Szent István Egyetem Állatorvos-tudományi Kar
Biomatematikai és Számítástechnikai Tanszék

11. előadás

Bevezetés a statisztikába

Fodor János
Biomatematika 2001

Copyright © 2001 jfodor@univet.hu
Last Revision Date: November 9, 2001 Version 1.10

Table of Contents

1 Statisztika: Bevezetés, alapfogalmak 5

1.1 A minta 17

1.2 Az adatok 20

1.3 Mérési skálák 24

2 Adatok összegzése 28

2.1 Gyakorisági táblázatok 28

2.2 Adatok grafikus reprezentációja . . . 34

2.3 Kvalitatív adatok 34

- Oszlopdiagram (bar chart) 35
- Tortadiagram (pie chart) 38

2.4 Kvantitatív adatok 38

- Pontdiagramok (dotplots) 39
- Hisztogramok 42

2.5 Középértékek 47

- Átlag 48
- Medián 51
- Módusz 57

2.6 A szóródás mérőszámai 62

- Terjedelem 63
- Interkvartilis terjedelem 64
- Tapasztalati szórás és variancia . . 70

1. Statisztika: Bevezetés, alapfogalmak

A statisztika olyan tudomány, amely adatok összegyűjtésével, rendszerezésével, és értelmezésével foglalkozik abból a célból, hogy valamilyen, a valós életben felmerült kérdésre választ találjon.

Például:

- Okoz-e tüdőrákot a dohányzás?
- Több tanítási módszer közül melyik a legjobb?
- Egy vakcinától azt várták el, hogy a juhok 90%-át megóvja egy bizonyos betegségtől. 30 állatot

beoltottak, majd kitették őket a fertőzésnek. Bizonyítja-e a vakcina hatásosságát, ha csak egy kapta meg a betegséget? Hatásos-e a vakcina abban az esetben, ha 5 kapta meg? Mi történt volna, ha nem oltottuk volna be őket?

- 6 bikával és 6 ökörrel 7 hónapon keresztül ugyanazt a tápot etették. A napi átlagos súlygyarapodás a bikák esetében 690, 750, 640, 630, 610, 620 gramm, az ökrök esetén 770, 630, 680, 790, 610, 660 gramm volt. Az átlagos súlygyarapodás bikák esetén 657, az ökrök esetén 690 gramm volt. Bizonyítja-e ez azt, hogy a

növekedés nagyobb az ökrök esetén?

A statisztikának két jól elkülöníthető része van: **leíró**, illetve **induktív** statisztika.

Leíró statisztika: olyan módszerek, amelyek az összegyűjtött adatok leírására szolgálnak.

Példa.

Tipikus példa a **népszámlálás**, amikor mindenkinek felveszik az adatait.

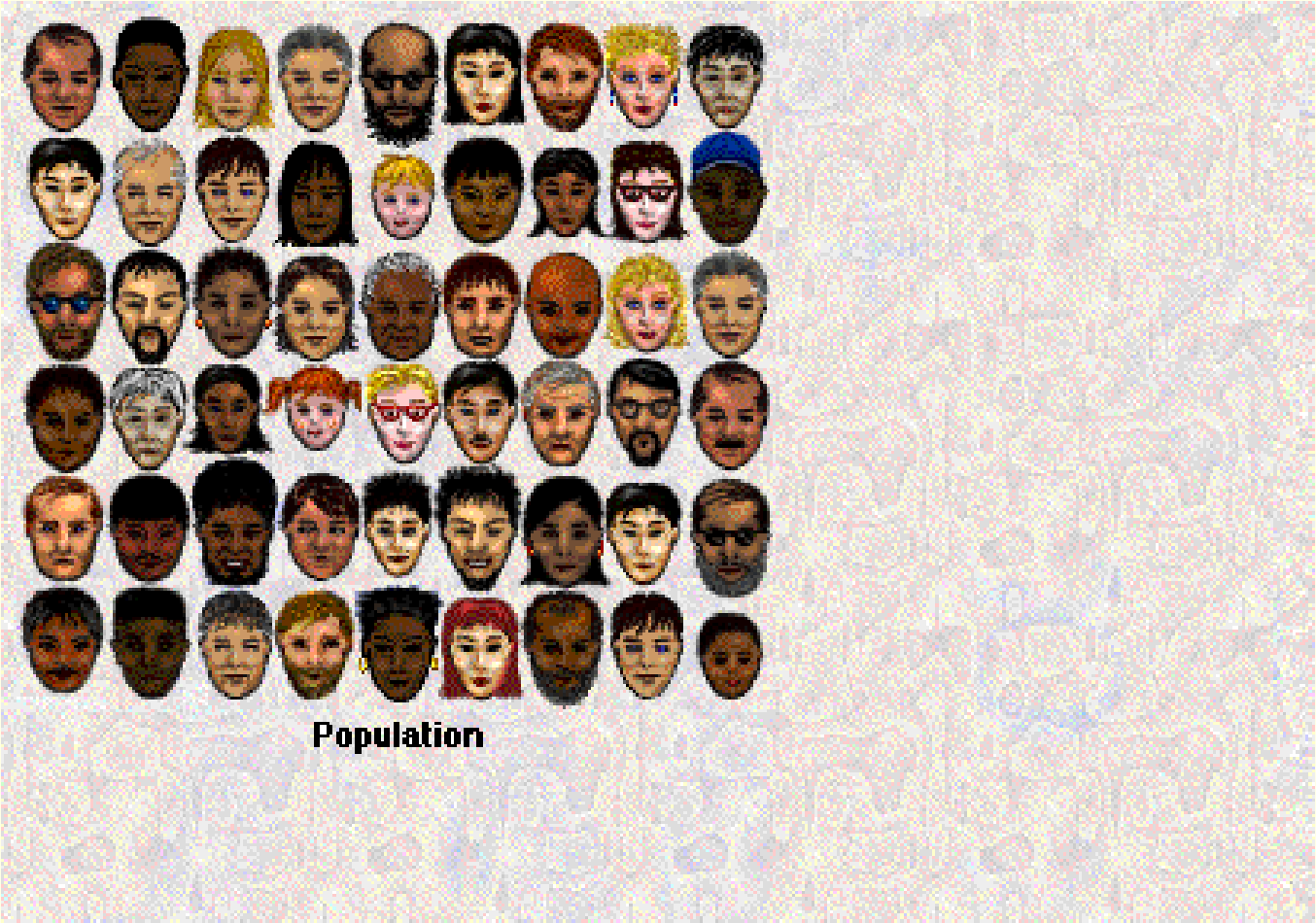
Ha a 10 millió adatot elénk teszik, azzal semmire sem megyünk. Ezt az óriási adathalmazt sűrített

formában ábrázolják: grafikonok, átlagok és egyéb statisztikai mutatók segítségével.

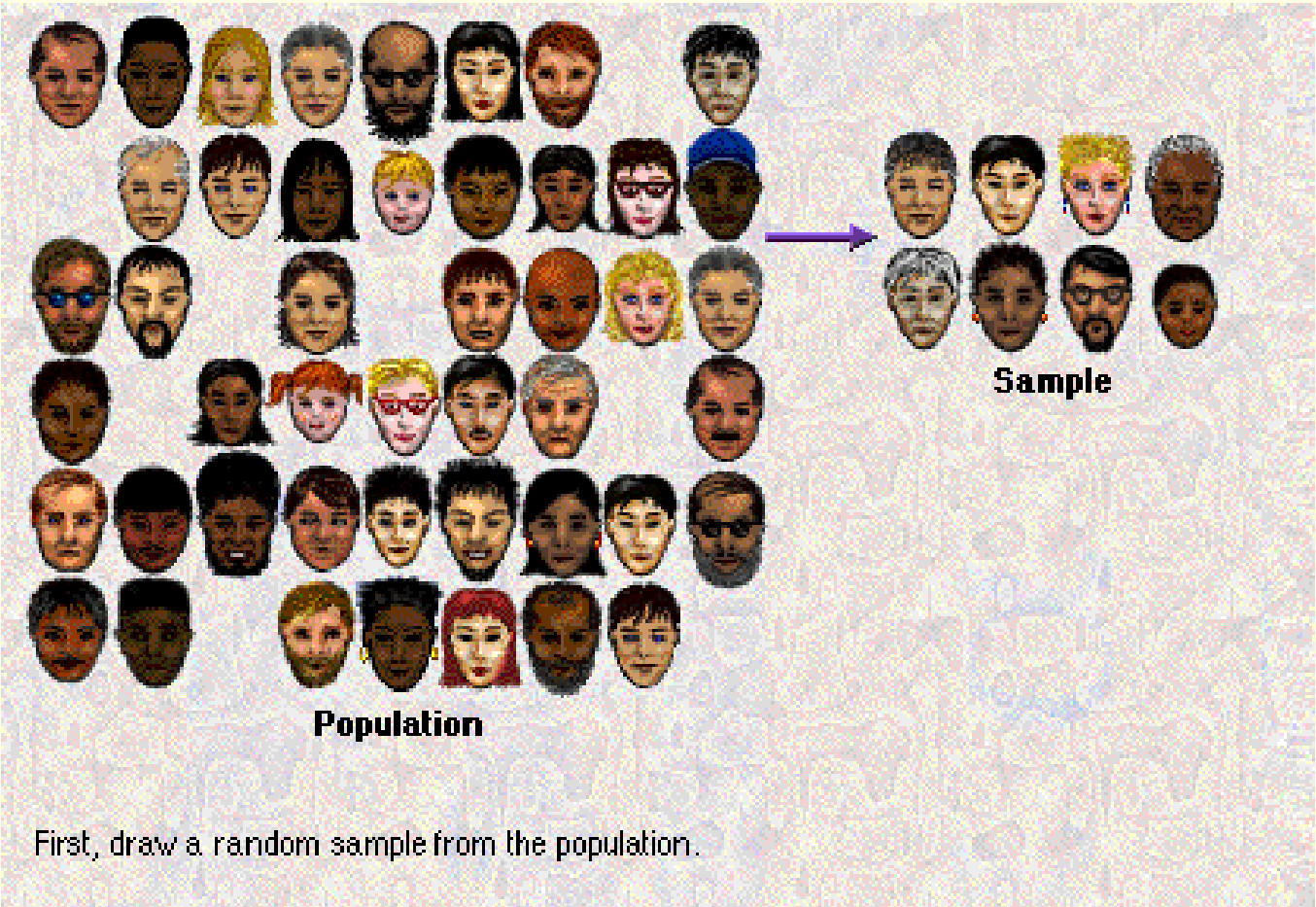
Induktív statisztika: olyan módszerek, amelyeket arra használnak, hogy egy minta alapján következtessenek a populáció tulajdonságaira.

Példa. az egyes tévéműsorok nézettségének vizsgálata; választás előtti közvéleménykutatás.

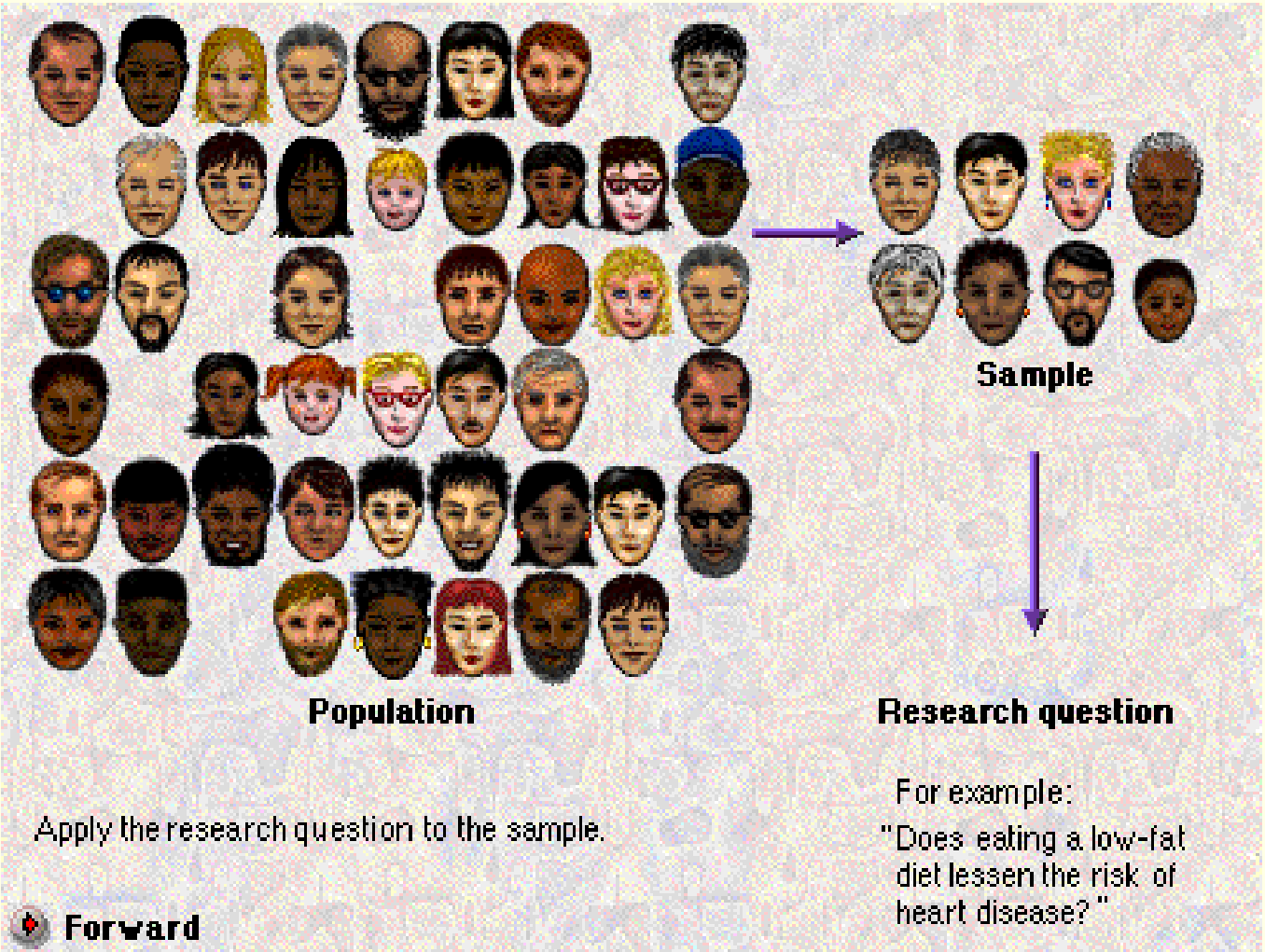
Populáció: a vizsgálat tárgyát képező egyedek összessége (alapsokaságnak is hívják).



