Matematika Statisztika

1. előadás

Arató Miklós

2022.09.12.

Tudnivalók a tantárgyról

- Kötelező irodalom: az előadásokon és gyakorlatokon elhangzottak – a bemutatott módszerek, definíciók, tételek, bizonyítások, példák.
- Ajánlott irodalom:
  - Korpásné: Általános statisztika I. $\rightsquigarrow$ tankönyv leíró statisztikához
  - Molnárné-Tóthné: Általános statisztika példatár I. $\rightsquigarrow$ példatár leíró statisztikához
  - Bolla-Krámli: Statisztikai következtetések elmélete. $\rightsquigarrow$ tankönyv matematikai statisztikához
  - Fazekas (szerk.): Bevezetés a matematikai statisztikába. $\rightsquigarrow$ tankönyv matematikai statisztikához
  - Móri-Szeidl-Zempléni: Matematikai statisztika példatár.
  - Pröhle-Zempléni: Statistical Problem Solving in R. Elérési helye: `http://zempleni.elte.hu/Stat_R_Prohle_Zempleni` $\rightsquigarrow$ **R** programnyelv bevezető, a benne szereplő statisztikai témák erősen átfednek az előadással

Tudnivalók a tantárgyról, követelmények

- A tárgy felvételéhez a Valószínűségszámítás tárgy elvégzése szükséges
- A jelenlét kötelező az előadáson és a gyakorlaton is (3-3 hiányzás lehetséges)
- Gyakorlati jegy pontszámai
  - Dolgozatok (feladatok és elméleti kérdések is) az 5., 10. héten és a vizsgaidőszak első hetében, 50-50 pontért (a dolgozatoknak legalább 15 pontosoknak kell lenniük)
  - Lesz egy javítási lehetőség is
  - Beadandó önálló feladat (statisztikai elemzés). Az elemzés beadási határideje december 15. Legalább 20 pontot el kell érni!
  - 15 pont szerezhető órai munkával
  - Tervezett ponthatárok: 2-es 75 ponttól,. . . , 5-ös 150 ponttól

A tananyagról

- Tervezett tematika:
  - Leíró statisztika (röviden)
  - Becsléselmélet
  - Hipotézisvizsgálat
  - Többdimenziós statisztika elemei
- A matematika a táblán fog megszületni; a leíró statisztikai anyagrészek nagy része, közérdekű infók, feladatok szövegei, érdekességek, szimulációk, egyéb ábrák lesznek a diákon
- **A diák az anyagnak csak egy részét fedik le!!!**

Felhasznált szoftver/programnyelv: **R**

- Statisztikai modellezésre, adatok elemzésére kiválóan alkalmas programnyelv
- Nyílt forráskódú, ma már alig van probléma, feladat, aminek a megoldására ne lenne valamilyen package – akár több is
- Népszerűsége 2022 szeptemberében az összes programozási nyelv mezőnyében:
  - 7. hely – PYPL index
  - 18. hely – TIOBE index
- Python és Matlab mellett a legelterjedtebb matematikai célú programnyelv
- Letöltési helye: `https://cran.r-project.org/`
- Szövegszerkesztésre ajánlott szoftver: RStudio
  letöltési helye: `https://www.rstudio.com/products/rstudio/download3/`

A statisztika története

- Kezdetek: népszámlálások az ókorban (Kína, Római Birodalom)
- A statisztika szó eredete (vitatott):
  - *status* [latin]: állapot
  - *Staat* [német]: állam

  $\rightsquigarrow$ Sokáig a statisztika az állam állapotáról fontos információk begyűjtését jelentette.
- Tudománnyá válásának kezdete: 17. század – demográfia (népesség/társadalomstatisztika)
- A 19. századtól
  - a statisztika mindenféle információ begyűjtésének, feldolgozásának és értelmezésének a tudományává vált
  - Összekapcsolódás a valószínűségelmélettel
- A számítógépek megjelenésével fejlődése felgyorsult és jelentősége megnőtt
- A statisztika megítélése vegyes, az eredményeket mindig kritikusan kell szemlélni $\rightsquigarrow$ Churchill: "*I only believe in statistics that I doctored myself*" (Csak azoknak a statisztikáknak hiszek, amiket én magam hamisítottam.)

