Matematikai alapok és valószínűségszámítás

Statisztikai változók
Adatok megtekintése

Statisztikai változók

A statisztikai elemzések során a vizsgálati, vagy megfigyelési egységeket különböző jellemzők mentén vizsgáljuk, adatokat gyűjtünk (pl. az iskolai végzettség és a kreativitás).

Ezek a jellemzők általában vizsgálati egységről vizsgálati egységre változnak, ezért ezeket ***statisztikai változó***knak nevezzük.

Adatmátrix

	Nem	Iskolai végzettség	Intelligencia	Testmagasság
Személy 1	Férfi	alapfokú	95	190
Személy 2	Férfi	felsőfokú	123	173
Személy 3	Nő	alapfokú	111	178
Személy 4	Nő	középfokú	109	162
Személy 5	Férfi	középfokú	130	169
Személy 6	Nő	posztgraduális	95	165
Személy 7	Férfi	felsőfokú	104	184

Adatmátrix

A statisztikai változók *értékei*ről beszélünk, és ezeket az értékeket konvencionálisan számokkal reprezentáljuk.

	Nem	Iskolai végzettség	Intelligencia	Testmagasság
Személy 1	1	1	95	190
Személy 2	1	3	123	173
Személy 3	2	1	111	178
Személy 4	2	2	109	162
Személy 5	1	2	130	169
Személy 6	2	4	95	165
Személy 7	1	3	104	184

Statisztikai változók típusai I.

Számoknak számos különböző tulajdonsága van:

- Sorba rendezhetők
- Összegük is értelmes
- Hányadosuk is értelmes

A statisztikai változók megkülönböztethetők aszerint, hogy a fenti tulajdonságok közül melyekkel rendelkeznek!

Ez alapján négy ún. mérési skáláról beszélhetünk statisztikai változók esetén.

Nominális skála

A változó a számok egyetlen tulajdonságával sem rendelkezik

- Kit kedvel a leginkább a következő festők közül?
  1. Dalí
  2. Klimt
  3. Van Gogh
  4. Da Vinci
- Biológiai nem
  1. Férfi
  2. Nő
- Milyen cigarettát szív?

Ordinális skála

A számok tulajdonságai közül rendelkezik a sorba rendezhetőséggel.

- Pl. Iskolai végzettség:
  1. Alapfokú, vagy az sem
  2. Középfokú
  3. Felsőfokú
  4. Posztgraduális

Ordinális skála esetén igaz az, hogy 1<2<3<4

DE: nem mondhatjuk, hogy: 2-1=4-3

Intervallum skála

Értékei sorba rendezhetők és összegük (különbségük) is értelmes.

- Intelligencia teszten elért pontszámok
  - Sorba rendezhető: 99<100<101
  - A különbségük értelmes: 75-70=120-115

DE: az arányuk már nem értelmezhető:

NEM mondhatjuk, hogy a 90-es IQ a 180-as IQ fele

Arány skála

A változó értékei sorba rendezhetők, különbségük és arányuk is értelmes

- Testmagasság
  - Sorba rendezhető: 160<170<171
  - A különbségük értelmes: 75-70=190-185
  - Az arányuk is értelmes: 80cm éppen 160cm fele

Skála-típusok összefoglalása

	sorbarendezhető	különbség értelmes	arány értelmes
nominális	nem	nem	nem
ordinális	igen	nem	nem
intervallum	igen	igen	nem
arány	igen	igen	igen

Statisztikai változók típusai II.

Az alapján, hogy hány különböző értéket vehet fel egy adott változó két típust különböztethetünk meg:

- ***Diszkrét változó*:** véges számú különböző értéket vehet fel, és az értékek egymástól jól elkülönülnek.

  Pl.: iskolai végzettség, családi állapot

- ***Folytonos változó:*** értékei folytonosan helyezkednek el.

  Pl.: reakcióidő, testmagasság

Statisztikai változók típusai III.

Attól függően, hogy számszerűsíthető-e egy adott változó szintén két típust különböztethetünk meg:

- ***Kvalitatív változó*:** bár a változó értékei számokká konvertálhatók, ám ezek a számok nem rendelkeznek a számok egyetlen tulajdonságával sem. A változó értékei nem számszerűsíthető minőséget fejeznek ki.

