sor és oszlop változók függetlenek”.

Követelmények

1. **A minta adatokat véletlenül választjuk ki és két dimenziós gyakorisági táblázatban helyezzük el.**

2. **A null hipotézis H_0 az, hogy a sor és oszlop változók függetlenek; az alternatív hipotézis H_1 az, hogy az oszlop és sor változók függenek egymástól.**

3. **A kontingencia táblában minen várható gyakoriság E legalább 5. (Nem feltétel, hogy a megfigyelt esetek száma legalább 5 legyen. Nem feltétel, hogy a populáció normális eloszlású legyen.)**

Függetlenségi teszt

Teszt statisztika

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

Kritikus értékek

1. **A khí-négyzet eloszlás táblázatából**

   szabadsági fokok száma = $(r - 1)(c - 1)$

   r a sorok, c az oszlopok száma

2. **A függetlenségi teszt mindig jobboldali.**

Feltételezett/várható gyakoriság

$$E = \frac{(\text{sor összeg})\,(\text{oszlop összeg})}{(\text{összes eset})}$$

A megfigyelt gyakoriságok teljes száma az egész táblázatban

Függetlenségi teszt

Ez a procedúra nem alkalmas arra, hogy direkt ok-okozati kapcsolatot mutassunk ki a változók között.

A függőség csak azt jelenti, hogy kapcsolat van a két változó között.

A kontingencia tábla várható gyakorisága

$$E = \text{összes eset} \cdot \frac{\text{sorösszeg}}{\text{összes eset}} \cdot \frac{\text{oszlopösszeg}}{\text{összes eset}}$$

$$E = n \cdot p$$

(cella valószínűség)

$$E = \frac{(\text{sorösszeg})\,(\text{oszlopösszeg})}{(\text{összes eset})}$$

Eset tanulmány

	Fekete	Fehér	Sárga/Narancs	Sorösszeg
Kontroll (nem sérült)	491	377	31	899
Balesetes	213	112	8	333
Oszlopösszeg	704	489	39	1232

A bal felső cellára:

$$E = \frac{(\text{sorösszeg})\,(\text{oszlopösszeg})}{(\text{összes eset})}$$

$$E = \frac{(899)(704)}{1232} = 513.714$$

Esettanulmány

	Fekete	Fehér	Sárga/Narancs	Sorösszeg
Kontroll	491	377	31	899
Várt esetszám	513.714			
Balesetes	213	112	8	333
Várt esetszám				
Oszlopösszeg	704	489	39	1232

$$E = \frac{(\text{sorösszeg})\,(\text{oszlopösszeg})}{(\text{összes eset})}$$

$$E = \frac{(899)(704)}{1232} = 513.714$$

Esettanulmány

	Fekete	Fehér	Sárga/Narancs	Sorösszeg
Kontroll	**491**	**377**	**31**	**899**
Várt	**513.714**	**356.827**	**28.459**	
Balesetes	**213**	**112**	**8**	**333**
Várt	**190.286**	**132.173**	**10.541**	
Oszlopösszeg	**704**	**489**	**39**	**1232**

Kiszámítottuk a várható esetszámot.

A bal felső cella interpretálása: azt mondhatjuk, hogy 491 fekete sisakos motoros sérült meg, de 513.714 lenne a várható szám, ha a sérülések függetlenek lennének a sisak színétől.

folyt.

A 0.05 szignifikancia szintet használva teszteljük azt a feltevést, hogy a csoport (kontroll vagy balesetes) független a sisak színétől.

H_0: Az, hogy valaki a kontroll vagy a balesetes csoportba esik független a sisak színétől.

H_1: A csoport és a szín összefüggnek.

folyt.

	Fekete	Fehér	Sárga/Narancs	Sorösszeg
Kontroll	**491**	**377**	**31**	**899**
Várható	**513.714**	**356.827**	**28.459**	
Balesetes	**213**	**112**	**8**	**333**
Várható	**190.286**	**132.173**	**10.541**	
Oszlopösszeg	**704**	**489**	**39**	**1232**

$$\chi^2 = \sum \frac{(O-E)^2}{E} = \frac{(491-513.714)^2}{513.714} + \ldots + \frac{(8-10.541)^2}{10.541}$$

$$\chi^2 = 8.775$$

folyt.