Motiváció

Kérdések, amikre statisztikai eszközökkel – bizonyos mértékig – választ tudunk adni:

- Idén nagyon erős aszály volt/van Magyarországon. Állíthatjuk-e, hogy ez példátlan szárazság?
- A dohányzás mennyivel növeli annak az esélyét, hogy valaki 70 éves koráig tüdőrákban betegszik meg?
- Az utolsó előtti USA-beli elnökválasztáson a közvélemény-kutatók Wisconsin államban közvetlenül a választás előtt átlagosan 6,5%-os Clinton-előnyt mértek. Mi az esélye, hogy Wisconsin-ban Trump fog győzni? [$\rightsquigarrow$ 0,7%-kal Trump nyert]
- Vajon állíthatjuk-e, hogy egy év során a bizonyos méretet meghaladó napfoltok száma Poisson-eloszlást követ? Előre tudjuk jelezni a múltbeli adatok alapján, hogy 2023-ban hány napfoltot fognak észlelni?

A statisztika fogalma és ágai

Statisztika: a valóság tömör, számszerű jellemzésére szolgáló tudományos módszertan, illetve gyakorlati tevékenység.
Ágai:

- **Leíró statisztika**: magába foglalja az információk összegyűjtését, összegzését, ábrázolását, tömör, számszerű jellemzését szolgáló módszereket
- **Matematikai statisztika**: matematikai tudomány, adatok feldolgozásáról, értelmezéséről és felhasználásáról szóló tudományos módszertan

Megjegyzés: a *statisztika* szó másik jelentése – matematikai statisztikai értelemben a statisztika egy valószínűségi (vektor)változó, amit a mintából számolunk (később bővebben)

Leíró statisztikai alapfogalmak I.

- Statisztikai egység: a statisztikai vizsgálat tárgyát képező egyed
- Statisztikai **sokaság**: a megfigyelés tárgyát képező egyedek összessége, halmaza. Röviden: sokaság.
- **Statisztikai adat**: valamely sokaság elemeinek száma vagy a sokaság valamilyen másféle számszerű jellemzője, mérési eredmény.
- Statisztikai **ismérv**: a sokaság egyedeit jellemző tulajdonság. Röviden: ismérv.
- **Ismérvváltozatok**: az ismérvek lehetséges kimenetelei.
- **Minta**: a sokaság véges számosságú részhalmaza. [A minta más értelmezéseiről később...]

Statisztikai következtetés: a valóságban a teljes sokaságot nem tudjuk vagy akarjuk megfigyelni, ezért csak az egyedek egy szűkebb csoportját figyeljük meg. A viszonylag kisszámú egyedre vonatkozó információk alapján szeretnénk a teljes sokaság egészére, egyes jellemzőire, tulajdonságaira érvényes következtetéseket kimondani.

Leíró statisztikai alapfogalmak II.

Példák:

Sokaság:	most a teremben lévő homo sapiensek
Statisztikai egység:	a teremben lévő oktató
Adat:	a legmagasabb hallgató testtömegindexe
Ismérv:	nem
Ismérvváltozatok:	férfi ($\to 1$), nő ($\to 0$)
Minta:	5 véletlenül választott hallgató

Sokaság:	az ELTE TTK Matematikai szakgyűjteményében lévő könyvek
Statisztikai egység:	a BF 13873 raktári jelzetű könyv
Adat:	a szakgyűjteményben lévő könyvek száma
Ismérv:	oldalak száma
Ismérvváltozatok:	631, 321, 153, 463, . . .
Minta:	a Rényi: Valószínűségszámítás című könyvek