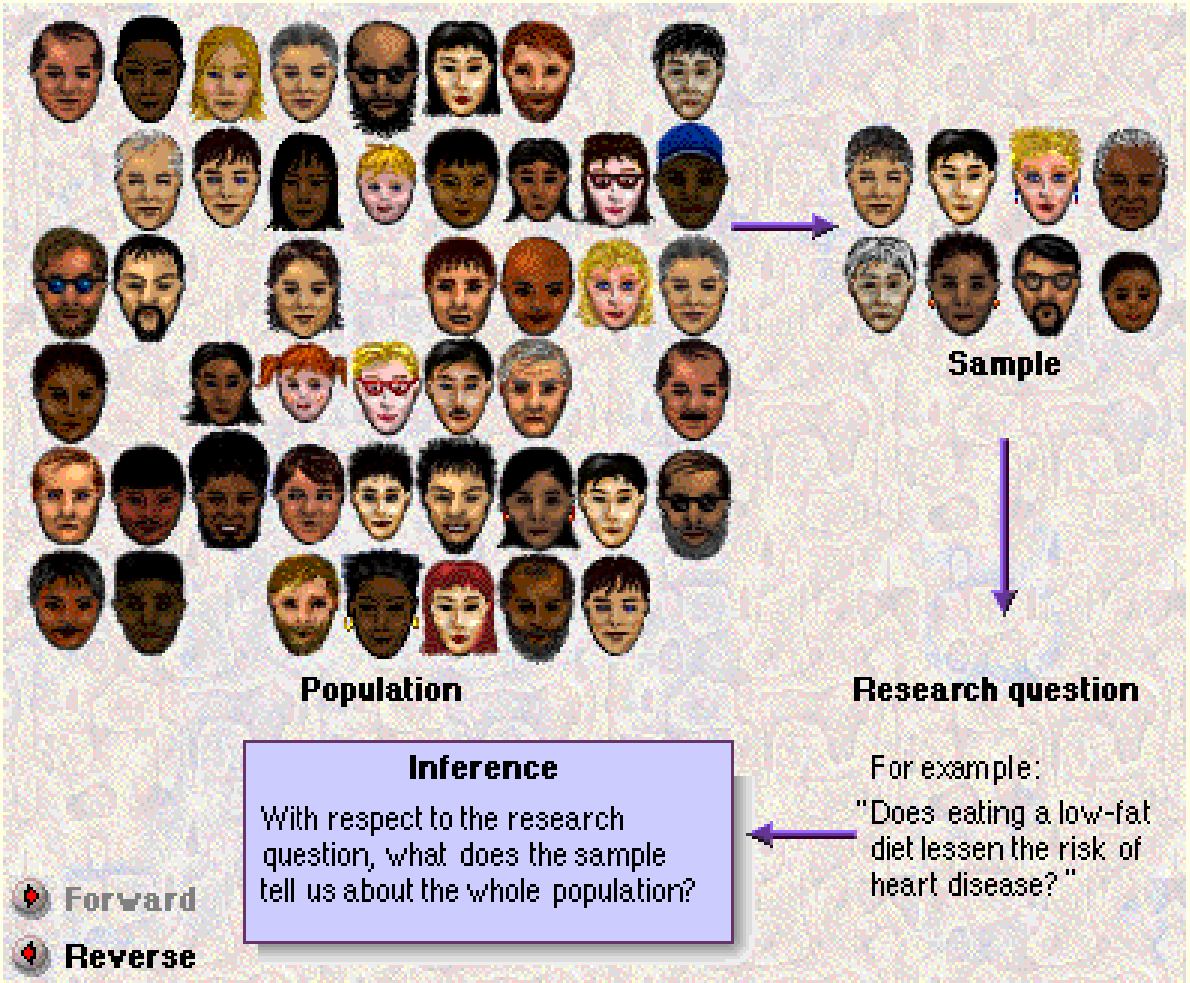
Minta: a populációból a vizsgálat céljára kiválasztott egyedek összessége (a populáció egy részhalmaza).



Megválaszolandó kérdés: Alacsony zsírtartalmú étrend csökkenti-e az infarktus veszélyét?

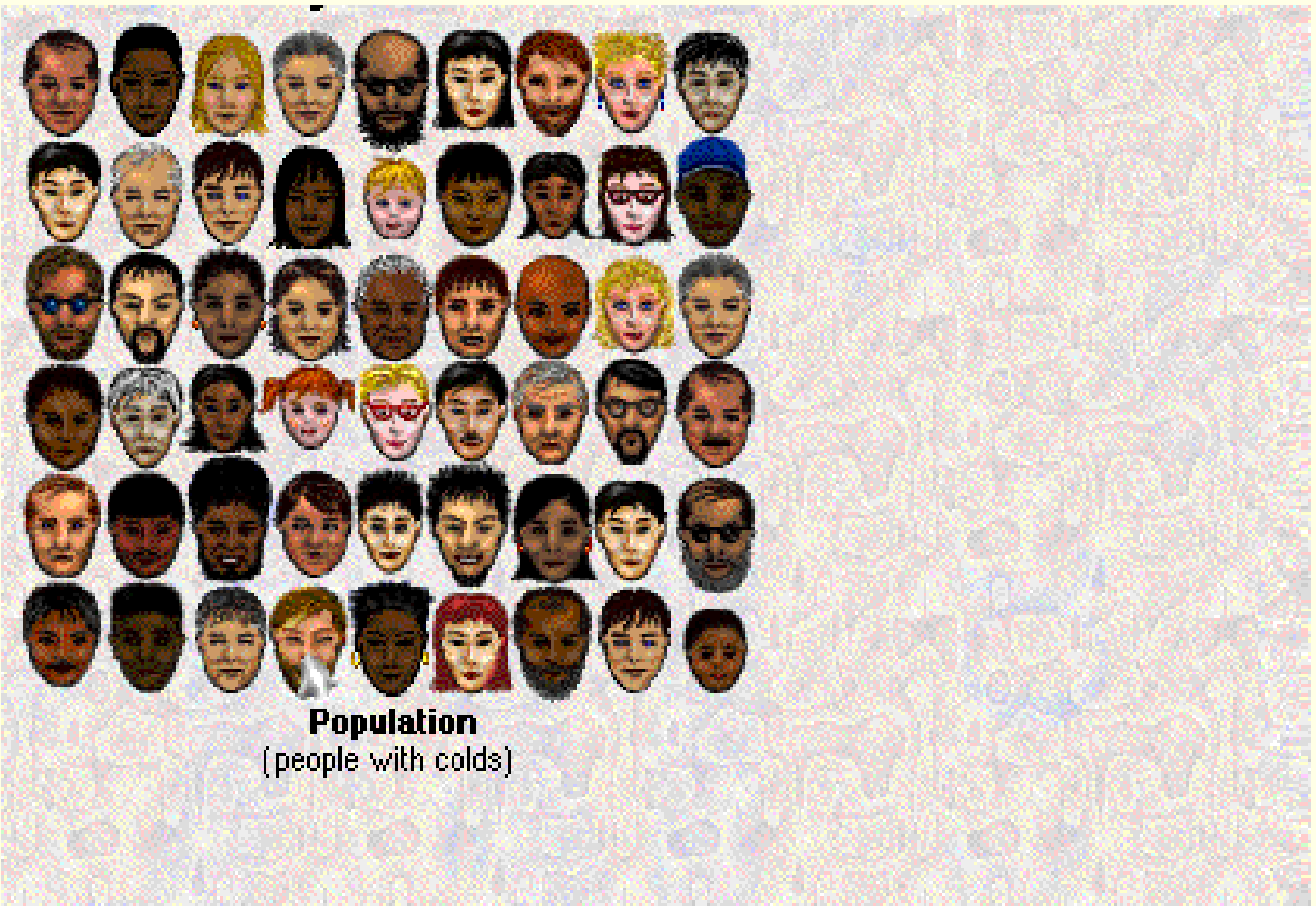


Következtetés: a feltett kérdés szempontjából mit mond számunkra a minta az egész populációról?

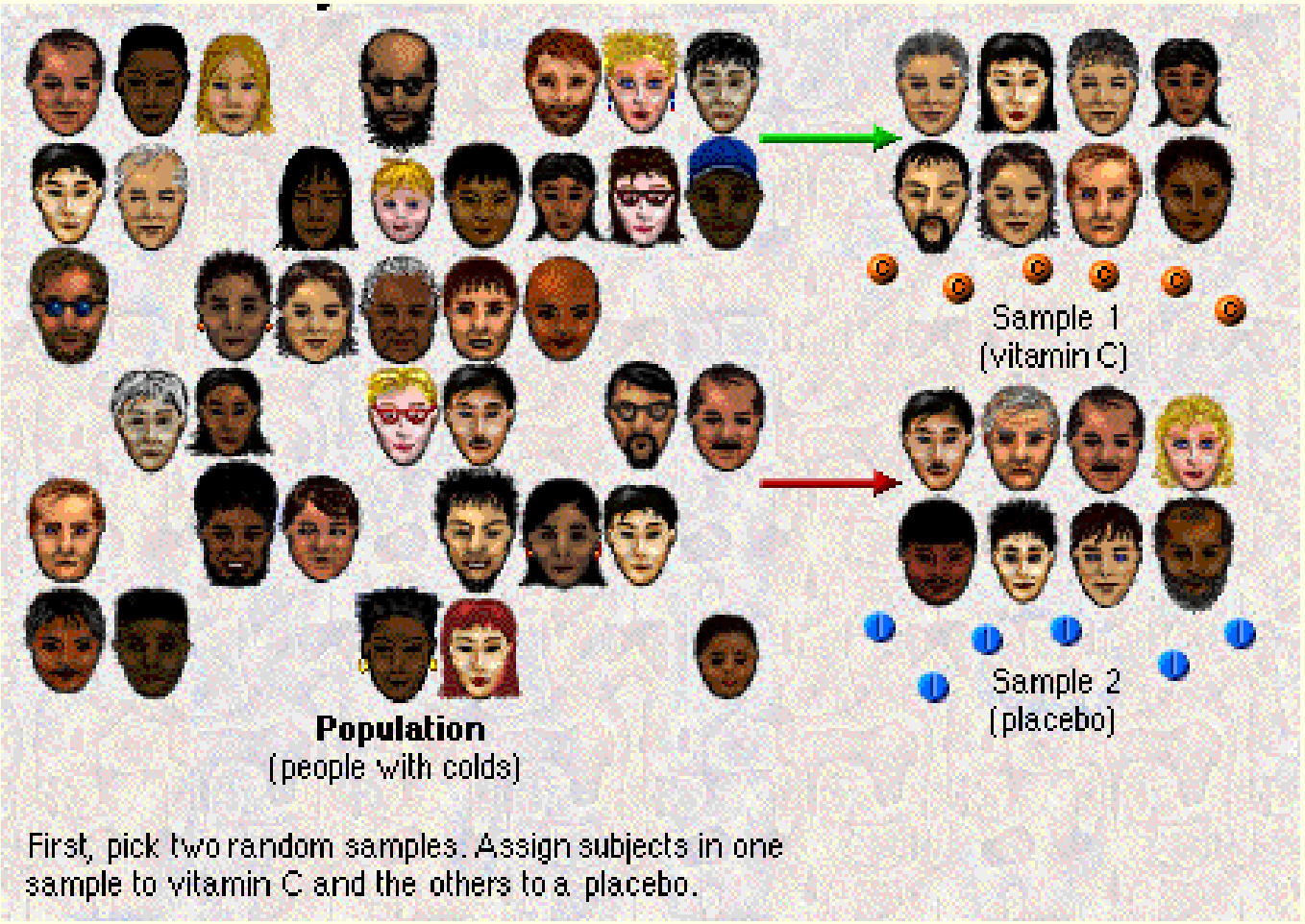


Példa.