H_0: Sor és oszlop változók függetlenek.

H_1: Sor és oszlop változók összefüggnek.

A teszt statisztika $\chi 2 = 8.775$

$\alpha = 0.05$

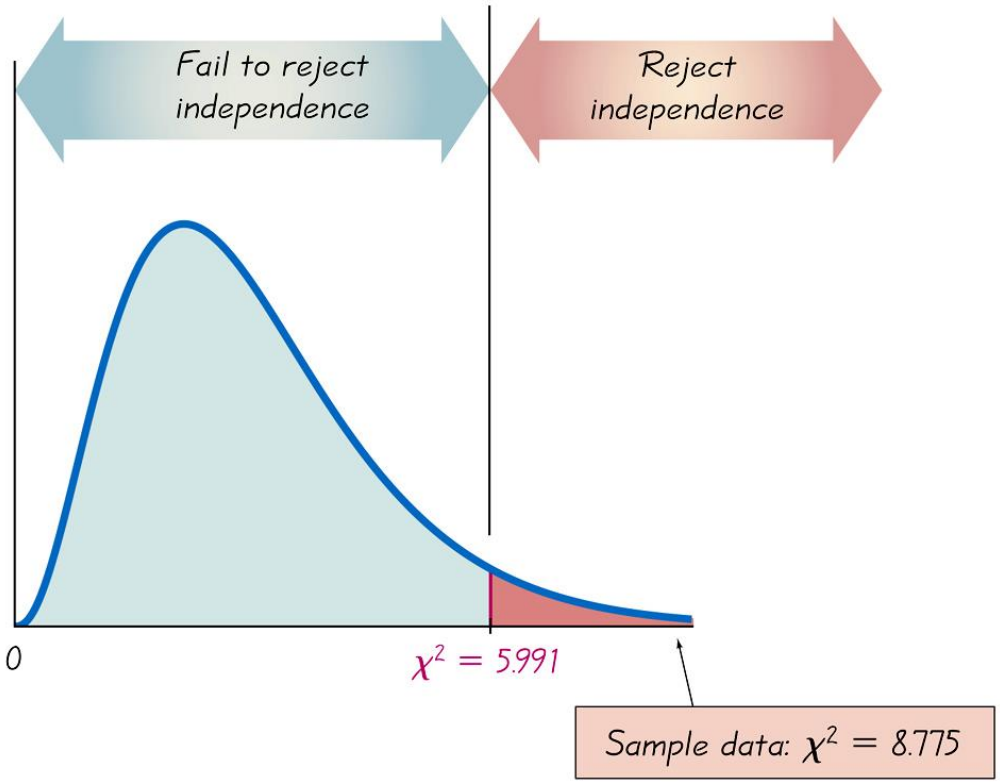
A szabadsági fokok száma:

$(r-1)(c-1) = (2-1)(3-1) = 2.$

A kritikus érték a táblázatból $\chi^2_{.05,2} = 5.991.$

folyt.