  Pl.: biológiai nem, családi állapot

- ***Kvantitatív változó:*** A változó értékei számszerűsíthető minőséget fejeznek ki.

  Pl.: reakcióidő, testmagasság

Statisztikai változók típusai

Skálatípus?
Kvalitatív vagy kvantitatív változó?
Diszkrét vagy folytonos változó?

- Dohányzási szokásokat mérő kérdőív alapján a függőség számszerűsített mértéke (20 fokú skálán).
- Dohányzással töltött idő naponta
- Milyen cigarettát szív?
- Dohányzik?

Gyakran, ritkán, soha.

A változók eloszlása a populációban és a mintában

- Az adatmátrixból minden személyről számos információt leolvashatunk

  pl. neme, végzettsége, depresszió mértéke

- Alapvetően azonban a populáció érdekel bennünket

  pl. mekkora hányada milyen nemű, végzettségű, depresszió tekintetében milyen az eloszlás

- Adataink azonban csak a mintáról vannak, ennek szerepe, hogy segítségével becsüljük a populáció jellemzőit. Minél reprezentatívabb a minta, annál sikeresebb a becslésünk

  pl. mekkora hányada milyen nemű, végzettségű, depresszió tekintetében milyen az eloszlás

Leíró statisztikák

Vizsgálataink során tehát kérdéseket fogalmazunk meg, meghatározzuk azon ***populáció***t, amelyre az érdeklődésünk irányul, majd a sokaságból ***mintá***t veszünk (lehetőleg nagyszámú, véletlen mintát) és a mintába került megfigyelési egységeinkről ***adat***okat gyűjtünk különböző statisztikai változók (a megfigyelési egységek jellemzői, ismérvei) mentén.

Az összegyűjtött adatokat ***adatmátrix***ban rögzítjük, melyben jellemzően egy megfigyelési egység (eset) egy sor, míg az egyes oszlopok egy-egy változót reprezentálnak.

Minthogy azonban az ideális minta nagyszámú megfigyelési egységet tartalmaz, az adatmátrix általában túlontúl terjedelmes ahhoz, hogy egyszerű megtekintés útján megállapíthassuk az adatok lényeges tulajdonságait.

Az adatok lényeges ismérveinek jellemzésére, leírására ún. ***leíró statisztikáka***t használunk. Erre a célra használhatunk:

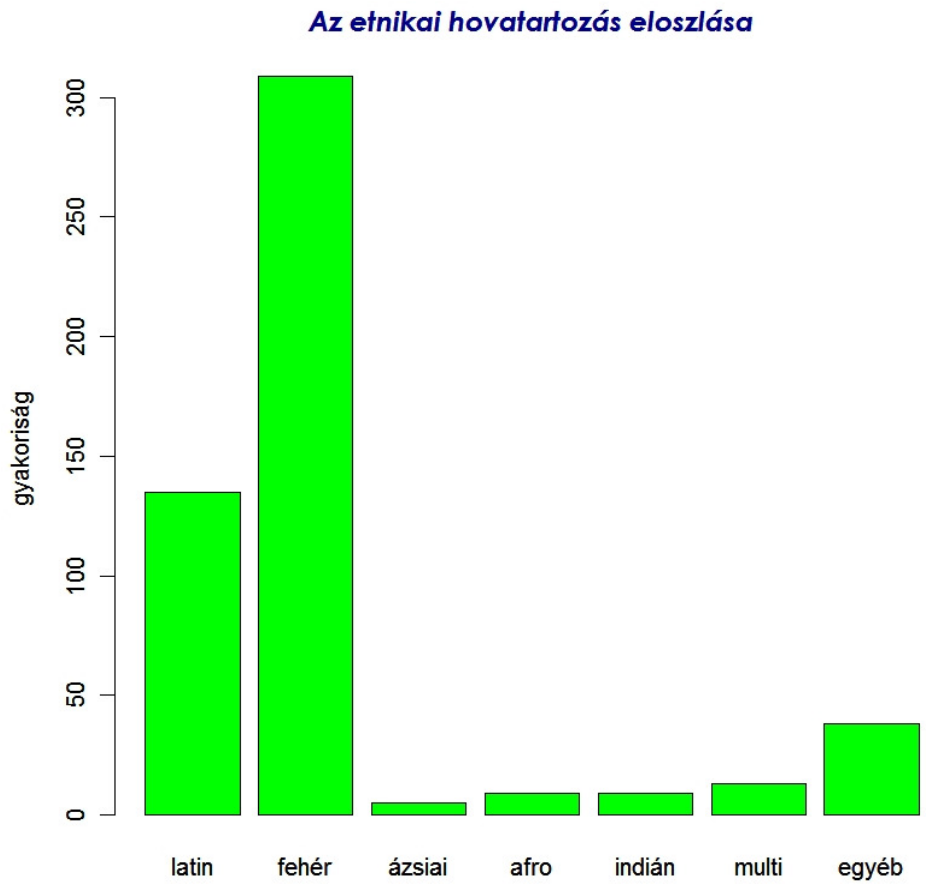
- Grafikus módszereket
- Az adatokból számolt összegző statisztikákat

