Elemi statisztika fizikusoknak

Pollner Péter

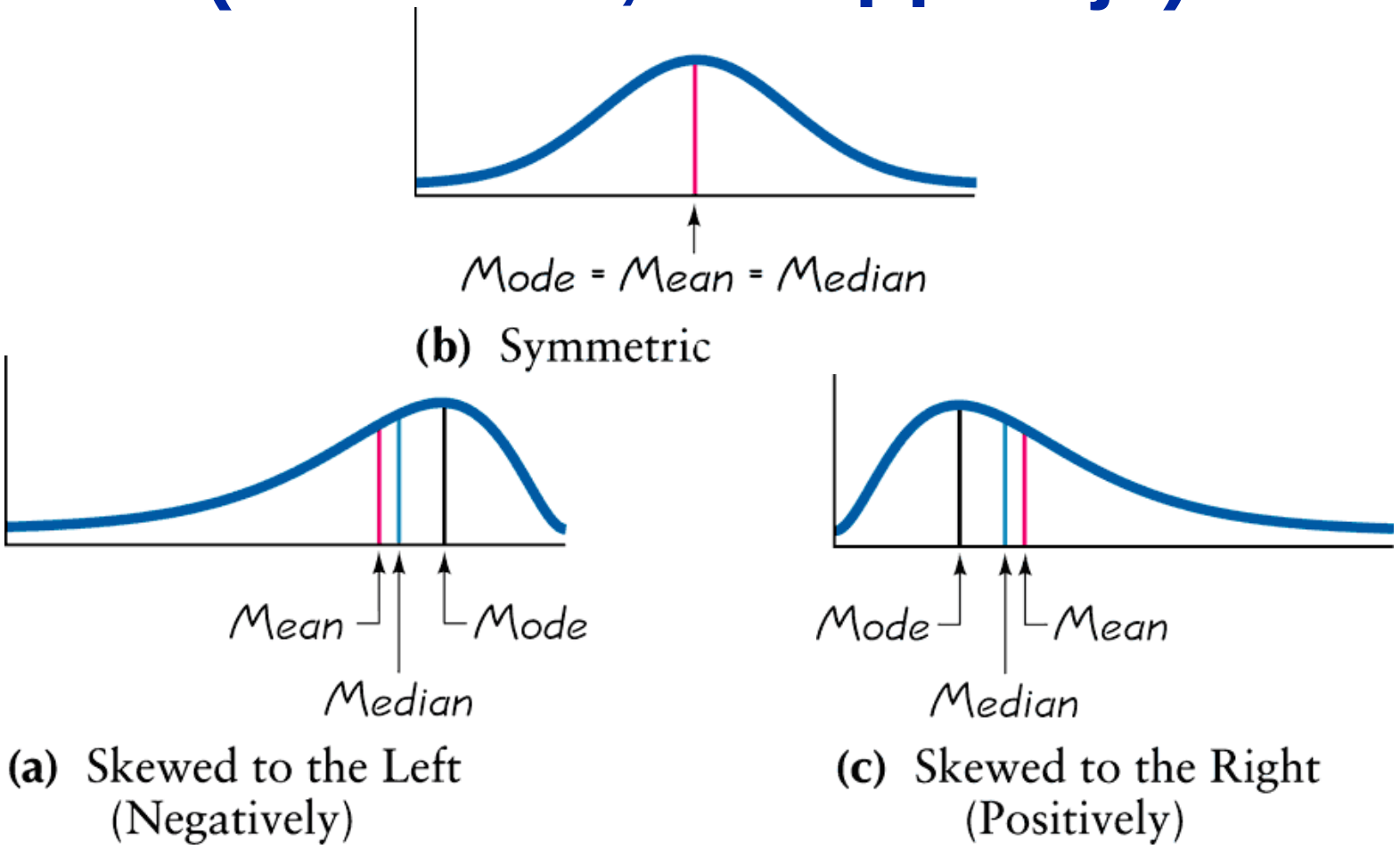
Biológiai Fizika Tanszék

pollner@elte.hu

Az adatok leírása, megismerése és összehasonlítása

Ismétlés:

az adat elhelyezkedése (centruma, középpontja)



A centrum legjobb jellemzése

4. oldal

Table 2-10 Comparison of Mean, Median, Mode, and Midrange

Measure of Center	Definition	How Common?	Existence	Takes Every Value into Account?	Affected by Extreme Values?	Advantages and Disadvantages
Mean	$\bar{x} = \frac{\Sigma x}{n}$	most familiar "average"	always exists	yes	yes	used throughout this book; works well with many statistical methods
Median	middle value	commonly used	always exists	no	no	often a good choice if there are some extreme values
Mode	most frequent data value	sometimes used	might not exist; may be more than one mode	no	no	appropriate for data at the nominal level
Midrange	$\frac{\text{high} + \text{low}}{2}$	rarely used	always exists	no	yes	very sensitive to extreme values

General comments:

- For a data collection that is approximately symmetric with one mode, the mean, median, mode, and midrange tend to be about the same.
- For a data collection that is obviously asymmetric, it would be good to report both the mean and median.
- The mean is relatively *reliable.* That is, when samples are drawn from the same population, the sample means tend to be more consistent than the other measures of center (consistent in the sense that the means of samples drawn from the same population don't vary as much as the other measures of center).