Csoportosítások, adatok fajtái

A sokaságok csoportosítása:

1.) A sokaság egységeinek megkülönböztethetősége szerint:
   - diszkrét: a sokaság egységei elkülönülnek egymástól
   - folytonos: a sokaság egységeit nem tudjuk természetes módon elkülöníteni (pl. bauxittermelés)

2.) A sokaság időpontra vagy időtartamra értelmezhető-e:
   - álló: csak egy adott *időpontra* értelmezhető
   - mozgó: csak egy adott *időtartamra* értelmezhető

3.) A sokaság számossága szerint:
   - véges (a gyakorlatban általában ilyenekkel foglalkozunk)
   - végtelen

A statisztikai adatok fajtái:

- Alapadatok: közvetlenül a sokaságból származnak (méréssel, megszámlálással)
- Leszármaztatott adatok: alapadatokból műveletek eredményeként adódnak (pl. átlagolással, osztással)

A statisztikai adatok nem mindig pontosak – a mért és a tényleges adat eltérhet egymástól, például kerekítési okokból.

Ismérvek

- Az ismérvek típusai I.
  - minőségi ismérv: az egyedek számszerűen nem mérhető tulajdonsága
  - mennyiségi ismérv: az egyedek számszerűen mérhető tulajdonsága. Két fajtájukat különböztetjük meg:
    - diszkrét: véges vagy megszámlálhatóan sok értéket vehet fel
    - folytonos: egy adott intervallumon belül kontinuum számosságú értéket felvehet
  - időbeli ismérv: az egységek időbeli elhelyezésére szolgáló rendezőelvek
  - területi ismérv: az egységek térbeli elhelyezésére szolgáló rendezőelvek
- Az ismérvek típusai II.
  - közös ismérvek: tulajdonságok, amik szerint a sok. egyedei egyformák
  - megkülönböztető ismérv: azok a tulajdonságok, amik szerint a sokaság egyedei különböznek egymástól

Ismérvek (példa)

Legyen a sokaság: a teremben lévő hallgatók. Példák ismérvekre:

minőségi:	szemszín, nem	közös:	orrok száma
diszkrét mennyiségi:	testvérek száma	megkülönböztető:	testsúly
folytonos mennyiségi:	testmagasság		
időbeli:	születési idő		
területi:	születési hely		