Gyakorisági eloszlás (nominális változó)

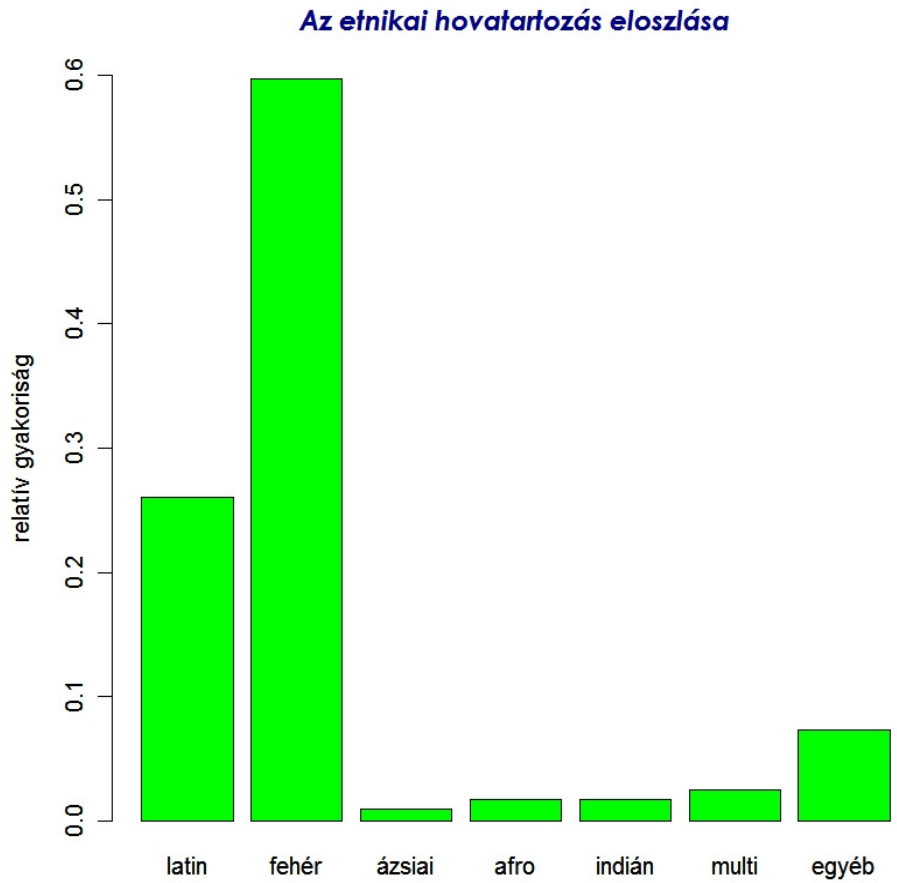
Egy változó jellemezhető azzal, hogy különböző értékei hányszor, milyen gyakran fordulnak elő az adott mintában. A gyakorisági eloszlás éppen ezt, a változó értékeinek gyakoriságát fejezi ki.