Kritikus szemlélet

Átlag számításnál:

Minek az átlagát számoljuk? Mi az alapsokaság?

csoportok átlagos mérete --- tagok által érzett átlagos méret

Minták átlagát számoljuk vagy csoportok átlagait?

Csoportosított adatokat mindig súlyozva átlagoljuk!

Melyik középérték-mutatót használjuk?

átlagot – csonkított átlagot – mediánt?

Pl. mekkora az átlagfizetés?

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)

Adatokkal a képlet

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

2-5. képlet

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

2-4. képlet

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

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

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

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

2-4. képlet

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 $\begin{cases} s^2 \\ \sigma^2 \end{cases}$

Minta variancia

Populáció variancia

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

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

2-4. képlet

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

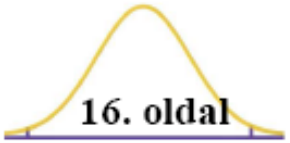
3,3 $\overline{x} = 3$ $s^2 = ((3-3)^2 + (3-3)^2)/1 = 0.0$

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?



$$\text{Átlag abszolút eltérés} = \frac{\sum |x - \bar{x}|}{n}$$

nem additív és nem torzítatlan becslése a populáció átlag abszolút eltérésének

Definíció

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

Minta

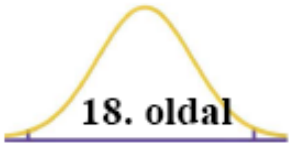
$$CV = \frac{s}{\bar{x}} \bullet 100\%$$

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:
- Magasság átlaga $\overline{x}$ = 173.58 cm
- Magasság szórása S= 7.67 cm
- Súly:
- Súly átlaga $\overline{x}$ = 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 magasság sokkal kevésbé változékony mint a súly!