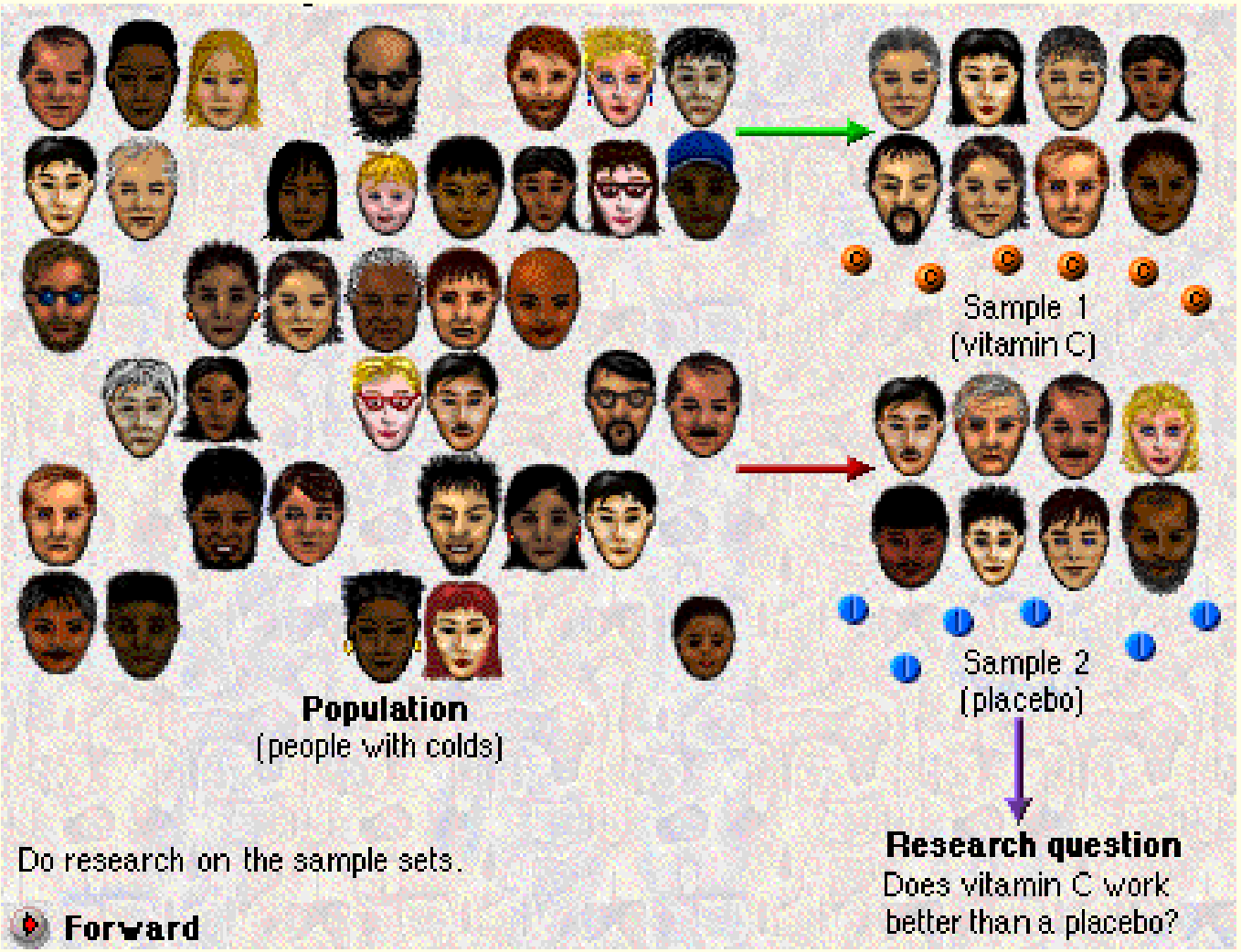
Populáció: a megfázással küszködő emberek



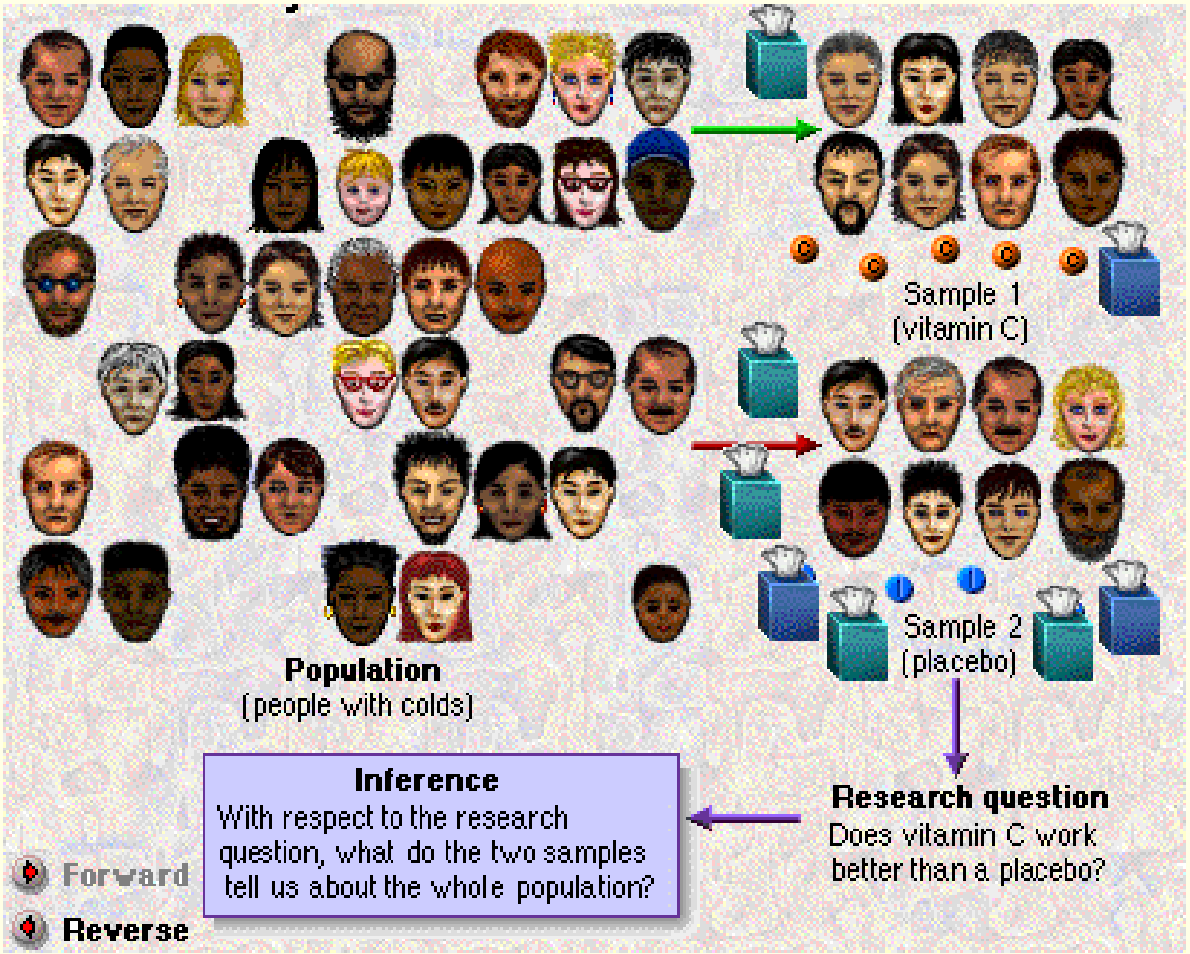
Válasszunk két véletlen mintát. Az egyikben adjunk C vitamint, a másikban placebót az embereknek.



Kérdés: hatásosabb-e a C vitamin, mint a placebo?



Következtetés: mit mutat a két minta a teljes populációra vonatkozóan?



1.1. A minta

Miért van egyáltalán szükségünk egy mintára? Miért nem vizsgáljuk a teljes populációt?

- túl költséges, időigényes
- nem elérhető az összes egyed
- a kísérlet rongálhatja vagy megsemmisítheti az egyedeket.

Egy mintától elvárjuk, hogy hűen **reprezentálja** a teljes populáció tulajdonságait, hiszen csak így lehet

korrekt következtetéseket levonni a vizsgálat eredményéből.

A mintánk legnagyobb valószínűséggel akkor lesz megfelelő, ha a mintaelemeket **véletlenszerűen** választjuk ki a teljes populációból, így a populáció minden tagja egyforma valószínűséggel kerülhet a mintába.

Orvosi kísérletek: gyakran nincs lehetőség a mintába kerülő egyedek véletlenszerű kiválasztására. Például, ha az állatorvos egy bizonyos kezelést csupán azon a húsz kutyán alkalmaz, amelyek az elmúlt három évben kerültek hozzá az adott tünetekkel, akkor egy-

szerűen nincs arra lehetősége, hogy ezt a húsz kutyát véletlenszerűen válassza ki az **összes**, ebben a betegségben szenvedő kutya közül.

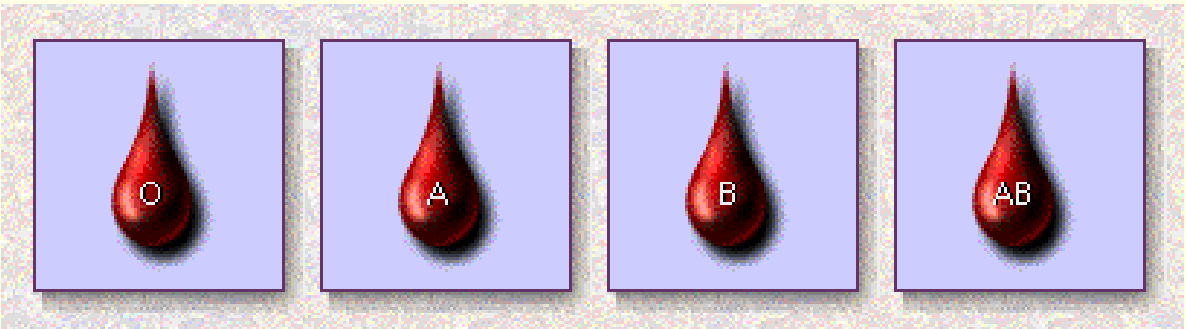
Ilyen esetekben a statisztikai következtetések nyilván nem azt mutatják meg, hogy milyen az ebben a betegségben szenvedő teljes kutyák populációja, hanem azt, hogy milyen az a populáció, amelyből ez a húsz kutya véletlenszerűen lehetett kiválasztva. Mindenféle extrapoláció ezentúl inkább az állatorvos szakmai ítélőképességére van bízva, mint a statisztika módszereire.

1.2. Az adatok

A mintabeli egyedekről a kísérlet során adatokat gyűjtünk. Két fő adattípust különböztetünk meg:

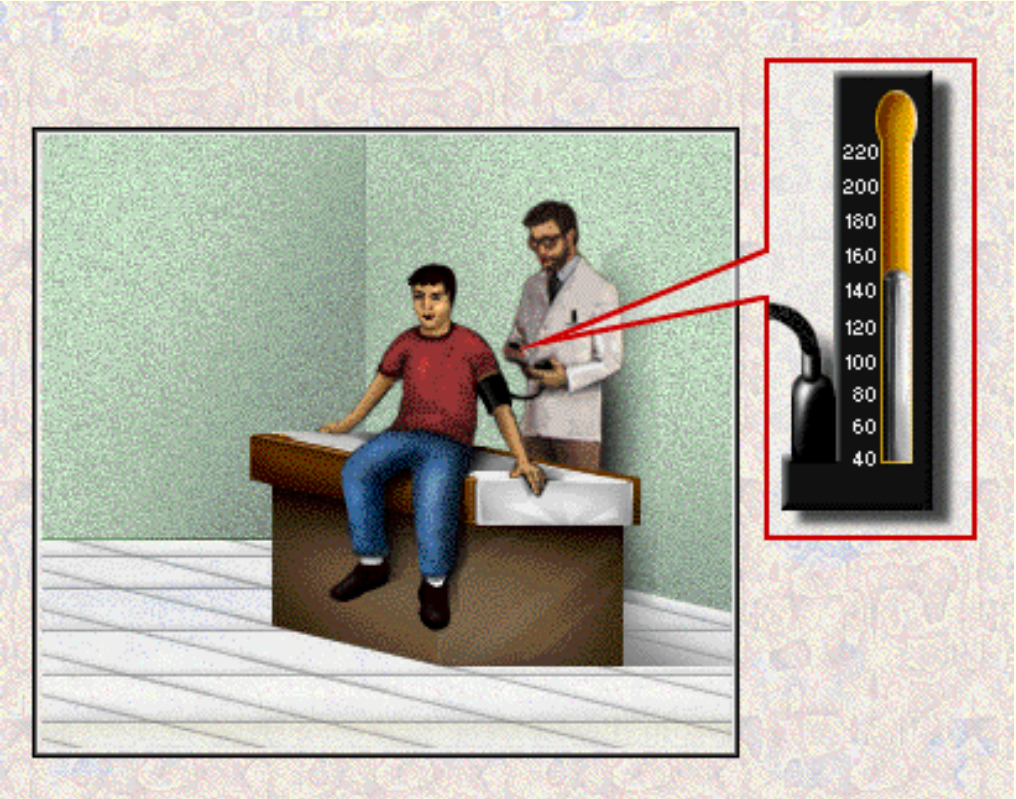
- **Kvalitatív adatok**: ezek nem fejezhetők ki mennyiségekként, inkább **kategóriák**.

Példa: vércsoportok



- **Kvantitatív adatok**: ezek valamilyen mérés eredményei (**mennyiségek**).

Példa: vérnyomás



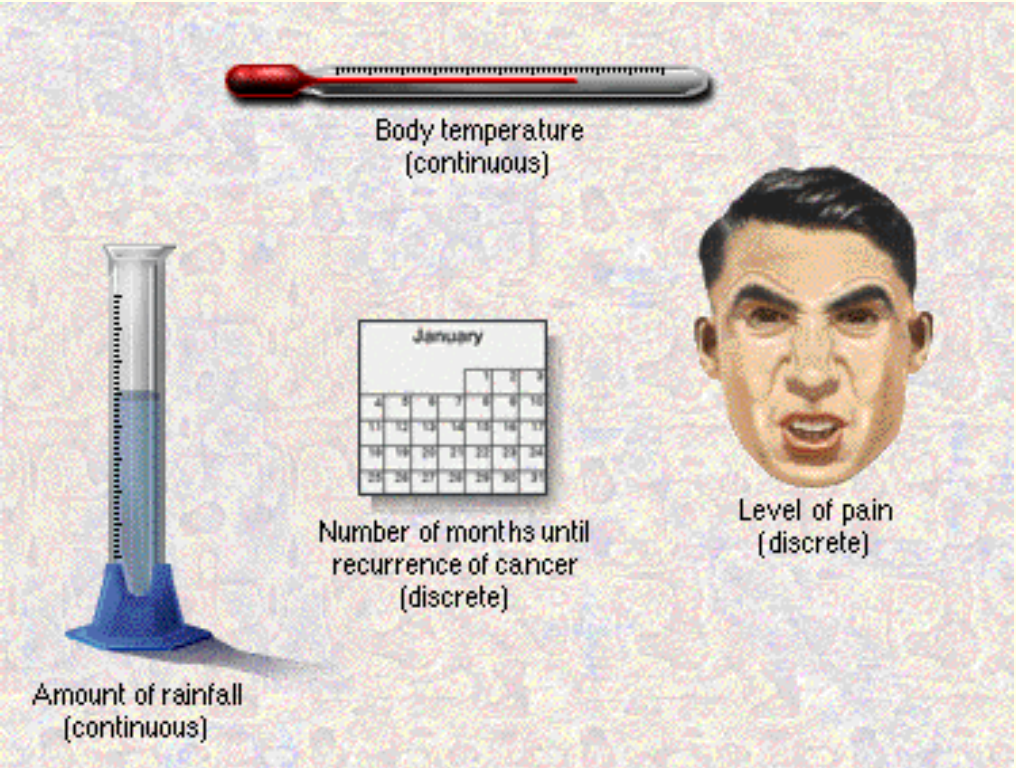
A kvantitatív adatok is kétfélék lehetnek:

- **folytonosak** (a mérési skála egy intervallum), vagy
- **diszkrétek** (a mérhető értékek a számegyenes izolált pontjai).