Pl. Etnikai hovatartozás	gyakoriság	relatív gyakoriság	százalék
1. Latin	135	.26	26 %
2. Fehér	309	.60	60 %
3. Ázsiai	5	.01	1 %
4. Afrikai	9	.02	2 %
5. Indián	9	.02	2 %
6. Multietnikai	13	.03	3%
7. Egyéb	38	.07	7 %

Oszlopdiagram



Oszlopdiagram

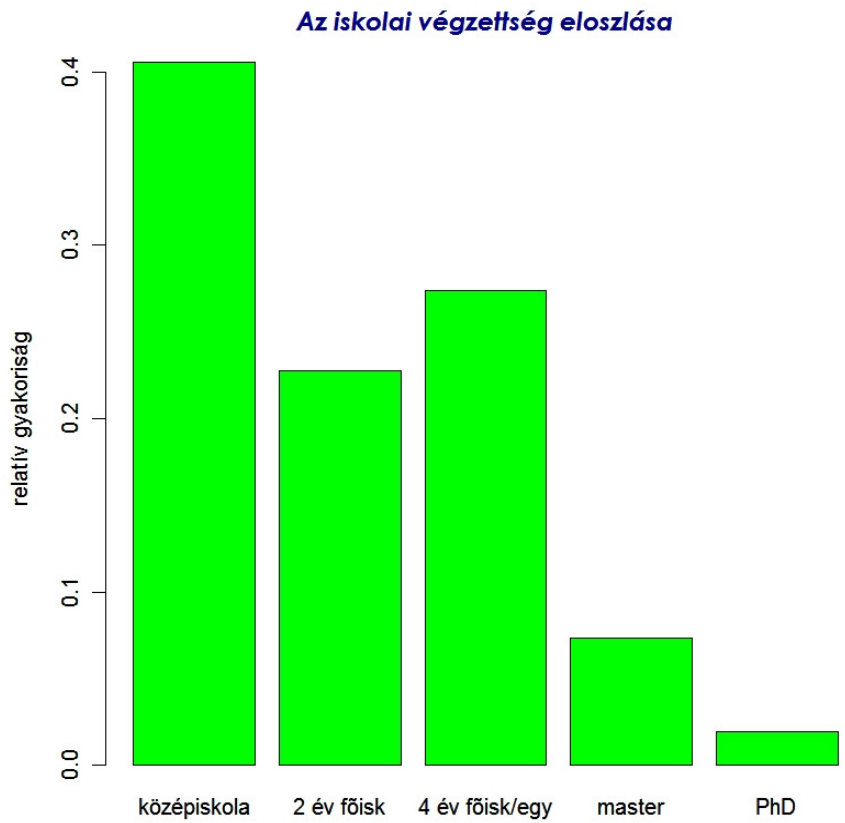


Gyakorisági eloszlás (ordinális változó)

Pl. iskolai végzettség:

	gyak	rel. gyak.	%	kum. gyak.	Kum. %
középiskola	210	.40	40	210	40
2 év főiskola	118	.23	23	328	63
4 év főiskola	143	.28	28	471	91
master	38	.07	7	509	98
PhD	10	.02	2	519	100