Definíció

Csebisev tétel

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

$$\#\{|x-\mu|<K\sigma\} > N(1-1/K^2)$$

- ***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**

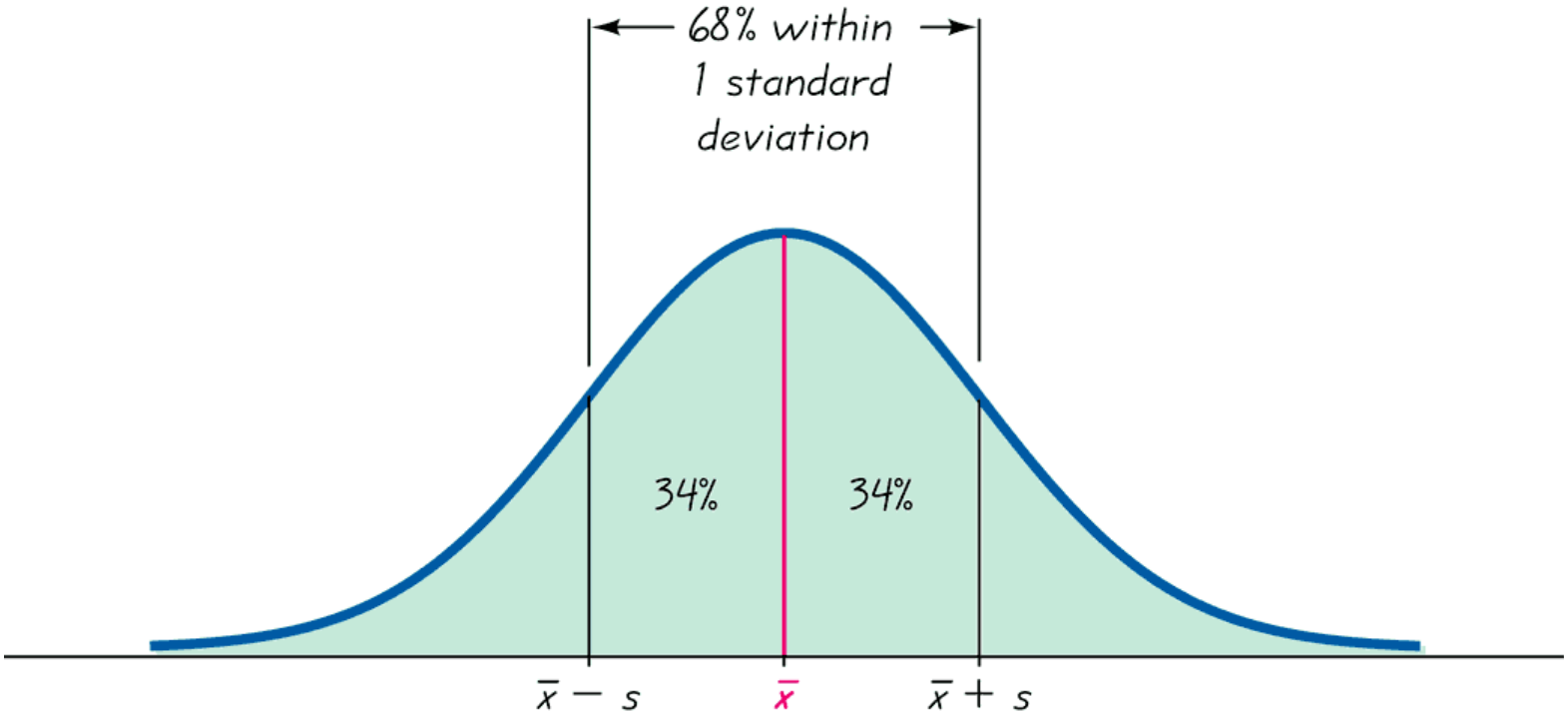
Definíció

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**

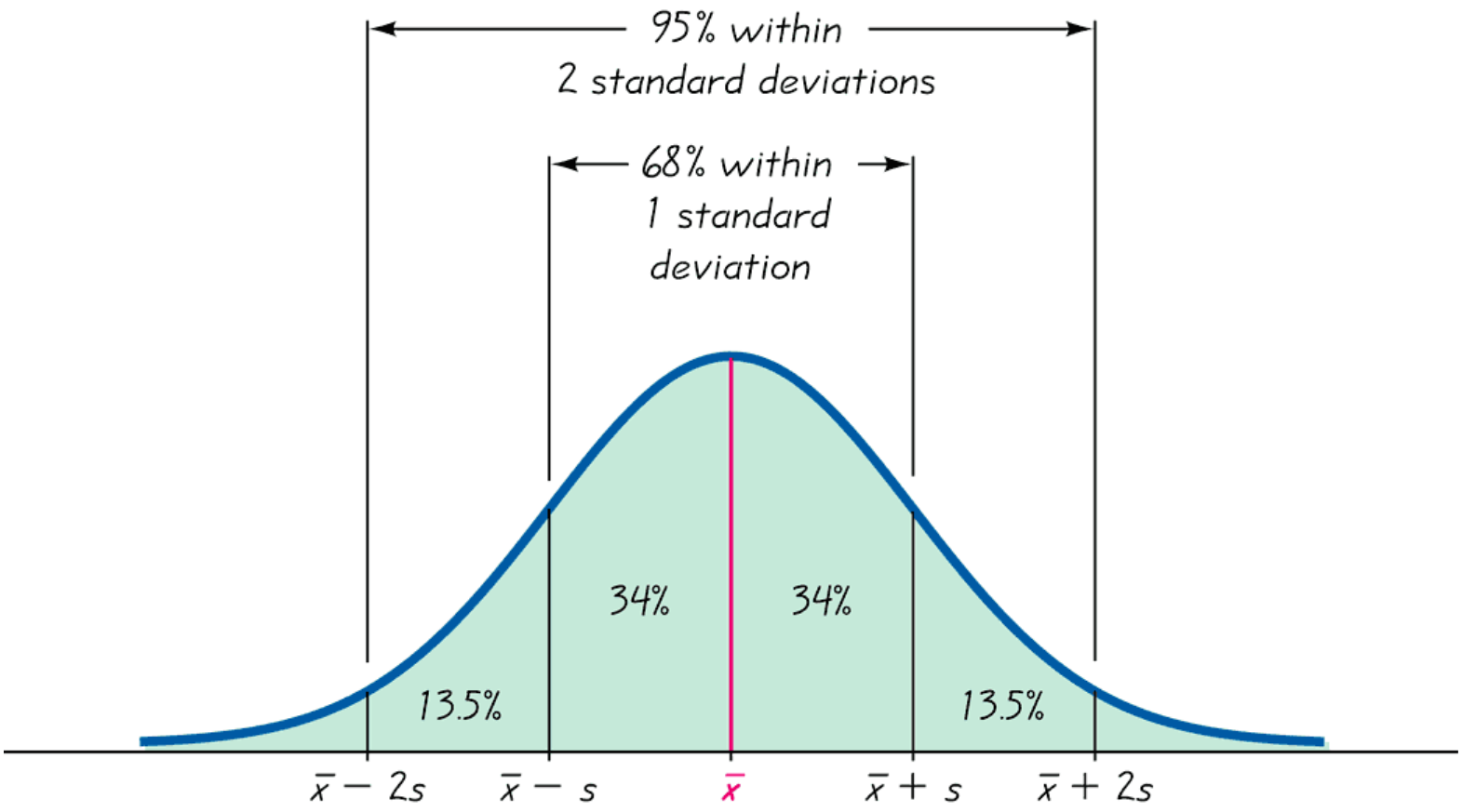
Az empirikus szabály



2-13. ábra

Az empirikus szabály

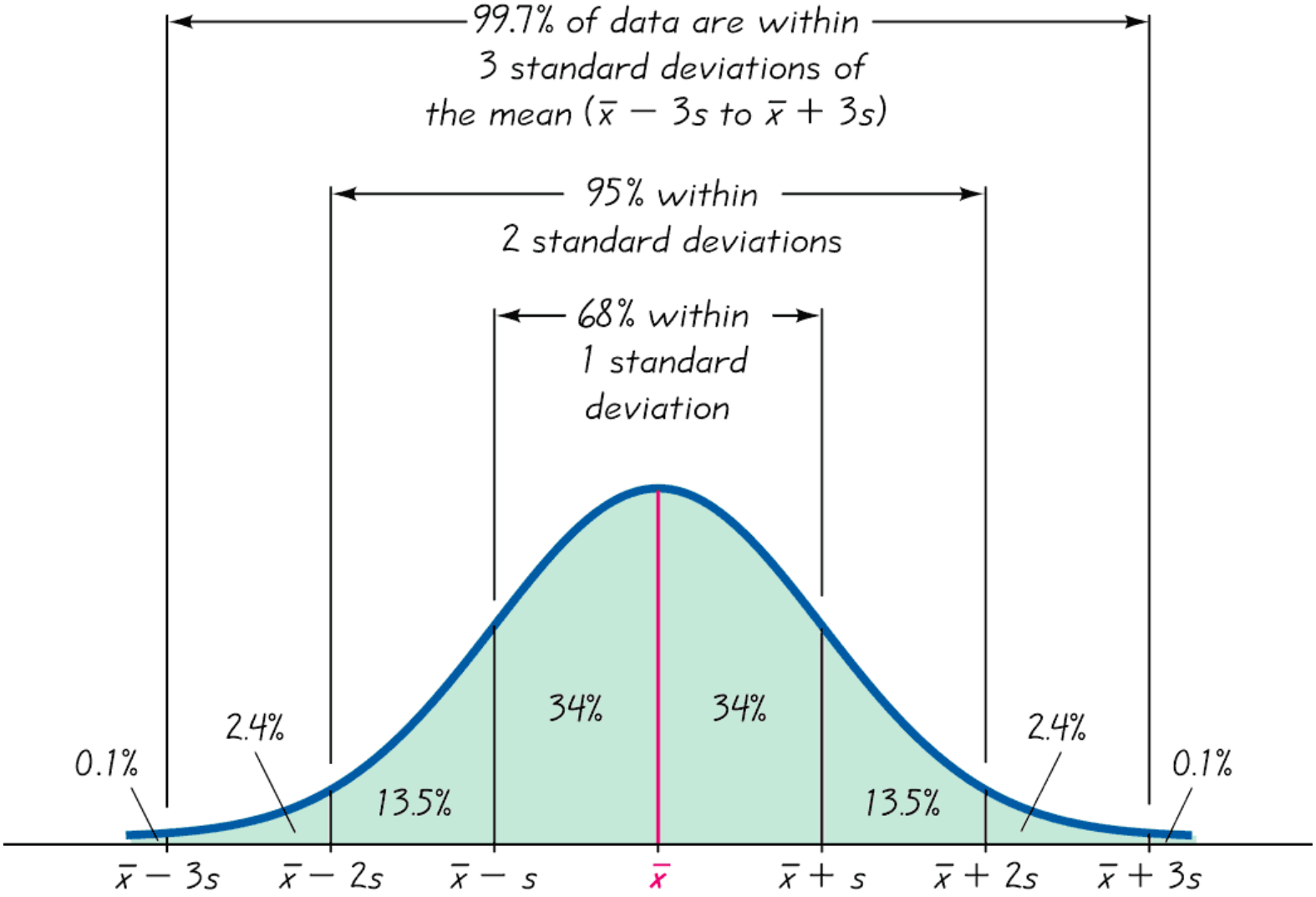
25. oldal



2-13. ábra