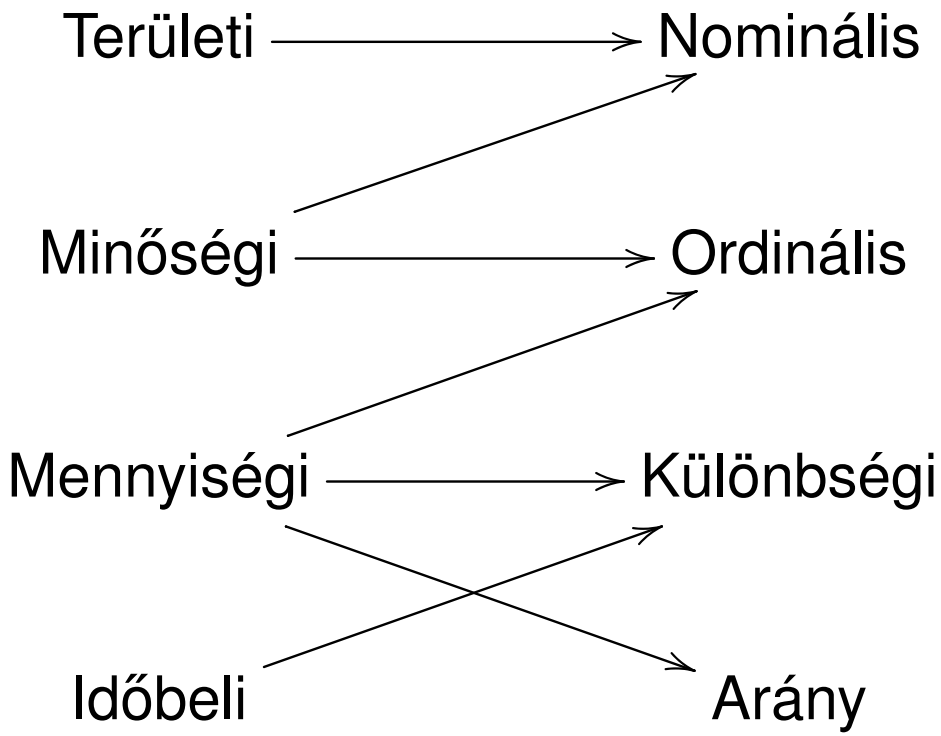
Skálák

Mérési skálák (mérési szintek):

- Névleges (nominális): a számok csak ún. kódszámok, amik a sokaság egyedeinek azonosítására szolgálnak. Ezek között matematikai relációkat és műveleteket nincs értelme végezni. Pl. a hallgatók neme.
- Sorrendi (ordinális): a sokaság egyedeinek valamely tulajdonság alapján sorba való rendezése. Az egyedek tulajdonsága közötti különbséget nem lehet mérni. Pl. a hallgatók jegyei egy tárgyból.
- Intervallumskála: a skálaértékek különbségei is valós információt adnak a sokaság egyedeiről. A skálán a nullpont meghatározása önkényes. Ilyen skálákhoz mértékegység is tartozik. Pl. hőmérséklet.
- Arányskála: a skálának van valódi nullpontja is. Minden matematikai művelet elvégezhető ezekkel a számokkal. Pl. a hallgatók magassága.

[Metrikus skála: intervallum- és arányskála közös neve – ritkábban használatos elnevezés]

Az ismérvek és a mérési skálák kapcsolódása:



Sorok

Statisztikai sor: a sokaság egyes jellemzőinek felsorolása.
Az ismérvek fajtája szerint beszélhetünk minőségi, mennyiségi, időbeli és területi sorokról.
A statisztikai sorok további csoportosítása:

- Csoportosító sor: a sokaság egy megkülönböztető ismérv szerinti osztályozásának eredménye; az adatok összegezhetők (van 'Összesen' sor)
- Összehasonlító sor: a sokaság *egy részének* a sokaságot egy megkülönböztető ismérv szerinti osztályozásának eredménye; az adatok nem összegezhetők
- Leíró sor: különböző fajta, gyakran eltérő mértékegységű statisztikai adatokat tartalmaz

Például ha egy statisztikai sor tartalmazza az osztályteremben a hallgatókat nemek szerint, akkor ez a sor minőségi csoportosító sor.

Táblák

Statisztikai tábla: a statisztikai sorok összefüggő rendszere.
A statisztikai táblák fajtái:

- Egyszerű tábla: nem tartalmaz csoportosítást, nincs benne összegző sor
- Csoportosító tábla: egyetlen csoportosító sort tartalmaz
- Kombinációs tábla vagy *kontingenciatábla* vagy kereszttábla: legalább két csoportosító sort tartalmaz

Viszonyszámok

A statisztikai elemzések egyik legfontosabb eszközei a viszonyszámok (alias: indikátorok). A **viszonyszám** két statisztikai adat hányadosa. Jelölések:

$$V = \frac{A}{B}$$

ahol V: viszonyszám; A: a viszonyítás tárgya; B: a viszonyítás alapja.

A viszonyszámok fajtái:

- Megoszlási: a sokaság egy részének a sokaság egészéhez való viszonyítása
- Koordinációs: a sokaság egy részének a sokaság egy másik részéhez való viszonyítása
- Dinamikus: két időpont vagy időszak adatának hányadosa
- Intenzitási: különböző fajta adatok viszonyítása egymáshoz; gyakran a mértékegységük is eltérő.