Példák:

- testhőmérséklet (folytonos)
- egy betegség újbóli megjelenéséig eltelt hónapok száma (diszkrét)

- lehullott csapadékmennyiség egy adott napon (folytonos)
- a fájdalom szintjei (pl. enyhe, közepes, nagy) (diszkrét).



1.3. Mérési skálák

Nominális skála: Egyszerű kategorizálási lehetőséget jelent. Az egyes tulajdonságoknak nevet adunk, és az objektumokat e nevek valamelyikével kapcsoljuk össze, (növényevő ill. húsevő állatok stb.).

Ez a csoportosítás csupán kvalitatív jellegű, számszerű információt nem tartalmaz, és összehasonlításra sem alkalmas.

Példa: **hideg** vagy **meleg** csap

Ordinális skála: a mért vagy megfigyelt értékeket **rangsorolhatjuk**, **összehasonlíthatjuk**.

Példa: az intelligencia-hányados (I.Q.=Intelligence Quotient), vagy az 1-től 5-ig terjedő osztályzati skála.

Nem véletlen, hogy nem létezik sem nulla I.Q., sem nullás osztályzat, mert egyik sem nulla intelligenciához vagy tudáshoz viszonyít, és a rangszámok közötti különbségek sem egyenlőek a különböző szinteken.

Intervallumskála: az adatok közötti **különbség** értelmes. Nincs természetes zérus pont.

Példa: hőmérsékletmérés Celsius fokokban. Itt a nulla fokot önkényesen választottuk (a víz fagyáspontja 1 atm nyomásnál), de az 1 °C hőmérséklet emelkedés már nem önkényes, ui. a hőmérő higanyszála mindig ugyanannyit emelkedik 1 °C hőmérséklet emelkedésnél, akármilyen szinten is mérjük ezt.

Abszolút skála: ez olyan intervallumskála, ahol van természetes nulla pont. Az adatok különbsége és hányadosa is képezhető.

Példa: a **súly**, ahol a mérés mindig a nullától indul; a **hőmérsékletet Kelvin fokban**, (az abszolút kezdet a -273 °C, ami 0 °K).

2. Adatok összegzése

2.1. Gyakorisági táblázatok

Ahhoz, hogy mért, illetve megfigyelt numerikus adataink **áttekinthetőek** és értékelhetőek legyenek, először bizonyos szempontok szerint csoportosítani, illetve összefoglalóan ábrázolni kell őket.

A csoportosítás legtöbbször **gyakorisági táblázat** formájában történik.

1. lépés: az **osztályok**, vagy **osztályintervallumok** kialakítása (az adatokat diszjunkt csoportokba soroljuk).

Diszkrét változóra a lehetséges értékek (ha nincs belőlük túl sok) adnak egy-egy osztályt.

Folytonos változókra (vagy sok értéket felvevő diszkrétekre):

- nagyság szerint növekvőleg rendezzük az adatokat,
- eldöntjük, hogy hány osztály legyen (5 és 15

közötti szám célszerű, mindegyikbe kívánatos legalább 6 adatnak esnie),

- meghatározzuk az osztályintervallumok végpontjait,
- az osztályokba az alsó határ is beletartozik.

2. lépés: összeszámoljuk, hogy az egyes osztályokba hány adat esik (gyakoriság).

3. lépés: kiszámítjuk a relatív, és ha szükség van rá, a százalékos gyakoriságokat.

Ha kiszámítjuk egy minta esetén ezeket a gyakoriságokat, akkor ezzel megadjuk a minta **gyakorisági eloszlását**. Ezzel a gyakorisági eloszlással közelítjük az elméleti valószínűségi eloszlást.

Példa: anyakocák szaporaságának vizsgálata. 105 almot figyeltünk meg, és azokban a születéskori alomszámot jegyeztük fel. A gyakorisági táblázat:

Osztály (alomszám)	Egyedszám (db)	Relatív gyakoriság	Rel. %
6	3	0.029	2.9
7	7	0.067	6.7
8	12	0.114	11.4
9	20	0.190	19.0
10	31	0.295	29.5
11	17	0.162	16.2
12	8	0.076	7.6
13	5	0.048	4.8
14	2	0.019	1.9
Összesen:	105	1.000	100

2.2. Adatok grafikus reprezentációja

A grafikus reprezentációnál az adatokat szemléletes módon összegezzük.

2.3. Kvalitatív adatok

Kvalitatív adatokat összegző grafikus megjelenítésnek két dolgot feltétlenül mutatnia kell: a **kategóriákat**, és hogy ezek **milyen gyakoriságúak**.

- **Oszlopdiagram (bar chart)**

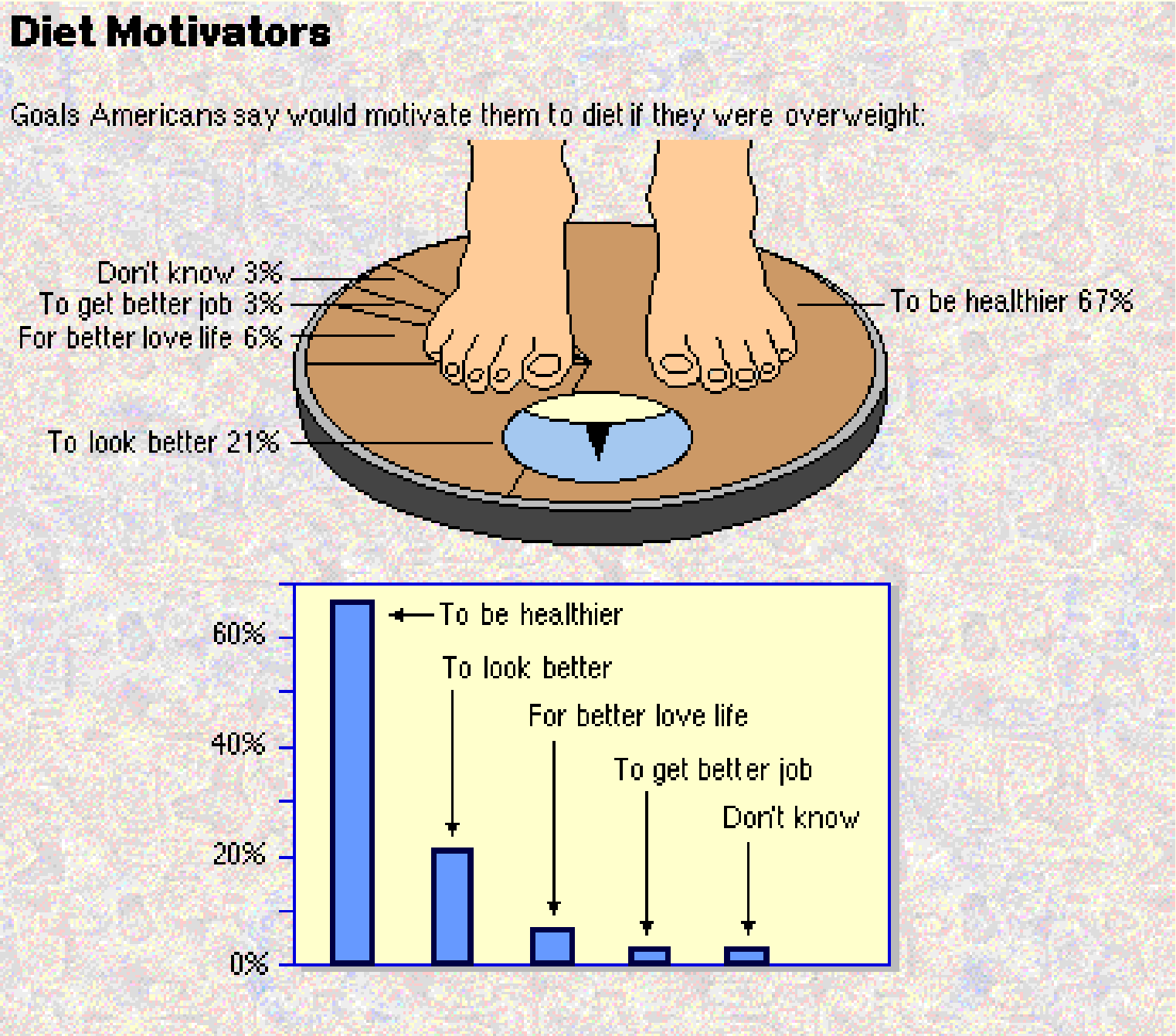
Minden kategóriának megfelel egy téglalap (oszlop), amelynek magassága az adott kategóriába tartozó adatok arányát fejezi ki az összes között.

Példa.

„Ha Ön túlsúlyos lenne, milyen célok ösztönöznék arra, hogy fogyókúrázni kezdjen?”

- egészségesebb legyek
- jobban nézzek ki

- javuljon a szerelmi életem
- jobb állást kapjak
- nem tudom.


Diet Motivators
Goals Americans say would motivate them to diet if they were overweight:
Don't know 3%
To get better job 3%
For better love life 6%
To look better 21%
To be healthier 67%
60%
40%
20%
0%
To be healthier
To look better
For better love life
To get better job
Don't know

- **Tortadiagram (pie chart)**

A fenti ábrán is látható, hogy a torta egy-egy szelete reprezentálja az egyes kategóriákat, míg a szelet területe a kategória arányát.

2.4. Kvantitatív adatok

Ilyen adatokra vonatkozó grafikonok két dolgot mutatnak: a mért **mennyiségeket** (számokat), és hogy ezek **milyen gyakran** fordulnak elő.

- **Pontdiagramok (dotplots)**

Ezek minden egyes adatot egy a mért szám fölé rajzolt ponttal jelenítenek meg.

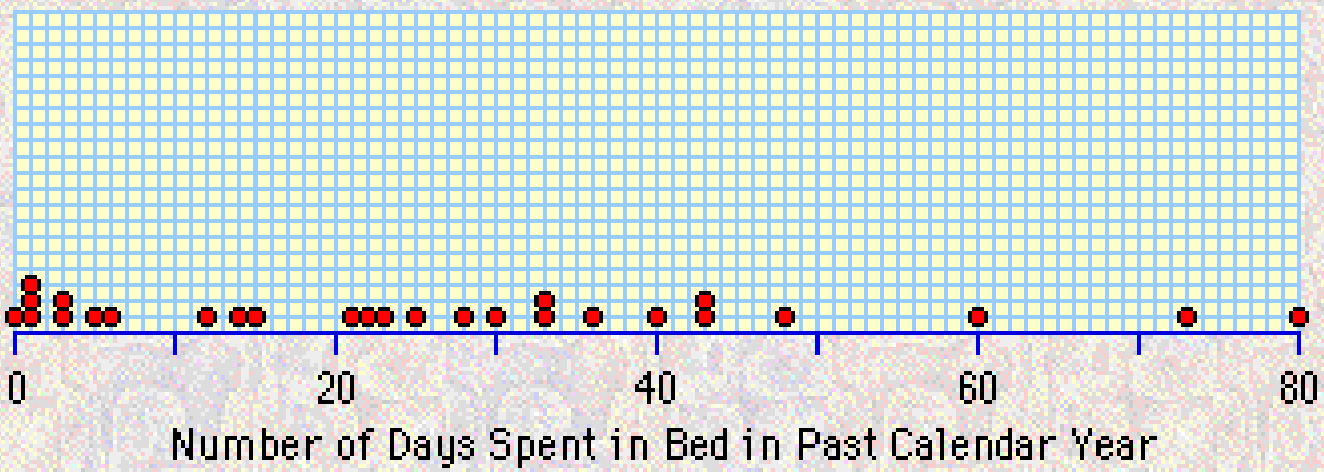
Példa.

Az elmúlt évben betegség miatt ágyban töltött napok száma egy nagyon idős populációban:

Building a Dotplot

Number of Days Spent in Bed Due to Illness in Past Calender Year in an Extremely Elderly Population

Individual	Days	Individual	Days	Individual	Days	Individual	Days
No. 1	14	No. 8	33	No. 15	43	No. 22	36
No. 2	60	No. 9	73	No. 16	80	No. 23	43
No. 3	01	No. 10	30	No. 17	03	No. 24	40
No. 4	33	No. 11	48	No. 18	05	No. 25	25
No. 5	15	No. 12	03	No. 19	01	No. 26	23
No. 6	00	No. 13	28	No. 20	12	No. 27	22
No. 7	21	No. 14	01	No. 21	06		



A pontdiagramok

- könnyen megrajzolhatók, ha nincs túl sok adatunk
- jól mutatják, hogy hol tömörülnek az adatok, vannak-e kiütő értékek, illetve olyan tartományok, ahol nincs adat
- a rendezésen alapuló jellemzőket jól jelenítik meg (pl. medián)
- hátránya, hogy a gyakoriságok nehezen olvashatók le az ábráról.