Oszlopdiagram

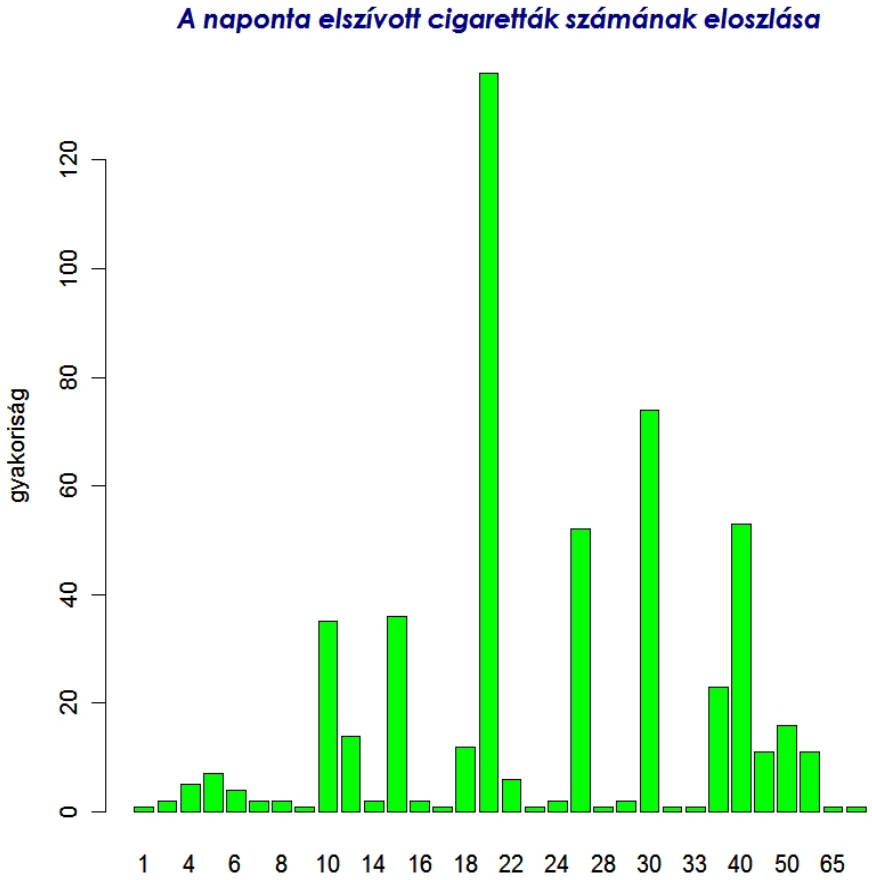


Gyakorisági eloszlás (folytonos változó)

Pl. naponta elszívott cigaretták száma

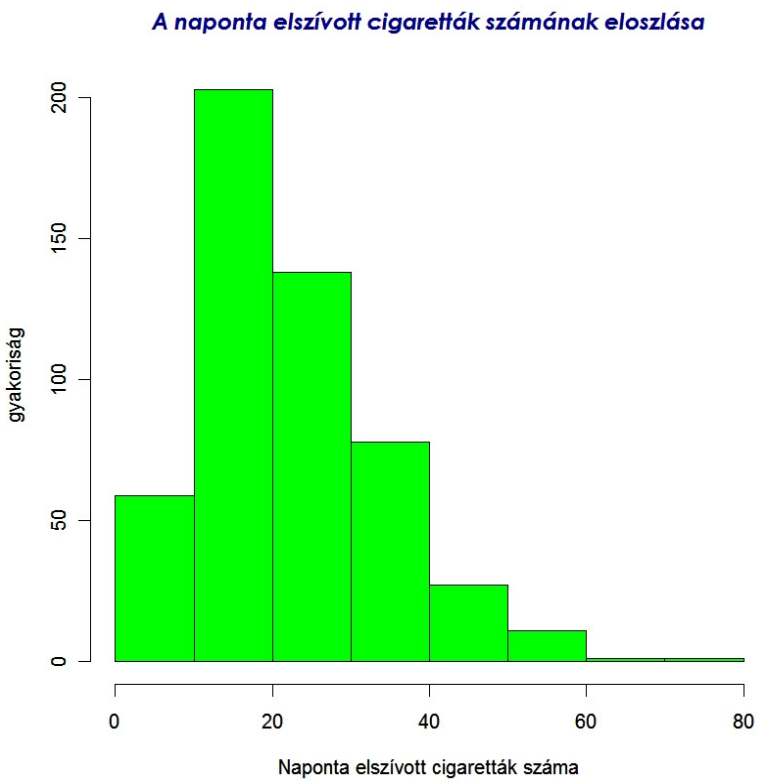
érték	gyakoriság	érték	gyakoriság
1	1	22	6
2	2	23	1
4	5	24	2
5	7	25	53
6	4	28	1
7	2	29	2
8	2	30	75
9	1	32	1
10	35	33	1
12	14	35	23
14	2	40	54
15	36	45	11
16	2	50	16
17	1	60	11
18	12	65	1
20	138	80	1

Oszlopdiagram



Folytonos változók esetén az oszlopdiagram nem a megfelelő megoldás az adatok grafikus megjelenítésére, mert a változónak túl nagyszámú értéke van, amelyek gyakorisága meglehetősen kicsi lesz, még nagy minta esetén is.