Az empirikus szabály

26. oldal



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) (z-score)**

  ***x* pozitív vagy negatív eltérése az átlagtól szórás egységekben mérve.**

Az eltérés mérése
z érték

Minta

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

Populáció

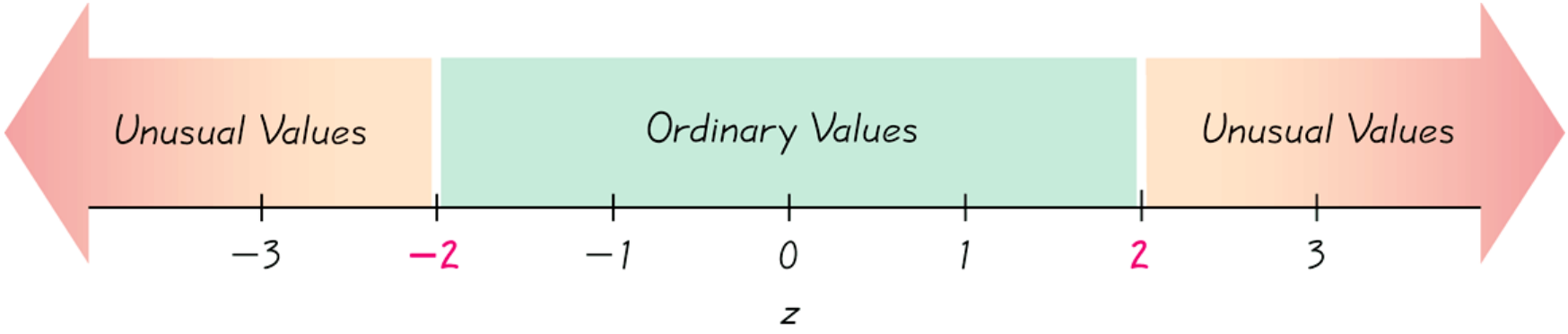
$$z = \frac{x - \mu}{\sigma}$$

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.**
- **$z=(190.5-181.6)/5.3=1.67$**
- **$z=(216-203.2)/8.4=1.52$**