- **Hisztogramok**

Emlékezzünk vissza a gyakorisági táblázatra a fenti példából:

Osztály (alomszám)	Egyedszám (db)	Relatív gyakoriság	Rel. %
6	3	0.029	2.9
7	7	0.067	6.7
8	12	0.114	11.4
9	20	0.190	19.0
10	31	0.295	29.5
11	17	0.162	16.2
12	8	0.076	7.6
13	5	0.048	4.8
14	2	0.019	1.9
Összesen:	105	1.000	100

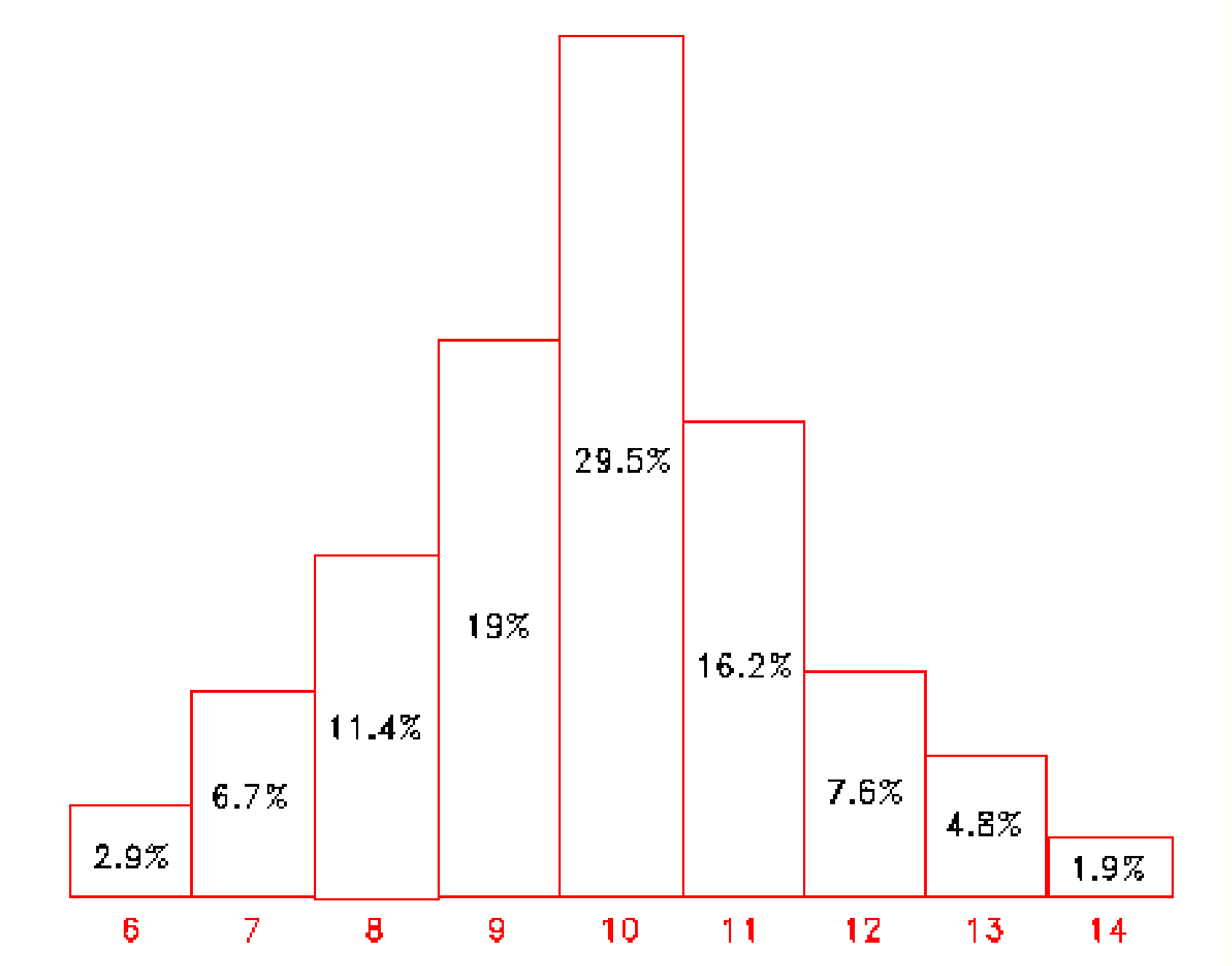
A **hisztogram** alapján próbálunk következtetni arra, hogy hogyan néz ki a változónk sűrűségfüggvénye.

A hisztogram vízszintes tengelyén ábrázoljuk az osztályintervallumokat, és ezek fölé olyan téglalapokat rajzolunk, melyek területe megegyezik a megfelelő relatív, vagy százalékos gyakorisággal (azaz az egyes téglalapok magassága egyenlő a megfelelő gyakoriság elosztva az intervallum hosszával). Így a hisztogram teljes területe 1, vagy 100% lesz.

Diszkrét változó esetén úgy járunk el, hogy a változó értékei az intervallumok közepén helyezkednek el.

A hisztogram – ha a minta elemszámát növeljük – közelíti a valószínűségi változó elméleti sűrűségfüggvényét. Az általunk tárgyalt módszerek többsége akkor használható jól, ha a hisztogramunk jól közelíti a normális eloszlás sűrűségfüggvényét, azaz nagyjából haranggörbe alakú.

A példában:


2.9%
6.7%
11.4%
19%
29.5%
16.2%
7.6%
4.8%
1.9%
6
7
8
9
10
11
12
13
14

A gyakorlatban a konkrétan végrehajtott mintavételezés eredményéül számok adódnak: $x_1, x_2, \ldots, x_n$. Ha a kísérletet újra meg újra megismételjük, akkor más és más számok. Ezek maguk is a véletlentől függnek.

Az előzőekben bemutatott néhány grafikus ábrázolás helyett csupán egy-két összesítő mennyiséget, ún. **statisztikát** használunk.

Adataink egy n-változós $s(x_1, \ldots, x_n)$ függvényét **statisztikai függvénynek** (röviden **statisztikának**) nevezzük.

Ennek értéke is a véletlentől függ, tehát valószínűségi változó.

Most áttekintjük a leggyakoribb statisztikákat.

2.5. Középértékek

Legyen a mintánk $x_1, x_2, \ldots, x_n$.

- **Átlag**

A megfigyeléseink **átlagát**, amelyet $\overline{x}$ jelöl, úgy számítjuk ki, hogy az értékek összegét elosztjuk a minta elemszámával:

$$\overline{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}.$$

Az anyakocák szaporaságáról szóló példában az átlag $\overline{x} = 9.8$.

Az átlag olyan érték, amelyik összességében a ,,leg-

közelebb" van a minta értékeihez. Ha kiszámítjuk a mintaértékek és az átlagérték közti összes eltérések összegét, akkor nullát kapunk eredményül:

$$\sum_{i=1}^{n}(x_i - \overline{x}) = \sum_{i=1}^{n} x_i - n\overline{x} = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i = 0.$$

Az átlagértéket nem mindig lehet tipikus értékként kezelni.

Példa. Egy kisvárosban 20000 ember él, akik közül kettő multimilliomos az összes többi pedig 8000 és

15000 Ft közötti összeget keres. Ekkor átlagkeresetként 100000 Ft körüli összeg jöhet ki. Ez azt sugallja, hogy a kisváros átlagos polgára ilyen jól keres, ami egyáltalán nincs így.

Ezek szerint az átlagérték csak azt mondja meg, hogy mennyi lenne a keresetük, ha az összkeresetet egyenletesen osztanánk el. Az átlag érzékeny a **kiütő értékekre** (angolul: **outliers**) (ezek az adatok zöméhez képest nagyon nagy, vagy nagyon kis értékek).

Ha a változónk normális eloszlású, akkor az átlag felfogható tipikus értékként.

- **Medián**

A **mediánt** úgy kapjuk, hogy először sorbarendezzük az adatokat:

$$x_1 \leq x_2 \leq \ldots \leq x_{n-1} \leq x_n,$$

majd az így sorbarendezett adatok közül kiválasztjuk a **középsőt**.
Ha két középső van, azaz n páros, akkor a **két középső átlaga** lesz a medián.

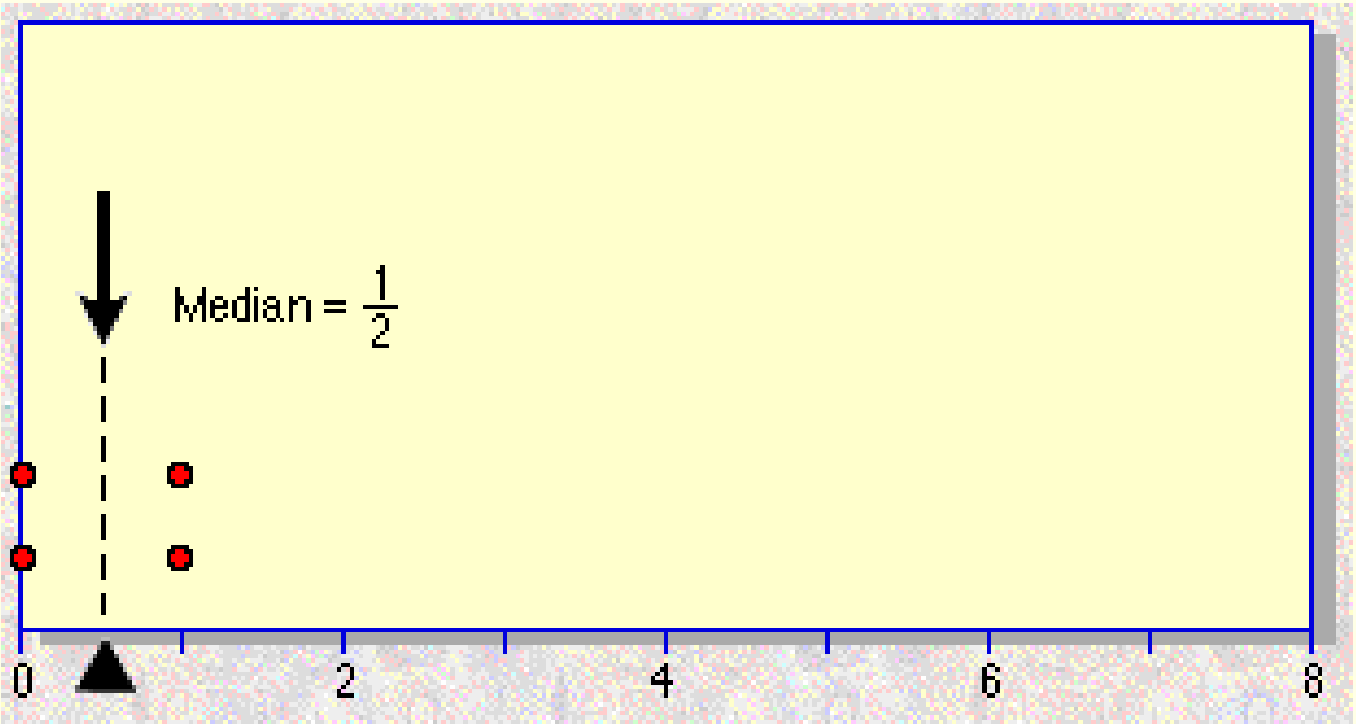
Formálisan:

$$x_{\text{med}} = \begin{cases} x_{k+1} & \text{ha } n = 2k+1, \\ \frac{x_k + x_{k+1}}{2} & \text{ha } n = 2k. \end{cases}$$

Az anyakocák szaporaságának példáját tekintve, a medián 10.

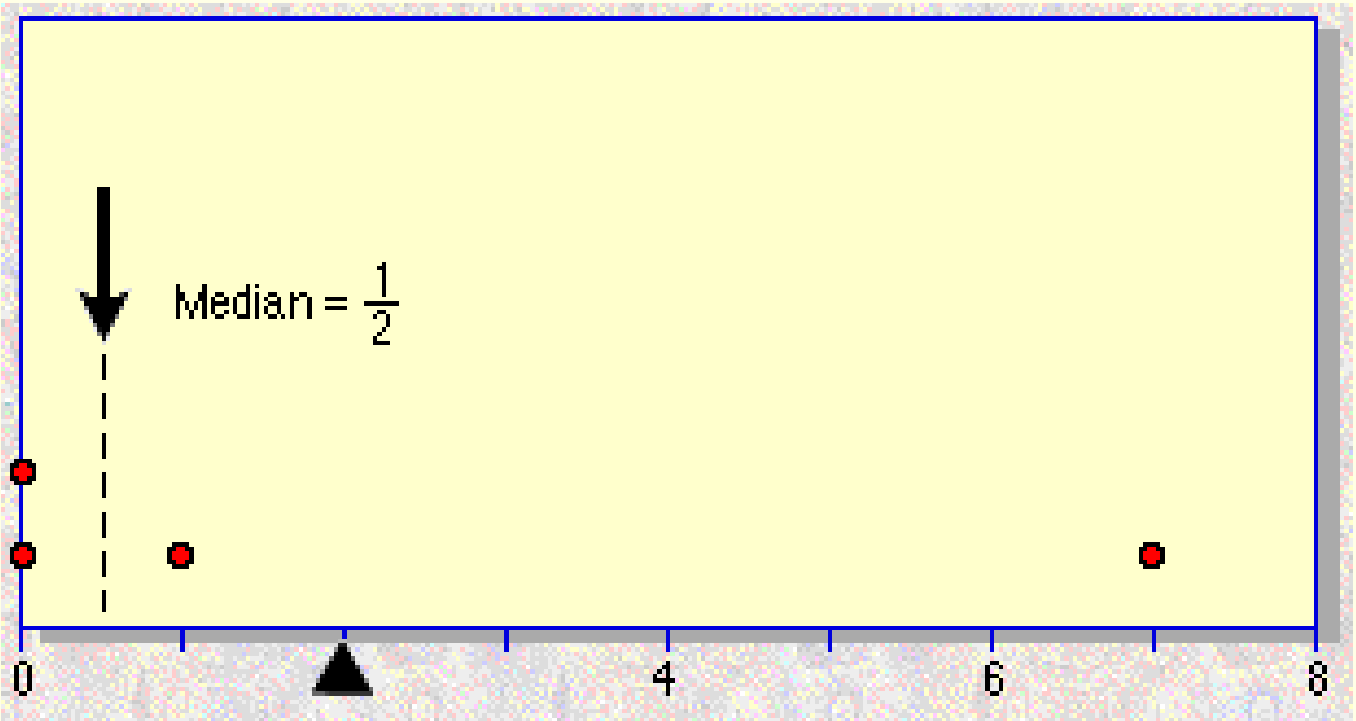
Szimmetrikus eloszlásokra az átlag és a medián egyenlő. Ferde eloszlásokra a medián a középponthoz közelebb helyezkedik el, az átlag pedig a hosszabb farok irányába tolódik ki.

Szimmetrikus eloszlás, páros sok adat.