A statisztikai elemzés lépései

1.) Tervezés
   a.) Mit vizsgálunk, mi a probléma/feladat
   b.) Hogyan gyűjtjük az adatokat
   c.) Előzetes sejtések, hipotézisek megfogalmazása
2.) Terepmunka – adatgyűjtés
3.) Adatbevitel, kódolás (ha szükséges)
4.) Adatok validálása (biztosan rossz értékek kiszűrése, mint például életkornál a 9999)
5.) Adatelemzés, adatellenőrzés: leíró statisztikákkal, grafikonok készítése
6.) Hibás adatok kijavítása vagy kihagyása
7.) Adatelemzés, statisztikai következtetések levonása – a matematikai statisztika módszereivel
8.) Az eredmények értelmezése, visszacsatolás

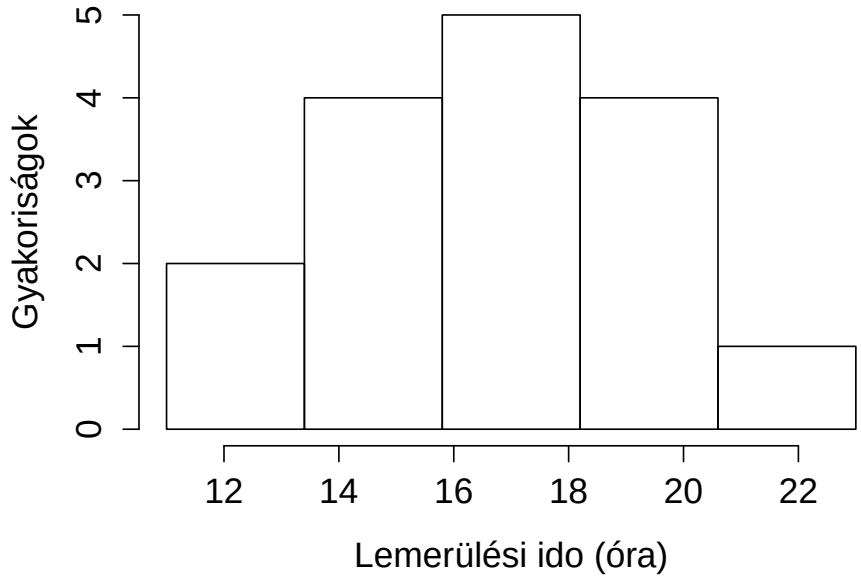
A grafikus megjelenítés szerepe

- A statisztikus legfőbb kommunikációs eszközei a diagramok.
- Az emberek többsége utálja a
  - barokkos körmondatokkal teletűzdelt statisztikai jelentéseket.
  - számokkal teli táblázatokat.
- Az adatokban rejlő információk gyorsabb kinyerését és feldolgozását segítik az azokból készített különféle ábrák, diagramok:
  - kördiagram: megoszlás érzékeltetésére
  - oszlopdiagram: idősorok ábrázolására
  - vonaldiagram: idősorok ábrázolására
  - hisztogram: mennyiségi sorok ábrázolására
  - stb.
- Milyen a jó diagram?
  - illeszkedik az ábrázolt adatok fajtájához és a probléma jellegéhez
  - a célközönség meg tudja érteni
  - áttekinthető, olvashatók rajta a feliratok, jelölések
  - kreatív, esztétikus

Hisztogram

Hisztogram – Ha a mennyiségi ismérv folytonos vagy sok ismérvérték van, akkor alkalmas módon osztályokat képezünk, majd minden egyes adatot pontosan egy osztályhoz rendeljük. A hisztogram az osztályok gyakoriságait ábrázolja.