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

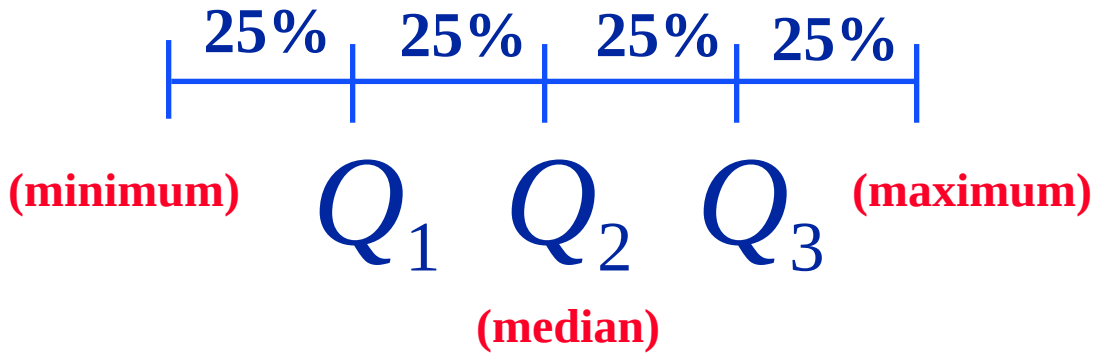
(szokatlan előfordulása: mintaméret-függő)

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).

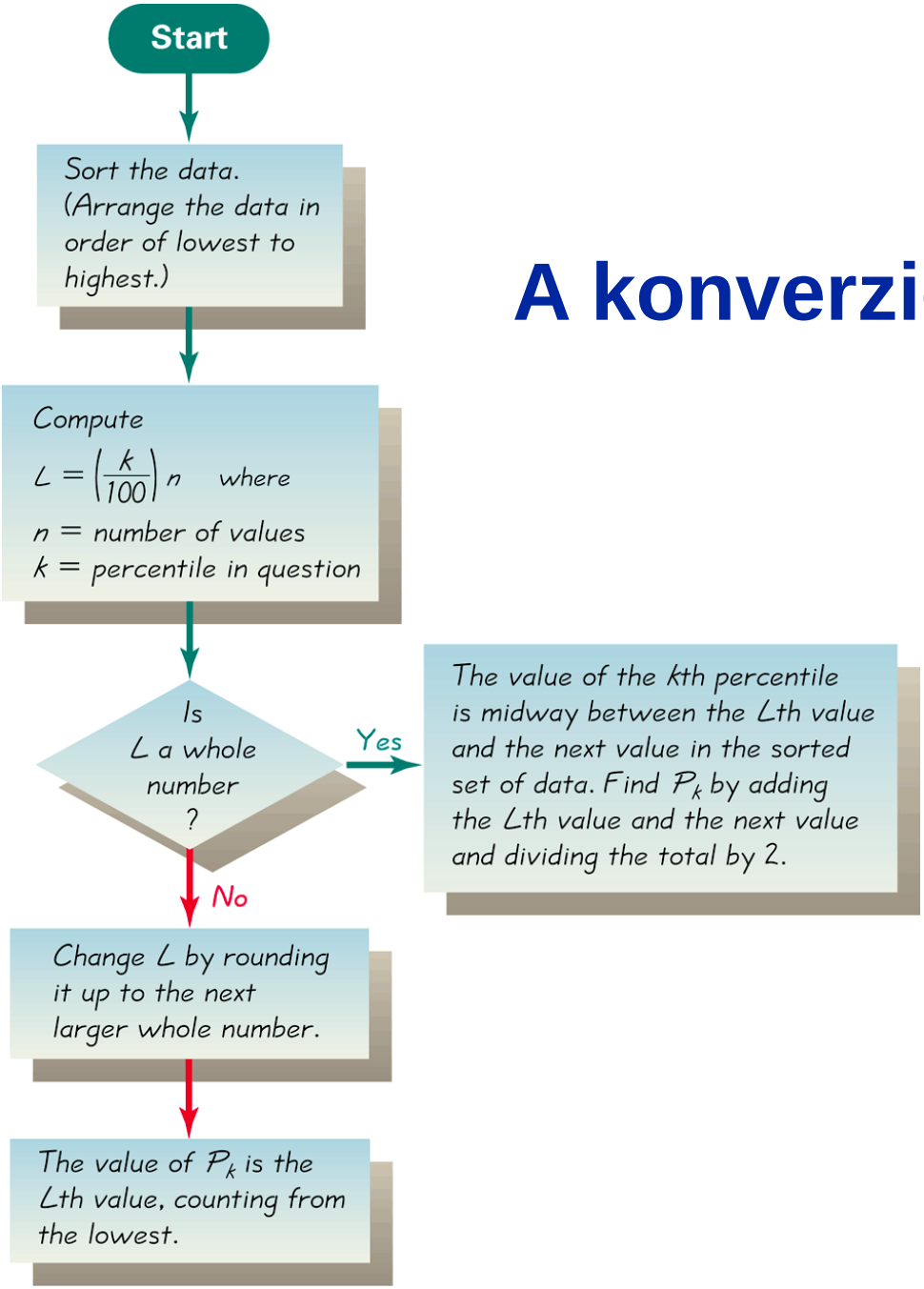
$$L = \frac{31}{100} \bullet 36 = 11.16$$

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 statisztikai 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