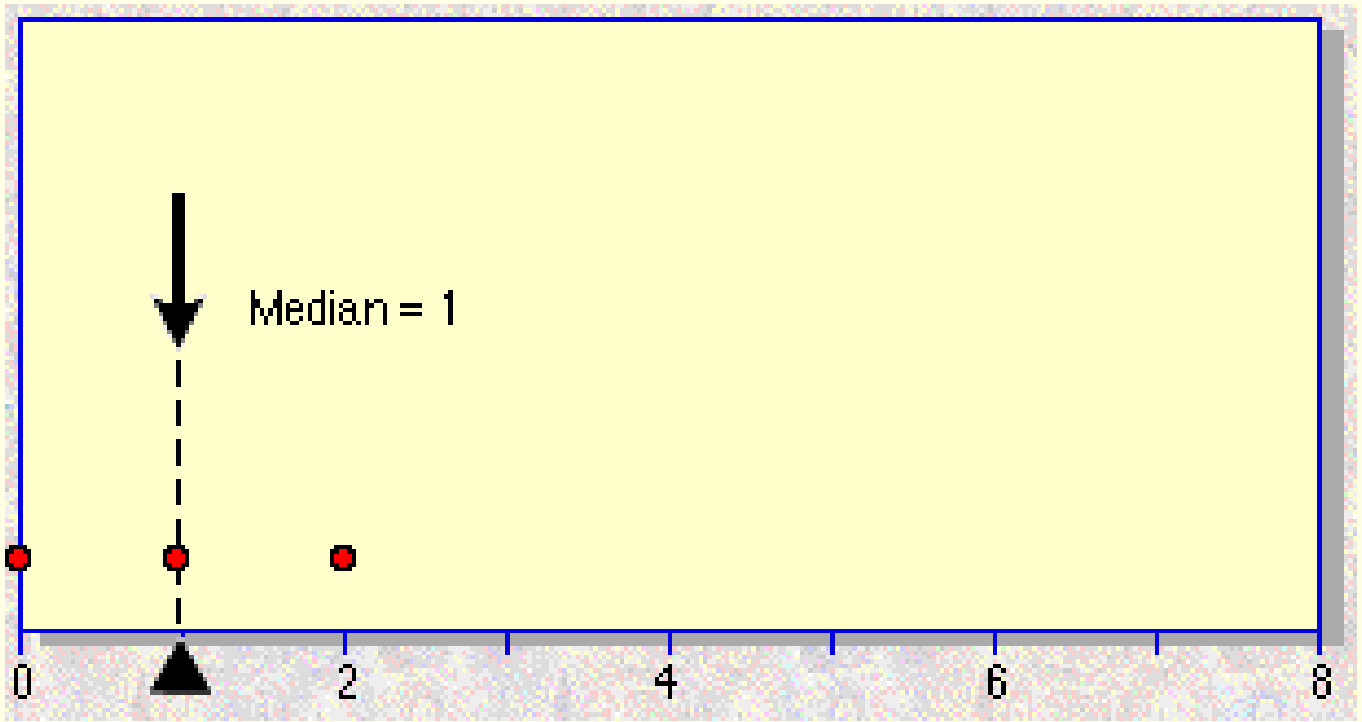
Átlag = medián = 1/2.

Ferde eloszlás, páros sok adat.



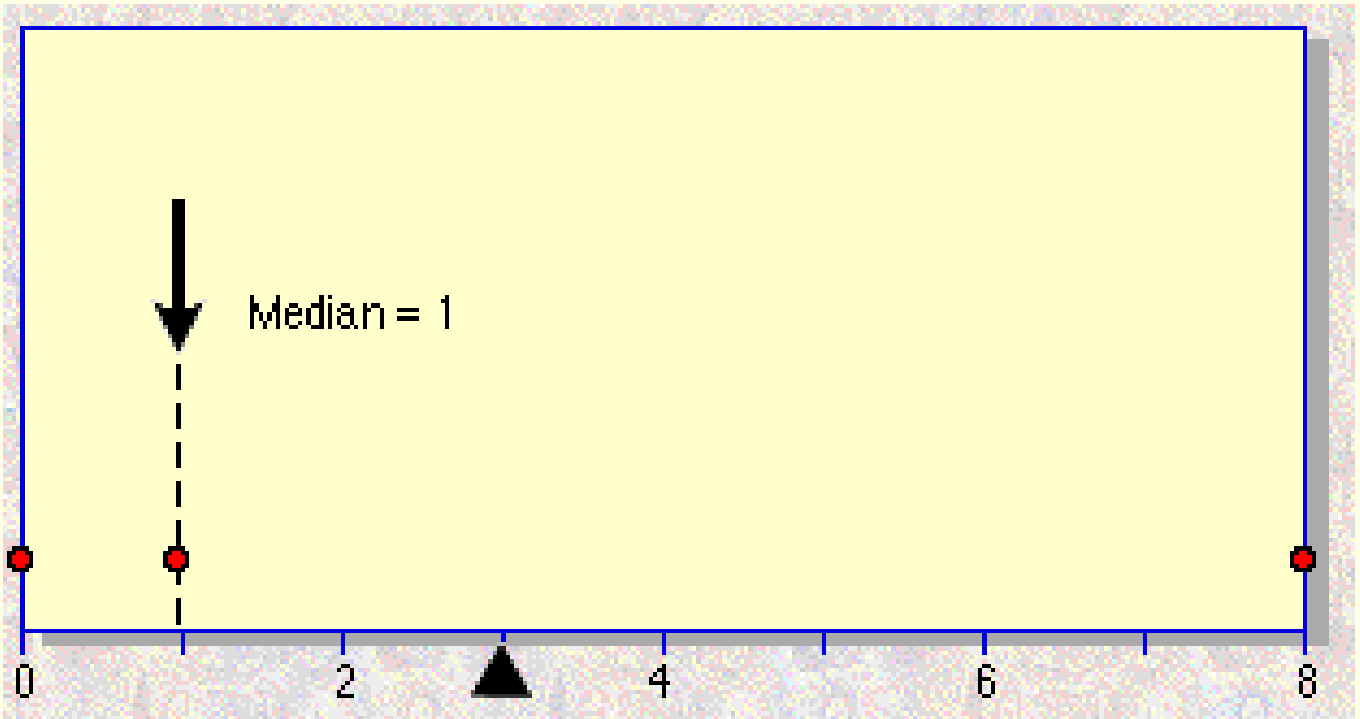
Átlag = 2, medián = 1/2.

Szimmetrikus eloszlás, páratlan sok adat.



Átlag = medián = 1.

Ferde eloszlás, páratlan sok adat.



Átlag $= 3$, medián $= 1/2$.

- **Módusz**

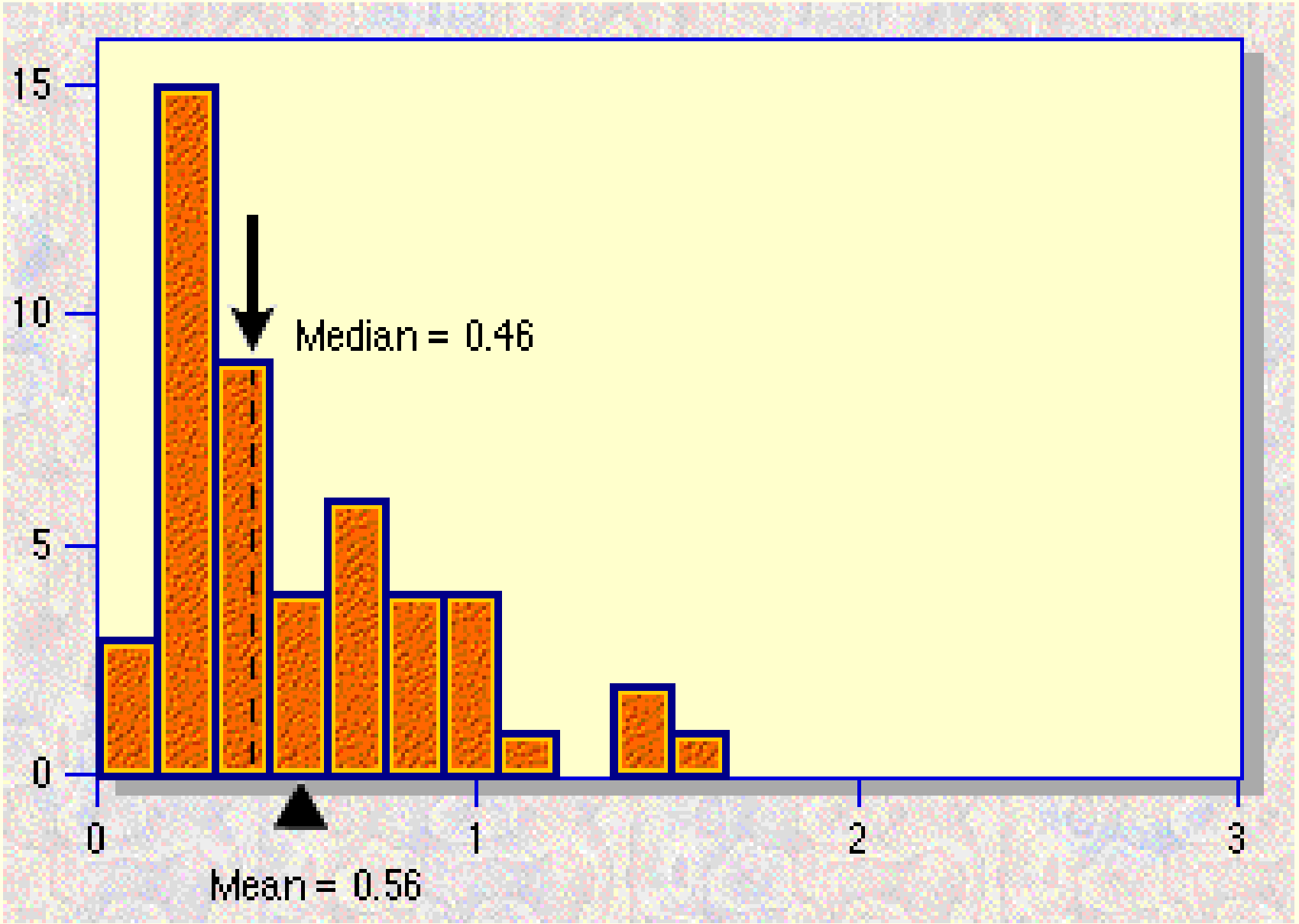
A **módusz** a mintában **leggyakrabban** előforduló érték.

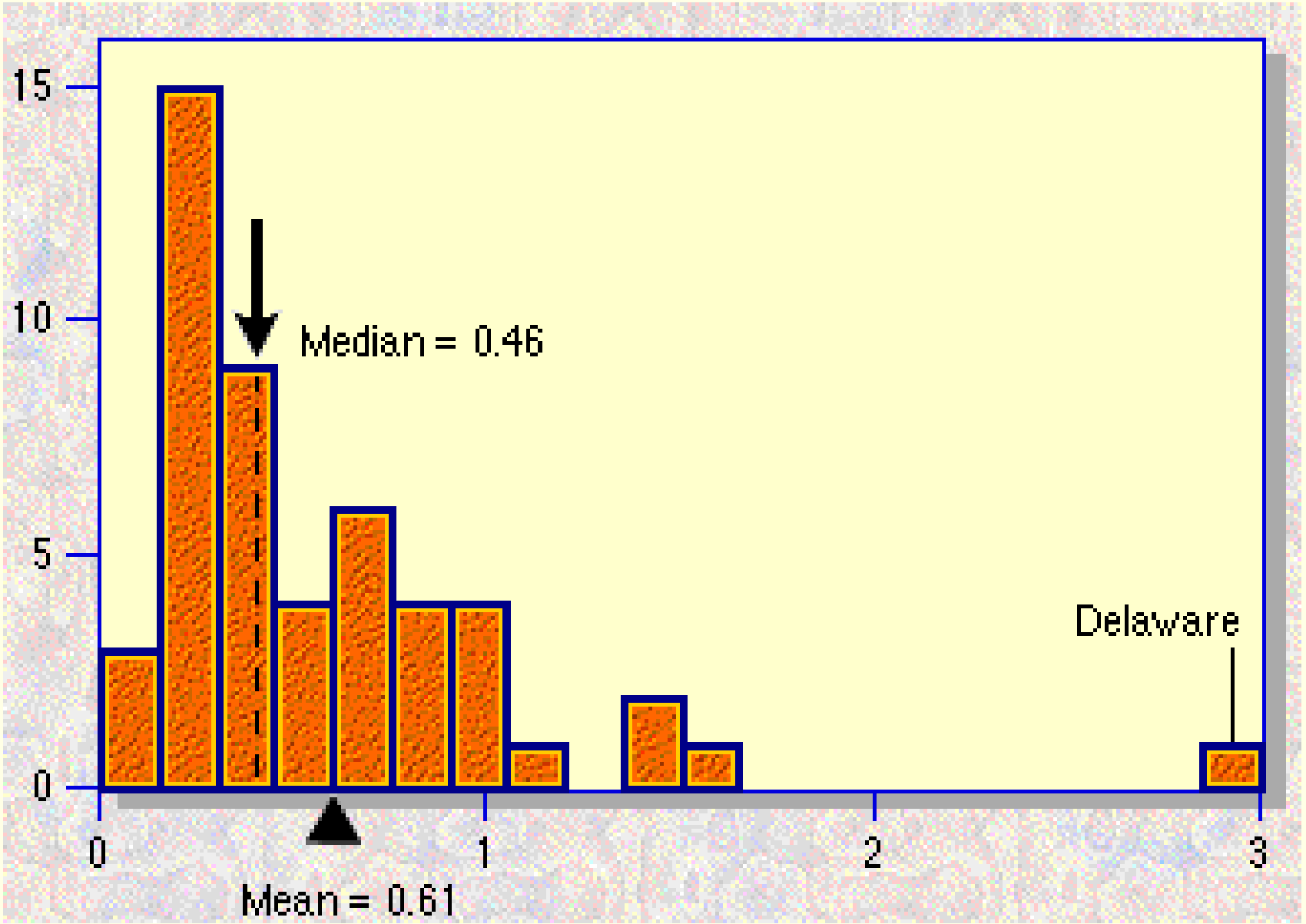
Az anyakocák szaporaságának példájában a módusz is 10.

A középértékek a hisztogramból becsülhetők, bár a becslés nagyon függ az osztályokba sorolástól:

- Az **átlag** az a pont, amelynél a hisztogram súlypont van (azaz ha ennél a pontnál alátámasztjuk a hisztogramot, akkor az egyensúlyban marad.)

- A **mediántól** balra és jobbra a hisztogram területének fele helyezkedik el.
- A **módusz** az az érték, amely fölött a legmagasabb téglalap van.
- Szimmetrikus és egy csúcsú hisztogram esetén a három középérték egybeesik (a szimmetria tengelyre).
- Ha a hisztogram több csúcsú, vagy nem szimmetrikus, akkor ezek az értékek nem esnek egybe.