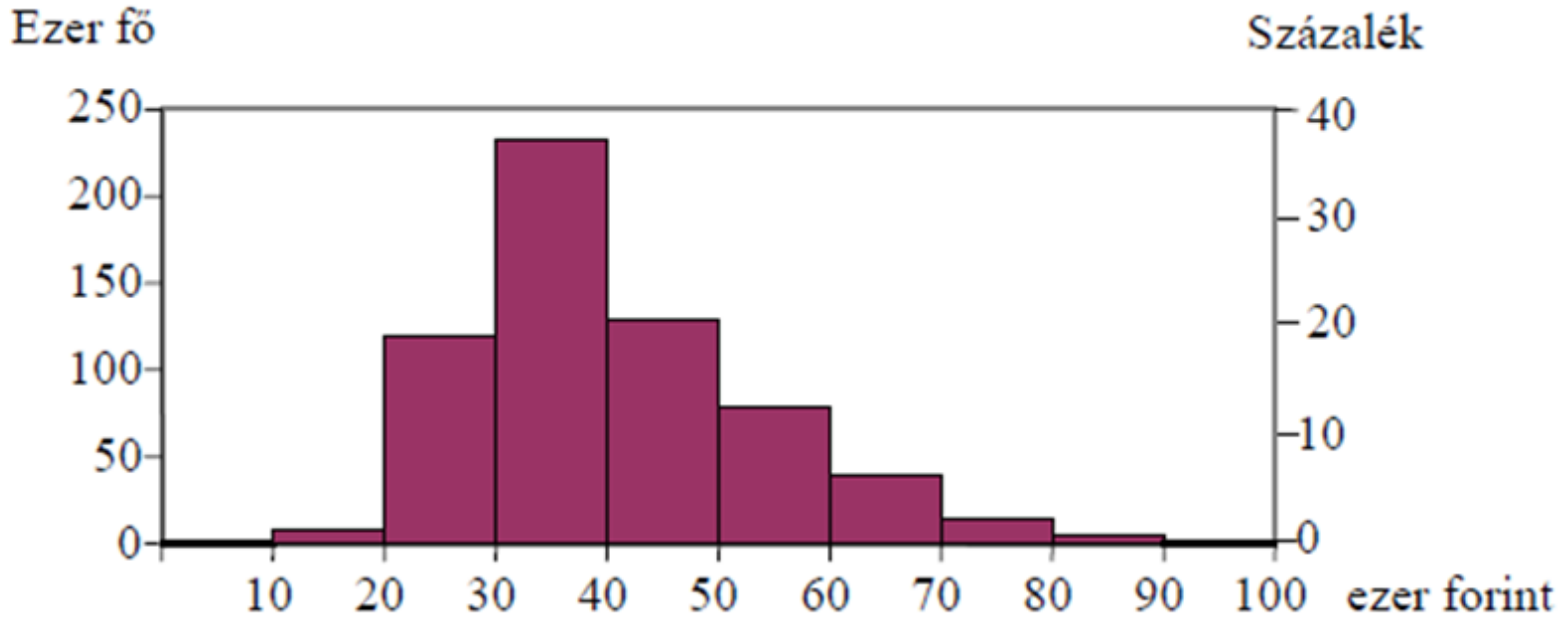
- javaslat az osztályok számára: $k = \lfloor \log_2 n \rfloor$
- ha azonos hosszúságú (h) osztályközöket akarunk létrehozni, akkor $h = \frac{x_n^* - x_1^*}{k}$
- az f_i gyakoriságokat ábrázoljuk a függőleges tengelyen
- sűrűséghisztogramnál a $g_i = \frac{f_i}{n}$ relatív gyakoriságokat ábrázoljuk a függőleges tengelyen



- **ha az osztályközök különböző hosszúságúak, akkor a gyakoriságokat egy közös hosszra kell arányosítani**

Hisztogram (folyt.)