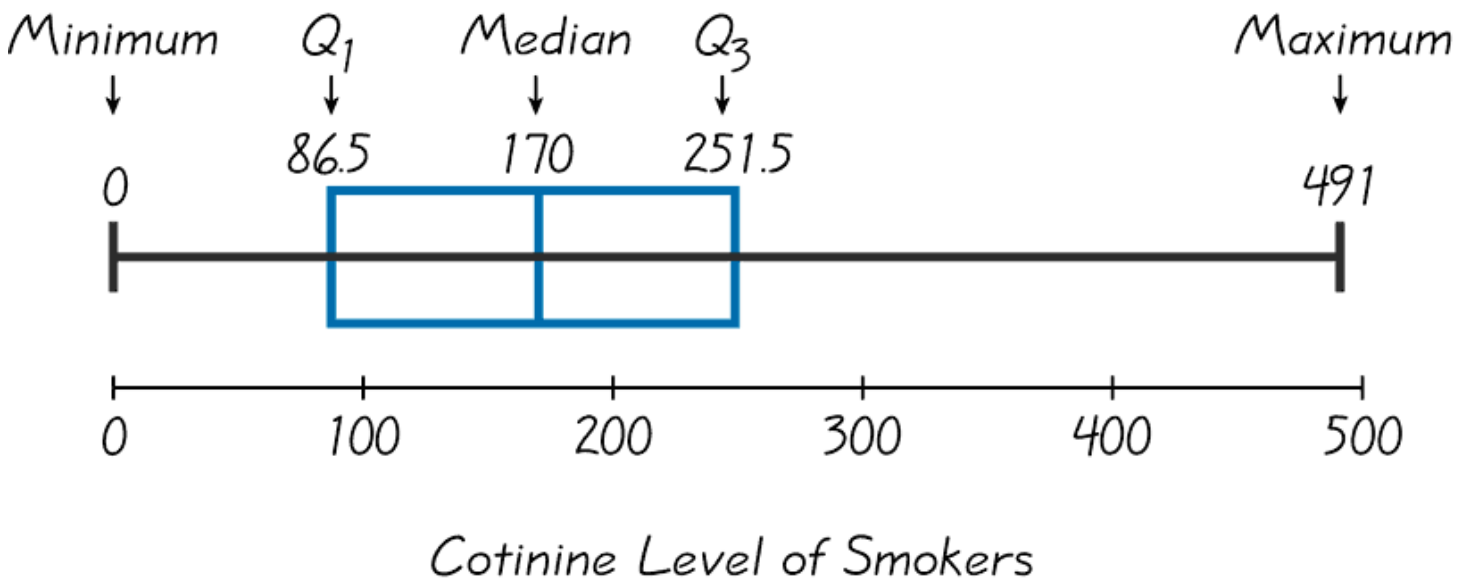
- **Egy outlier-nek drámai hatása lehet az átlagra**
- **Egy outlier-nek drámai hatása lehet a szórásra**
- **Egy outlier-nek drámai hatása lehet a hisztogramra, ami miatt az eloszlás teljesen zavaros lesz**

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.