15
10
5
0
Median = 0.46
0
1
2
3
Mean = 0.56


15
10
5
0
Median = 0.46
Delaware
0
1
2
3
Mean = 0.61

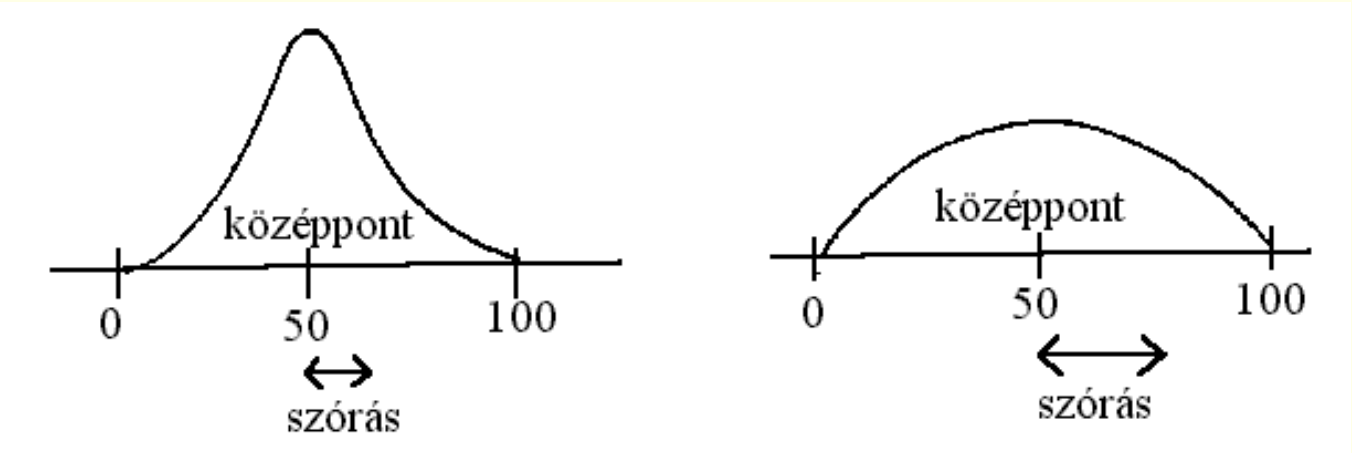
Ha adataink gyakorisági táblázatba vannak rendezve, és az osztályokat x_j-vel, az egyes osztályokban levő adatok számát f_j-vel, és az osztályok számát N-nel jelöljük, akkor ezekkel a mennyiségekkel kifejezve az átlag a következő:

$$\overline{x} = \sum_{j=1}^{N} \frac{f_j \cdot x_j}{n},$$

ahol $n = \sum_{j=1}^{N} f_j$.

2.6. A szóródás mérőszámai

Eddigi statisztikáink csak arra alkalmasak, hogy valamilyen értelemben az eloszlás középpontját jellemezzék. A következő ábrán látszik, hogy ezek **önmagukban** még olyan esetben sem jellemzik jól az eloszlást, ha az szimmetrikus, és csak egy csúcsa van.



Célszerű ezért olyan mérőszámokat is bevezetni, amelyek azt mutatják meg, hogy az adatok hogyan helyezkednek el az átlagérték körül.

- **Terjedelem**

A minta **terjedelme** a minta legnagyobb és legkisebb értéke közötti különbség.

Az anyakocák esetében a terjedelem 14-6=8.

Sajnos, a terjedelem megadása még mindig nem jellemzi elég jól az eloszlást, hiszen arról nem mond semmit sem, hogy **hol milyen sűrűn** helyezkednek

el az adataink.

- **Interkvartilis terjedelem**

Ez a jellemző az adatok középső 50 %-át tartalmazó intervallumnak a hossza.

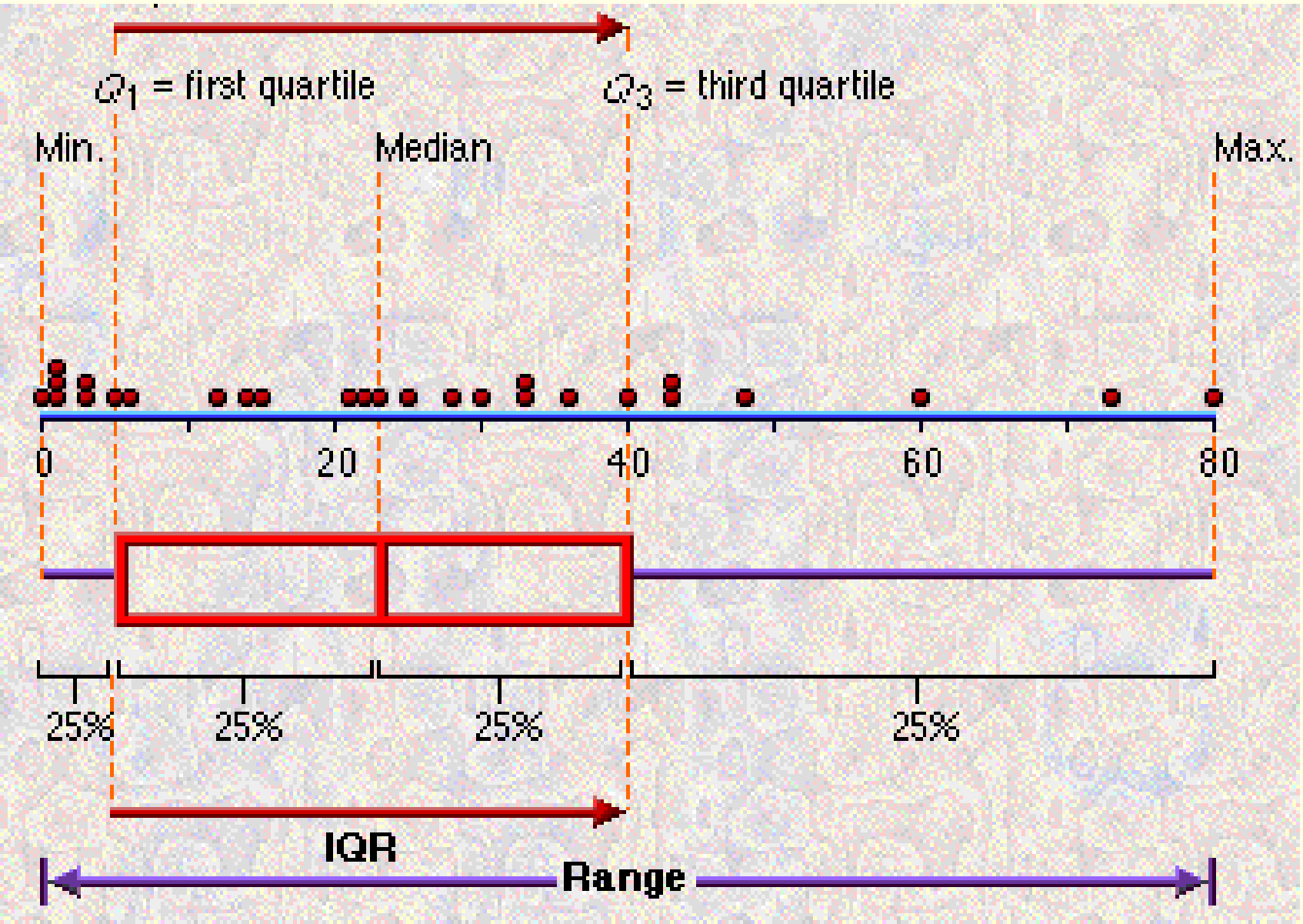
Rendezzük növekvő sorrendbe adatainkat.

Az **alsó kvartilis** egy olyan Q_1 adat, amelytől balra az adatok 25%-a, jobbra 75%-a található.

A **felső kvartilis** egy olyan Q_3 adat, amelytől balra az adatok 75%-a, jobbra 25%-a található.

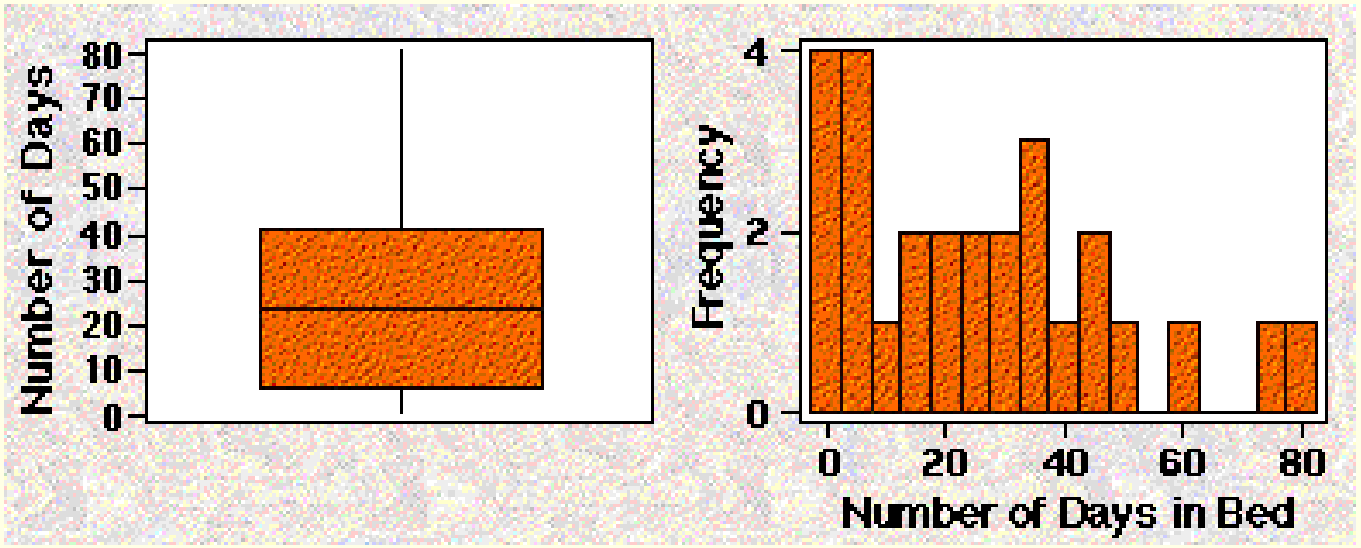
A minta **interkvartilis terjedelme** (röviden: **IQR**) a felső és alsó kvartilis különbsége:

$$IQR = Q_3 - Q_1.$$


Q_1 = first quartile
Q_3 = third quartile
Min.
Median
Max.
0
20
40
60
80
25%
25%
25%
25%
IQR
Range

Itt az ideje megemlíteni egy olyan grafikus adatösszegzési módot, amely a most bevezetett jellemzőket használja. Ez a **doboz diagram** (boxplot).

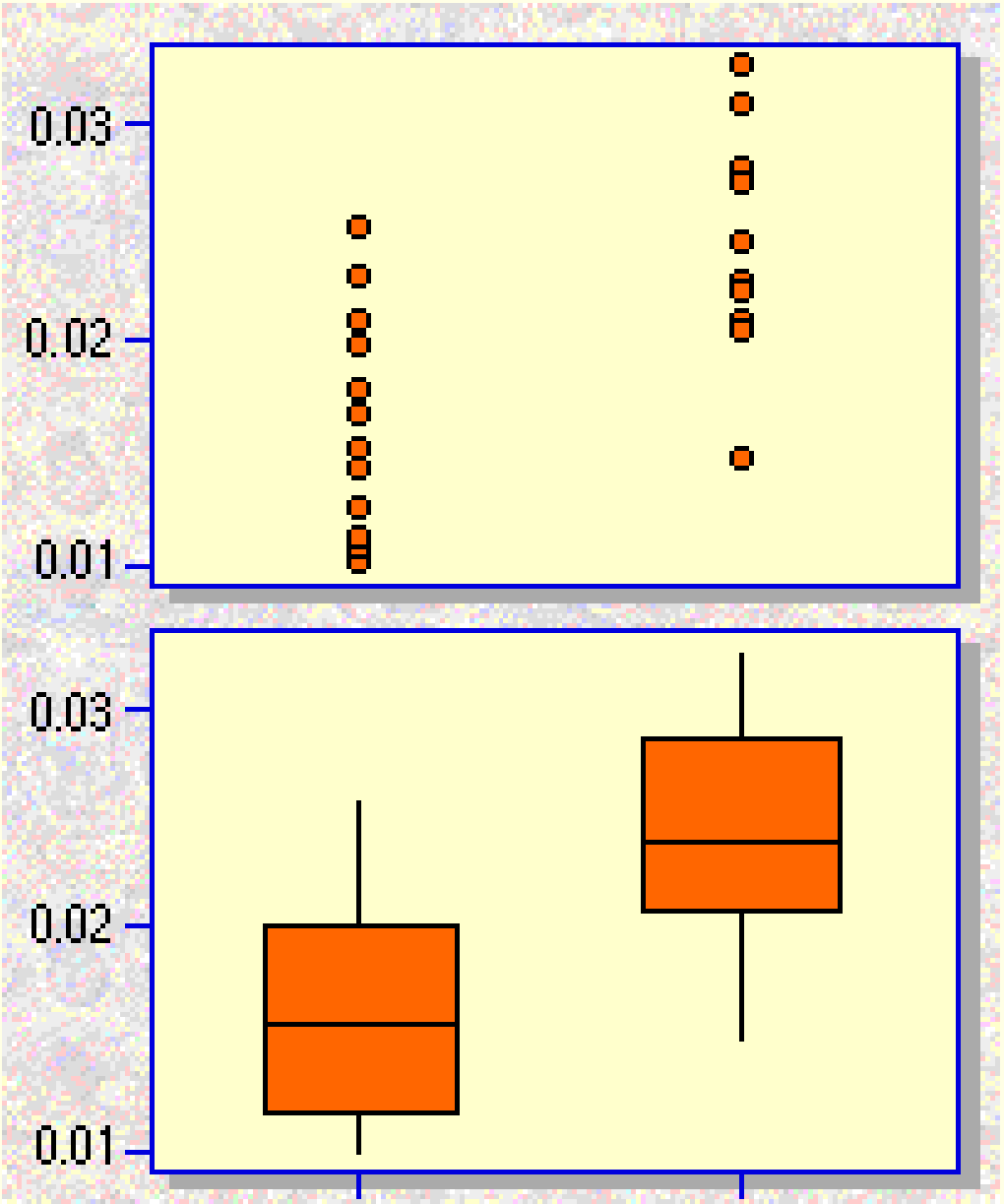
Egy korábban említett példa (betegség miatt ágyban töltött napok száma) doboz diagramja a hisztogrammal együtt:



A doboz diagram az adatok öt jellemzőjén alapul:

- a legkisebb érték (min)
- az alsó kvartilis (Q_1)
- a medián (Q_2)
- a felső kvartilis (Q_3)
- a legnagyobb érték (max).