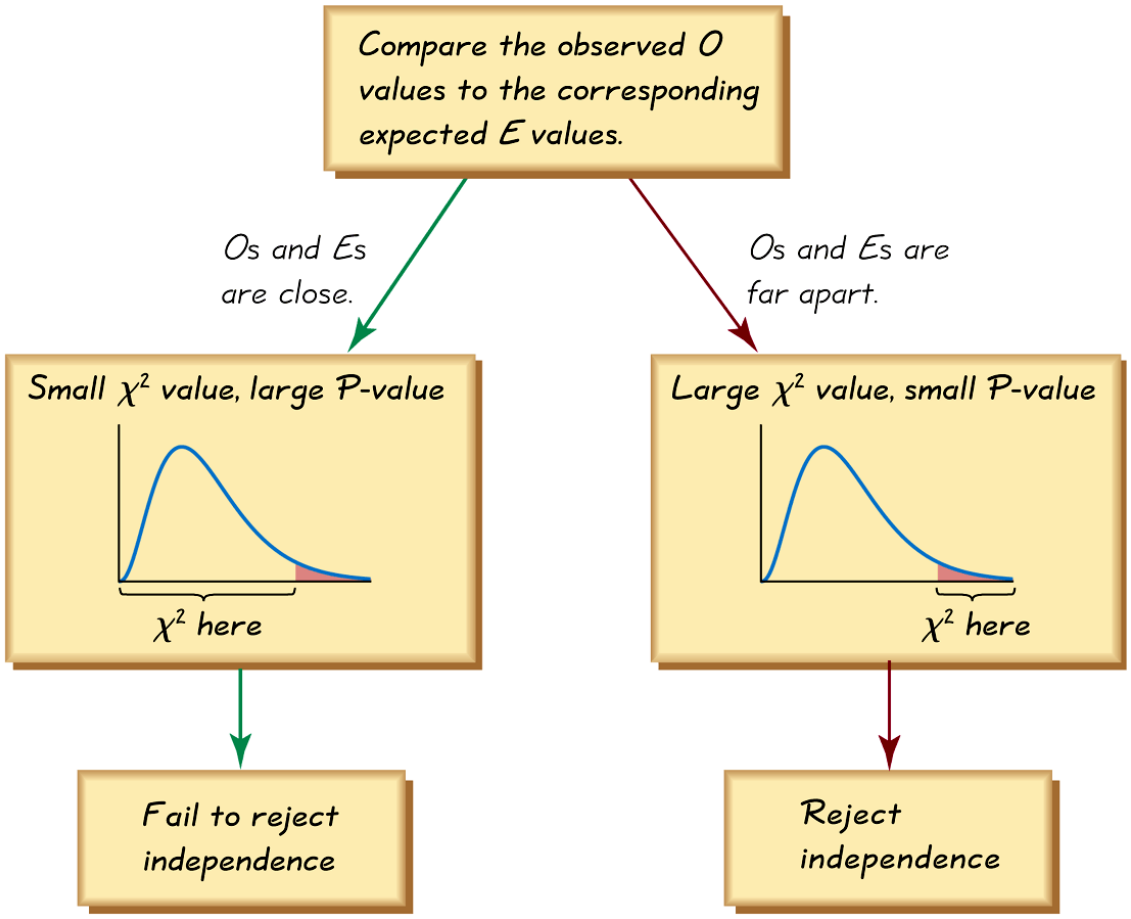
11-4. ábra



Elvetjük a null hipotézist. Úgy tűnik, van kapcsolat a sisak színe és a motorozás biztonsága között.

A tesztelés menete

11-8. ábra



Definíció

Homogenitás vizsgálat

A homogenitás vizsgálatban, azt a feltevést teszteljük, hogy *különböző populációk* bizonyos tulajdonságokat ugyanolyan arányban tartalmaznak.

Mi a különbség a homogenitás és a függetlenség vizsgálat között:

Egy *előre meghatározott* minta elemszámot használunk mindkét populációból (homogenitás vizsgálat), vagy egy nagy mintát használtunk, amiből a sor és az oszlopösszegek véletlenül jönnek ki (függetlenség vizsgálat)?

Példa: A nemek hatása

Az 11-6. táblázatot használva 0.05 szignifikancia szint mellett teszteljük, van-e hatása a kérdező nemének a férfi válaszolók válaszaira.

Table 11-6 Gender and Survey Responses

	Gender of Interviewer	
	Man	Woman
Men who agree	560	308
Men who disagree	240	92

Példa: folyt

H_0: Azok aránya, akik egyetértenek/nem értenek egyet ugyanakkora a férfi és a női kérdezők esetén is.

H_1: Az arányok különböznek.

Példa: folyt.

Minitab

```
Expected counts are printed below observed counts
Chi-Square contributions are printed below expected counts

            C1       C2  Total
    1      560      308    868
        578.67   289.33
         0.602    1.204

    2      240       92    332
        221.33   110.67
         1.574    3.149

Total      800      400   1200

Chi-Sq = 6.529, DF = 1, P-Value = 0.011
```

Összefoglalás

Ebben a fejezetben megvitattuk:

Kontingencia táblázatokat, ahol kategóriális adatok sorokba és oszlopokba vannak rendezve.

*** Függetlenség vizsgálattal teszteljük azt a feltételezést, hogy a sor és az oszlop változók függetlenek.**

*** Homogenitás vizsgálat teszteljük azt a feltételezést, hogy két populáció valamilyen tulajdonságot ugyanolyan arányban tartalmaz.**