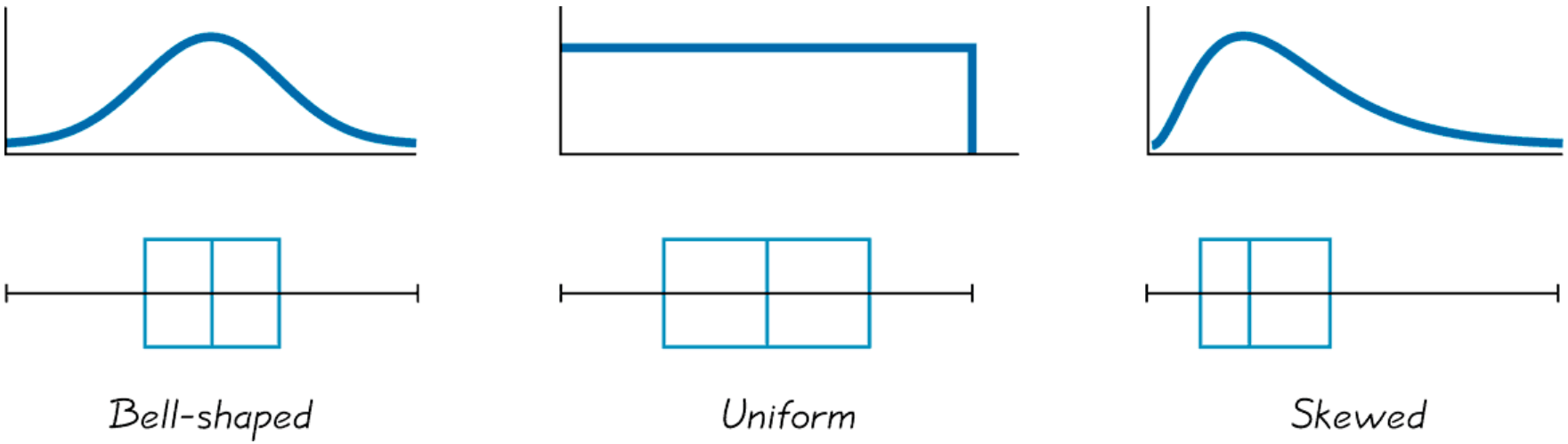
Boxplot



2-16. ábra

Boxplot-ok

49. oldal



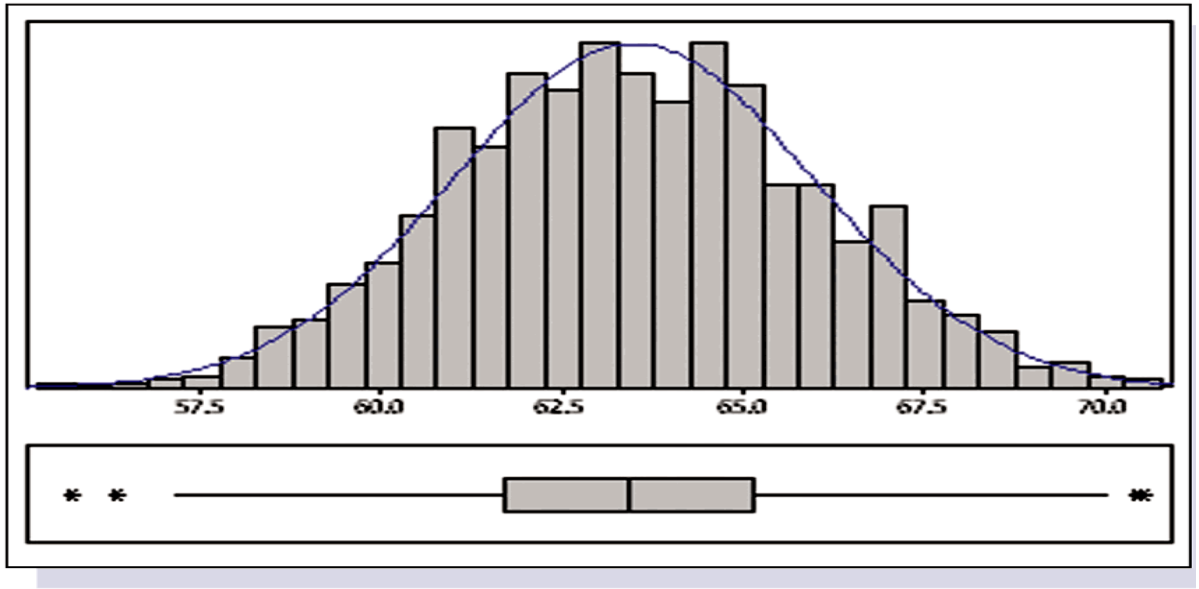
2-17. ábra

Módosított boxplot

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

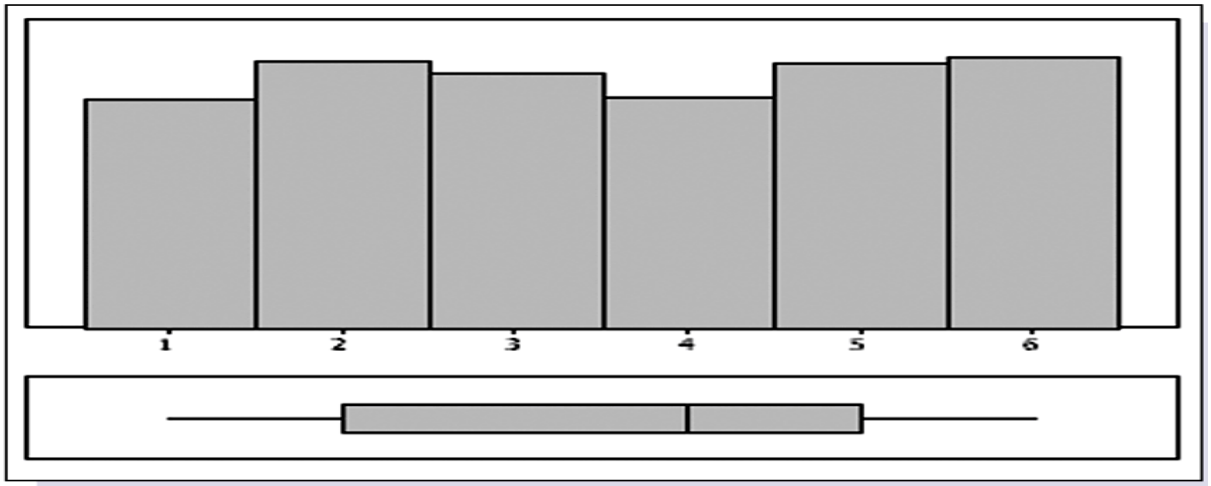
❖ **Interkvartilis terjedelem (IQR):** $Q_3 - Q_1$