Nem mutatja az egyes adatokat; nagyon tömör, mégis informatív; akkor a leghasznosabb, ha megfigyelések több csoportjának eloszlásait hasonlítjuk össze.


0.03
0.02
0.01
0.03
0.02
0.01

- **Tapasztalati szórás és variancia**

Az adatok átlag körüli szóródásának jól használható mérőszáma a **korrigált tapasztalati szórás** (angolul: standard deviation, SD) , és ennek négyzete, a **variancia** (vagy szórás).

Vigyázat, ez nem keverendő össze egy valószínűségi változó hasonló nevű paramétereivel. Később látni fogjuk a kapcsolatot a közöttük.

A szórás azt mutatja meg, hogy az adataink átlagosan milyen távol helyezkednek el a számtani középtől (átlagtól).

Adataink: $x_1, x_2, \ldots, x_n$, átlaguk $\bar{x}$.

Az adatok **korrigált tapasztalati szórása** az az s szám, amelyre

$$s := \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\mathbf{n-1}}}.$$

A szórásnak ugyanaz a mértékegysége, mint az eredeti adatainké.

Ha adataink gyakorisági táblázatba vannak rendezve (N csoport van), akkor a szórás:

$$s = \sqrt{\frac{\sum_{i=1}^{\mathbf{N}} f_i \cdot (x_i - \bar{x})^2}{\mathbf{n-1}}} \qquad (n = \sum_{i=1}^{N} f_i).$$

Bizonyos esetekben a szórás kiszámítása egyszerűbb, ha bevezetjük az eltérés négyzetösszeget: **SQ** (sum of squares of deviations):

$$SQ := \sum_{i=1}^{n} (x_i - \bar{x})^2.$$

Könnyen láthtó, hogy ekkor

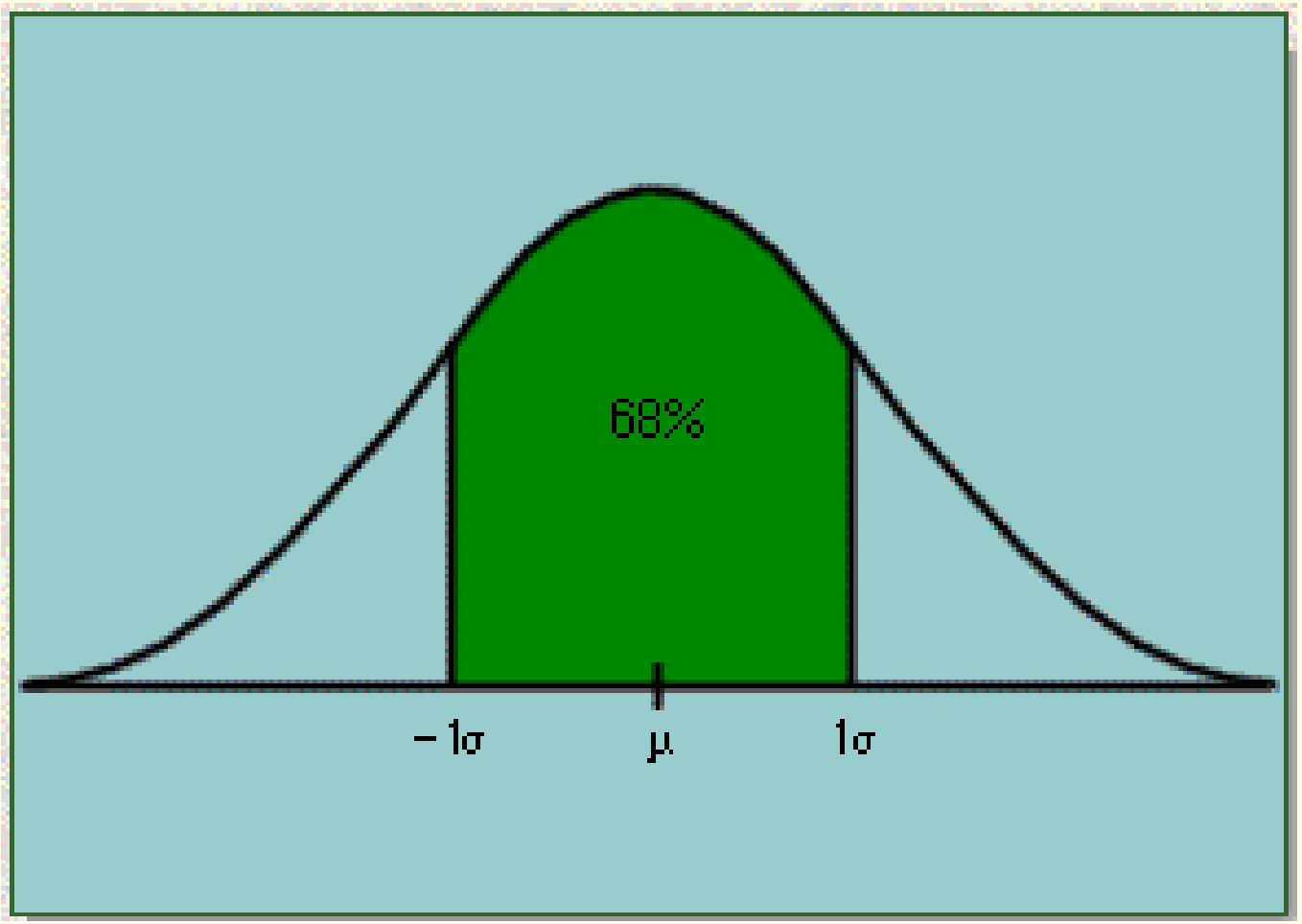
$$SQ = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n},$$

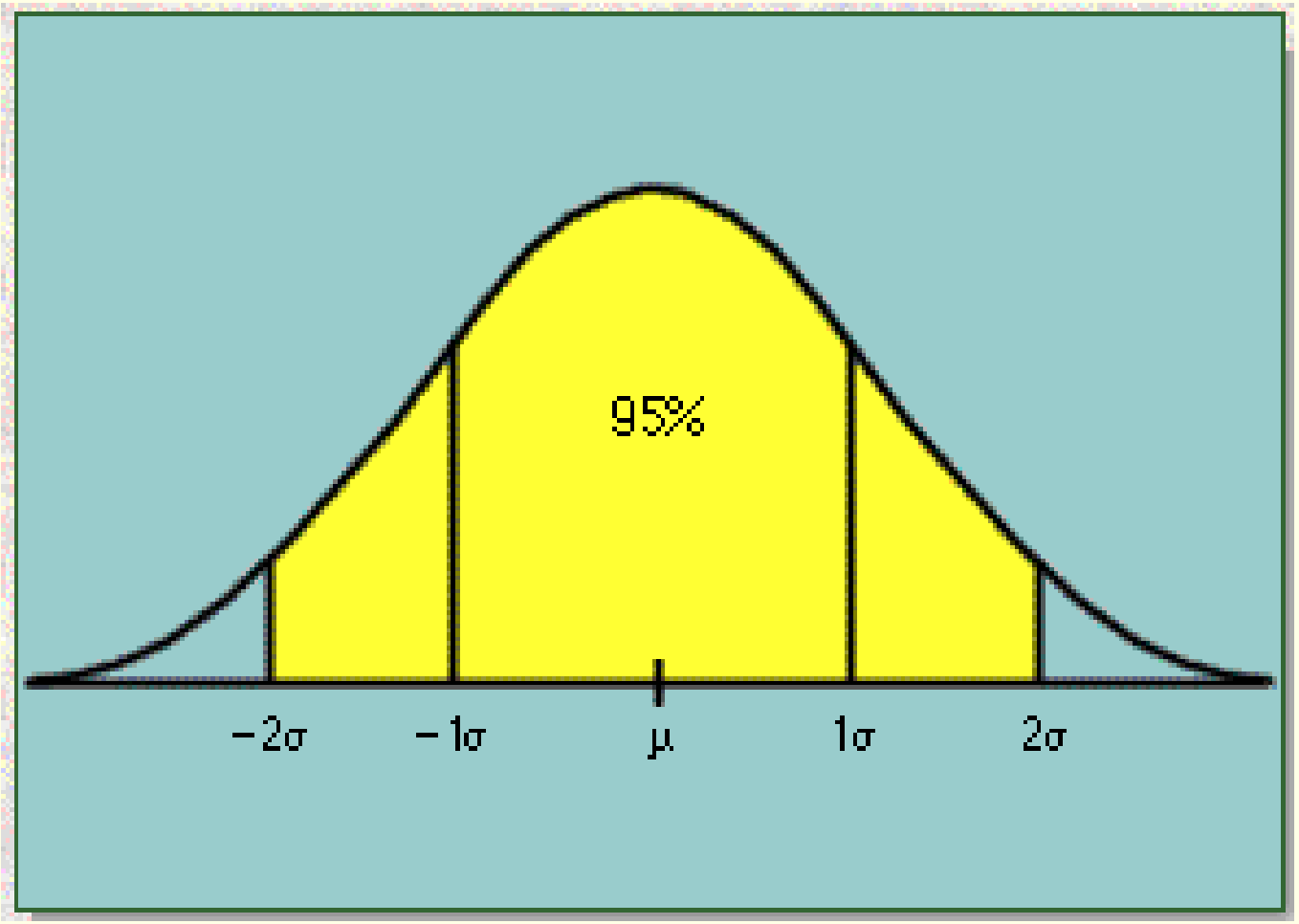
és nyilván

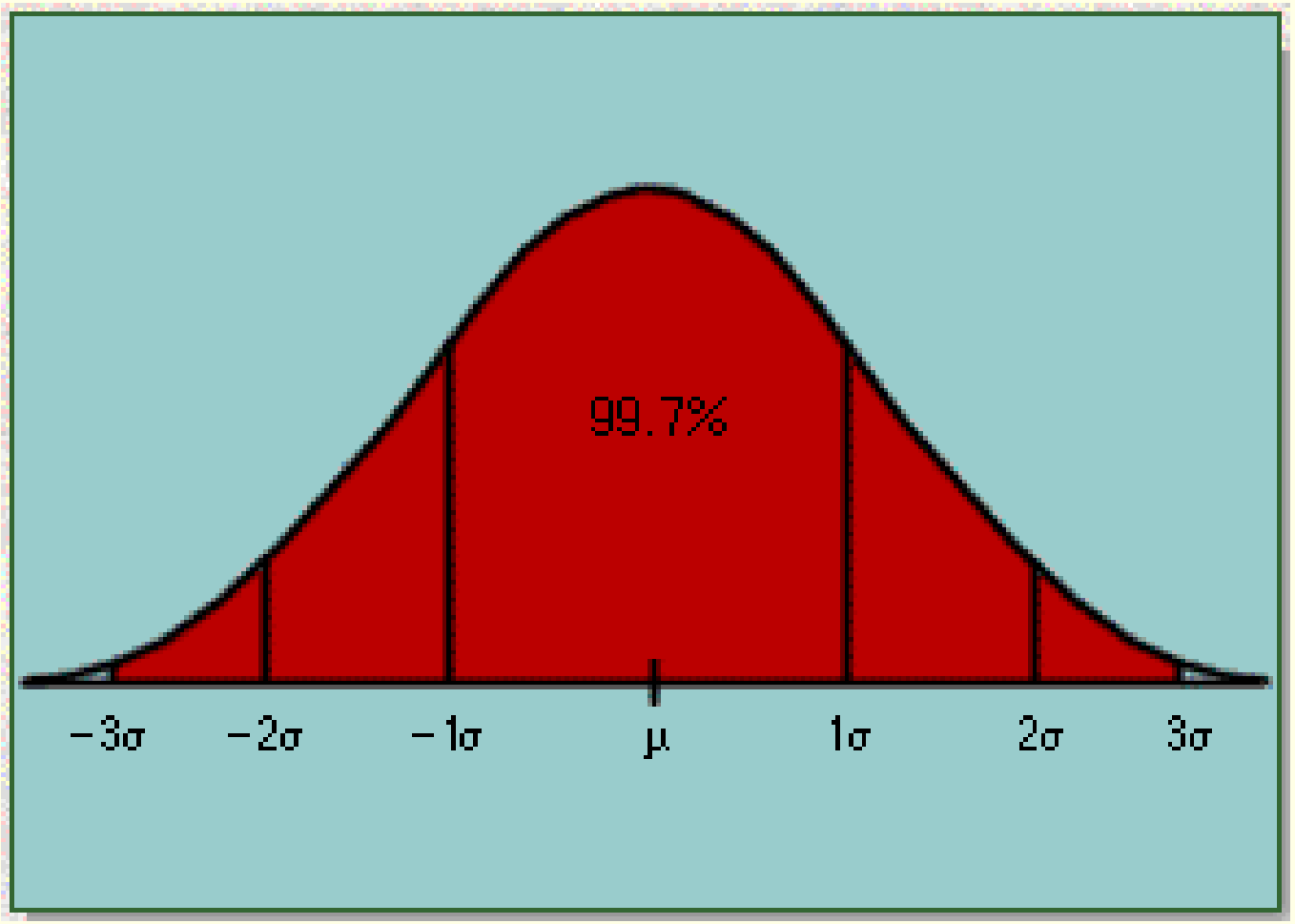
$$s = \sqrt{\frac{SQ}{n-1}}.$$

Ha a mintából készített hisztogram elég jól közelíti a normális görbét, akkor a normális eloszlás táblázatából kiolvasható, hogy az

- $]\bar{x} - s, \bar{x} + s[$ intervallumban van adataink kb. **68%**-a,
- $]\bar{x} - 2s, \bar{x} + 2s[$ intervallumban van kb. **95%**-a,
- $]\bar{x} - 3s, \bar{x} + 3s[$ intervallumba pedig kb. **99%**-a esik.


68%
– 1σ
μ
1σ


95%
−2σ
−1σ
μ
1σ
2σ


99.7%
−3σ
−2σ
−1σ
μ
1σ
2σ
3σ