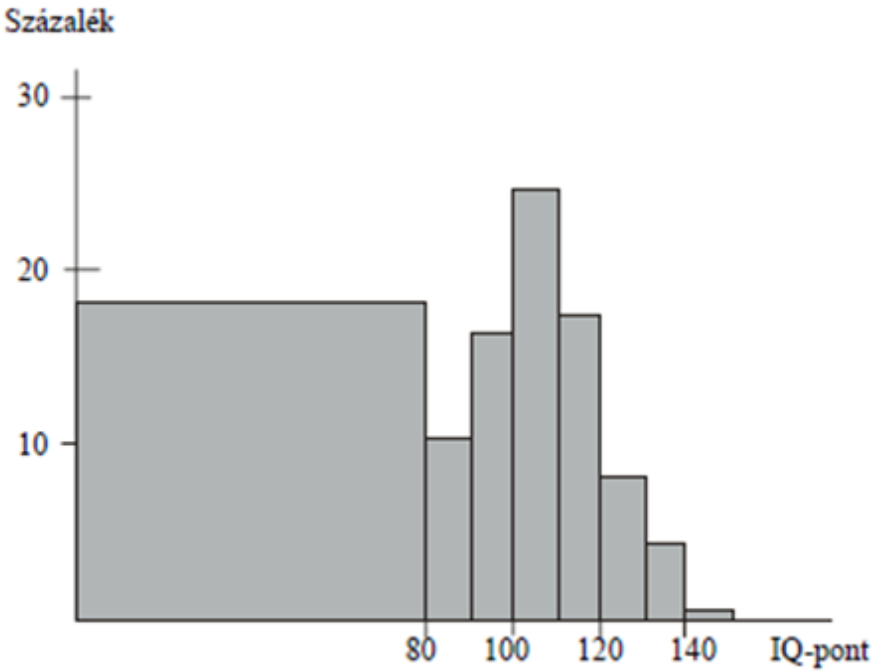
A férfi népesség megoszlása az öregségi nyugdíjak nagysága szerint
2000. január1.

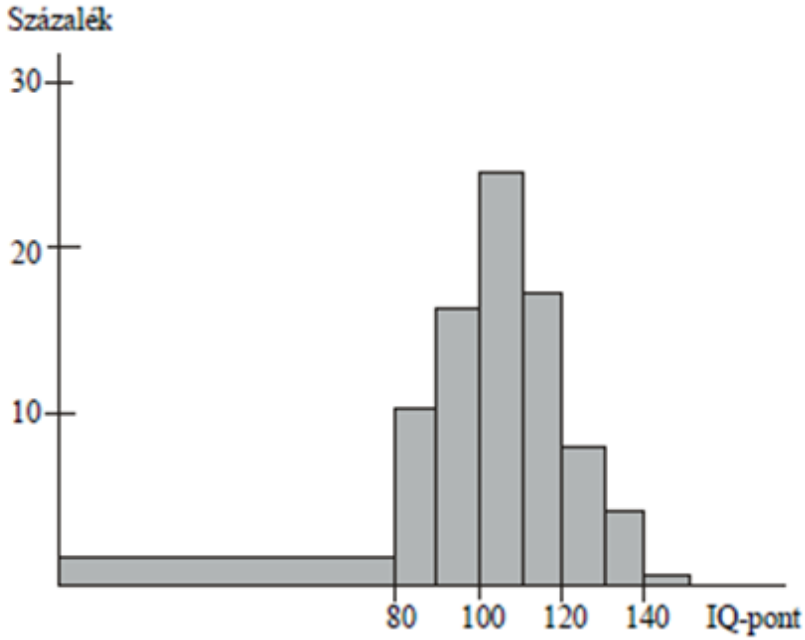


Adatforrás: *Magyar statisztikai évkönyv, 1999* (2000). Központi Statisztikai Hivatal, Budapest.

Hisztogram (folyt.)

15. ábra. Egy népességcsoport megoszlása IQ-pontok szerint





IQ-pont	Százalék	IQ-pont	Százalék	IQ-pont	Százalék
-80	18,2	101-110	24,7	131-140	4,1
81-90	10,8	111-120	17,3	141-150	0,5
91-100	16,2	121-130	8,2	*Összesen*	*100,0*

Hisztogram (folyt.)