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

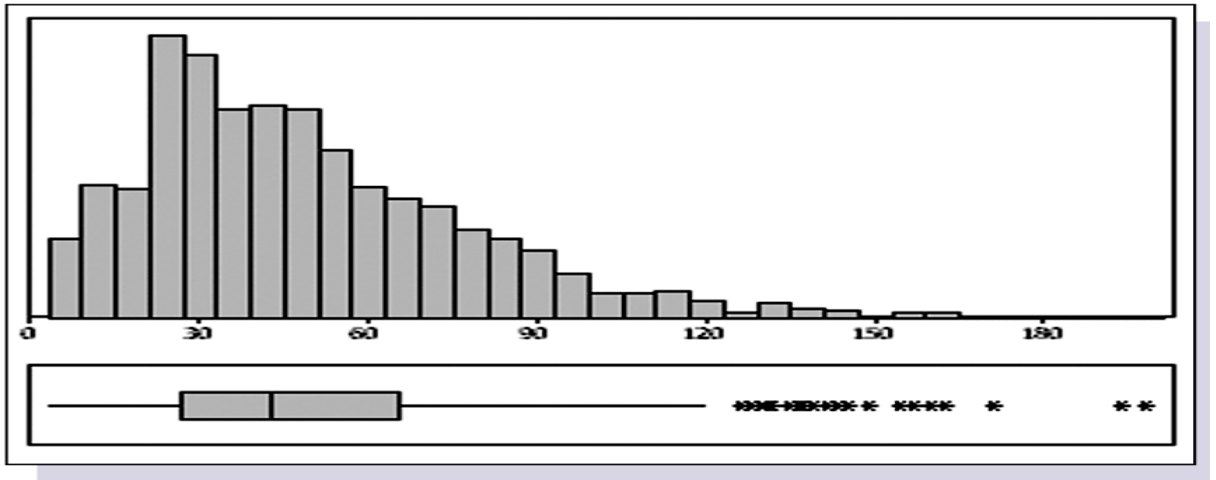


(a) Normal (bell-shaped) distribution
1000 heights (in.) of women

Módosított boxplot

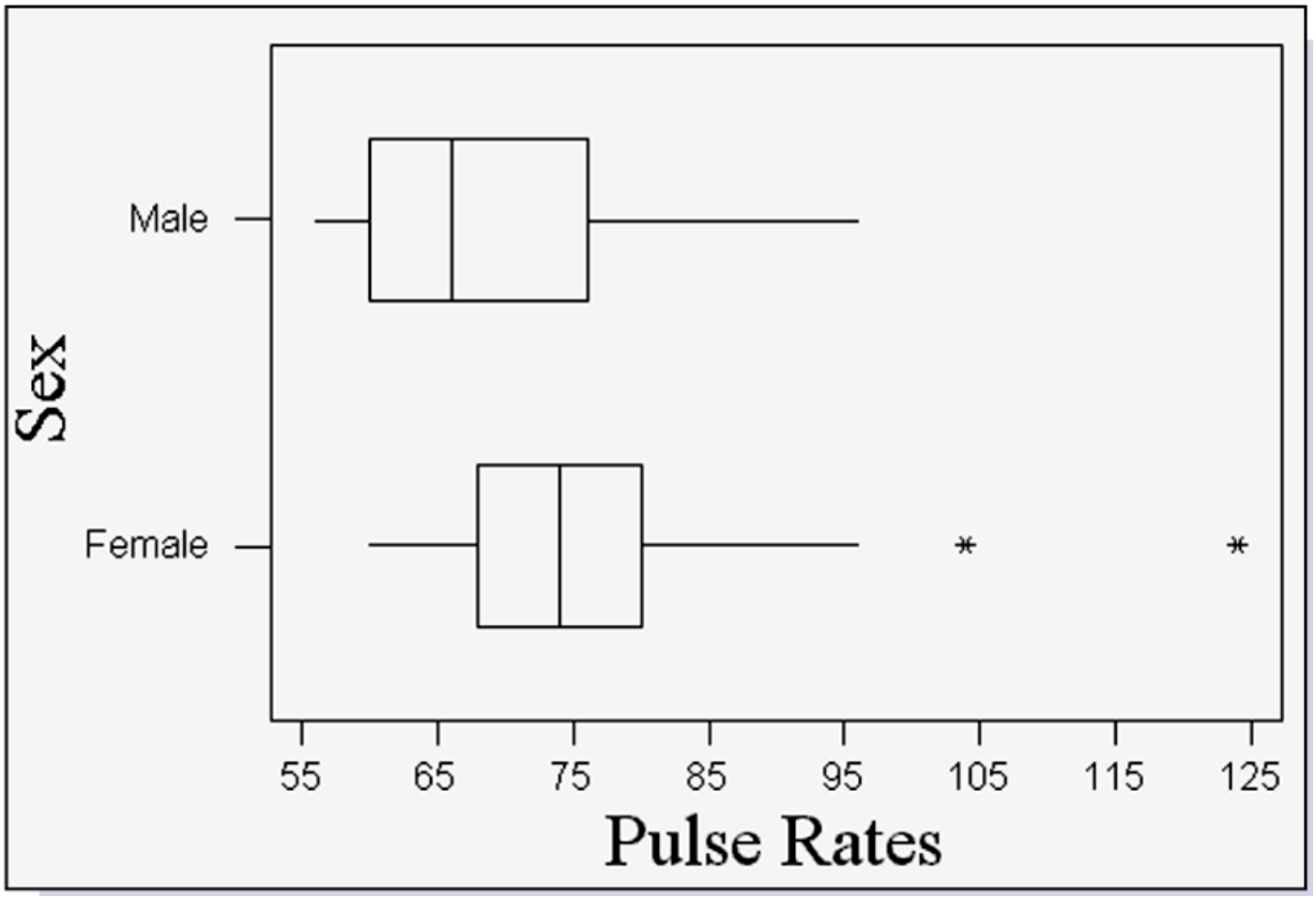


(b) Uniform distribution
1000 rolls of a die



(c) Skewed distribution
Incomes (thousands of dollars) of 1000 statistics professors

Módosított boxplot



Összefoglalás

Ebben a fejezetben áttekintettük:

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