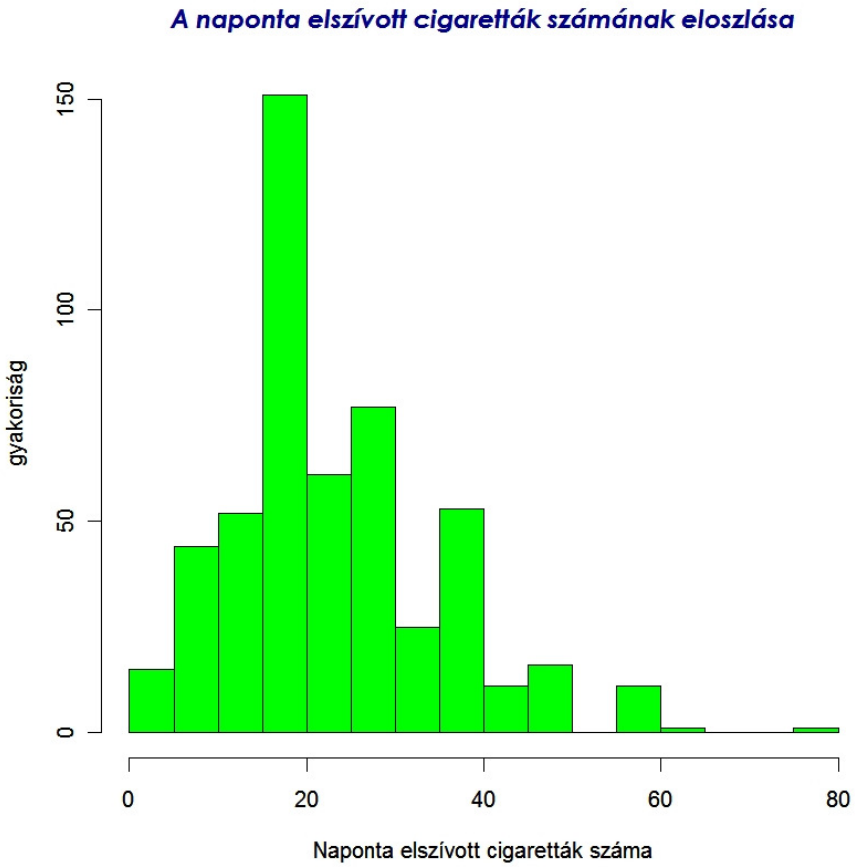
Hisztogram



Az értékeket intervallumokba soroljuk, és az egyes intervallumokba eső értékek gyakoriságára koncentrálunk. Ha pl. a naponta elszívott cigaretták számát 8 intervallumba (1-10, 11-20, 21-30,…, 71-80) soroljuk, a balra látható hisztogramot kapjuk.

Hisztogram

Az intervallumok száma (bizonyos határok között) önkényesen megválasztható, azonban a túl sok vagy túl kevés intervallum sem jó, mivel az előbbi esetben nem tömörítjük kellőképpen az információt, míg az utóbbi esetben pedig éppen fordítva, olyannyira tömörítjük, ami már információvesztést okozhat.



A sűrűségfüggvény

- Mint láttuk, a folytonos változók mintabeli eloszlását hisztogrammal reprezentáljuk, de ha a folytonos változók populációbeli eloszlását tekintjük, azt legjobban az ún. sűrűségfüggvénnyel lehet leírni.
- A sűrűségfüggvény jellemzői:
  - X változó minden x értékéhez egy nem negatív számot rendel (ez a szám nagyobb azon x-ek közelében, ahol az adatok jobban sűrűsödnek)
  - Az egy-egy szakaszra eső terület a populáció ezen szakaszra eső arányát adja meg
  - A sűrűségfüggvény alatti összterület 1, azaz 100%
  - más néven eloszlásgörbének is nevezzük

Példa sűrűségfüggvényre