Adatok rendezése és "első" elemzése

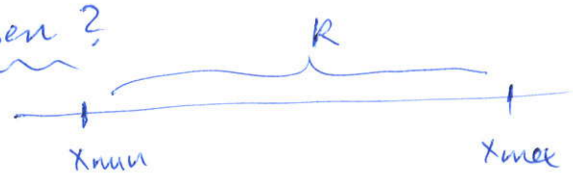
(Az adatgyűjtés fajtáival, lehetőségeivel és a minta elemek számával, ismételtségével stb. nem foglalkozunk.)

Kiinduló probléma: adott egy adatsor (minta) és ezekről szeretnénk minél többet megtudni.

$x_1, x_2, \dots, x_n$ n elemű adatsor (A) eset: mérés kül. érték van

1. lépés: Hogyan helyezkednek el a számegyenesen?

x_{min}, x_{max}, $R := x_{max} - x_{min}$ (adatsávszélesség) meghat.-a



2. lépés: rendezett minta meghat.-a: $x_1^* \le x_2^* \le \dots \le x_n^*$

3. lépés: $\boxed{\text{1. táblázat}}$ elkészítése:

i	x_i^*	H_i (abs. gyak.)	$h_i = \frac{H_i}{n}$ (rel. gyak.)	$\sum_{j=1}^{i} h_j$ (összegzett rel. gyak.)
1				
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
k				
		$\sum_{i=1}^{k} H_i = n$	$\sum_{i=1}^{k} h_i = 1$	

(k ~ a kül. mintaelemek száma)

4. lépés: a kvantilisek meghat.-a: $x_{0,25}$ (alsó kvartilis), $x_{0,5} = Me$ (medián), $x_{0,75}$ (felső kvartilis)

$$x_p := x^*_{p\cdot(n+1)} \quad \text{ha } p\cdot(n+1) \text{ egész szám} \qquad (0 < p < 1) \quad p\text{-kvantilis}$$

$$x_p := \frac{1}{2}\left[x^*_{[p(n+1)]} + x^*_{[p(n+1)]+1}\right] \quad \text{ha } p\cdot(n+1) \text{ nem egész}$$

D = módusz meghat.-a: (lehet több is) D ~ a legnagyobb gyakoriságú adat

(öt-számos összegzés: $x_{min}, x_{0,25}, Me, x_{0,75}, D$ megadása)

Már van fogalmunk arról, hogy az $[x_{min}, x_{max}]$ int.-ban hogyan helyezkednek el a mintaelemek.

5. lépés: empirikus sűrűség- és eloszlás fv.-k felrajzolása az 1. tábl. segítségével (hisztogram)

További jellemzők meghatározása

középérték (átlag, számtani közép)

$$\boxed{\overline{X} := \frac{\sum_{i=1}^{n} x_i}{n}}$$

(az 1. tábl. segítségével)

A szemináriumra javasolt feladatok típusai

(1.Pl.) x_{min}, x_{max}, R, D, $x_{0,25}$, Me, $x_{0,75}$ = ?
1. tábl.-t adja meg, $\overline{X}$ = ? (rajzolja be ezeket a számegyenesen!)
rajzolja fel az empirikus sűrűség és eloszlás fv.-k képét!

ie
a) 1, 3, 3, 3, 3, 4, 2, 3, 4, 4, 3, 4, 2, 3, 1, 4, 1, 3, 2, 2, 3, 2, 2, 5, 3, 1, 2, 4, 2, 2, 4, 1, 4, 1, 2, 2
3, 3, 2, 4, 3, 3, 4, 5, 1, 1, 5, 4, 2, 5 (stat. vizsgajegyek tavaly)

b) 4, 1, 4, 4, 1, 2, 2, 3, 5, 1, 4, 5, 5, 4, 5, 2, 4, 2, 1, 3, 2, 2, 2, - 5, 5, 1, 1, 4, 3, 2, 4, 1, 5,
2, 3, 4, 4, 1, 2, 2, 3, 4, 4, 1, - 4, 2, 3, 5, 1, 4, 2, 2, 3, 5, 2, 3 (hol van az adatok fele?)

c) egy röplabda-csapat tagjai magassága (cm)
183, 181, 183, 180, 182, 182, 185, 182, 184, 179, 182, 184, 180, 181, 179, 180,
182, 180, 181, 183

d) egy telefonfülke kihasználtsága (naponta hányan vették igénybe)
1, 1, 2, 0, 6, 4, 3, 1, 0, 0, 3, 4, 3, 2, 2, 1, 1, 2, 3, 3, 0, 0, 0, 4, 3, 2, 1, 2, 6, 3, 3, 3,
0, 4, 1, 2, 2, 6, 2, 2
(igaz-e, h. min 4 látogató van 60%-ban ?)

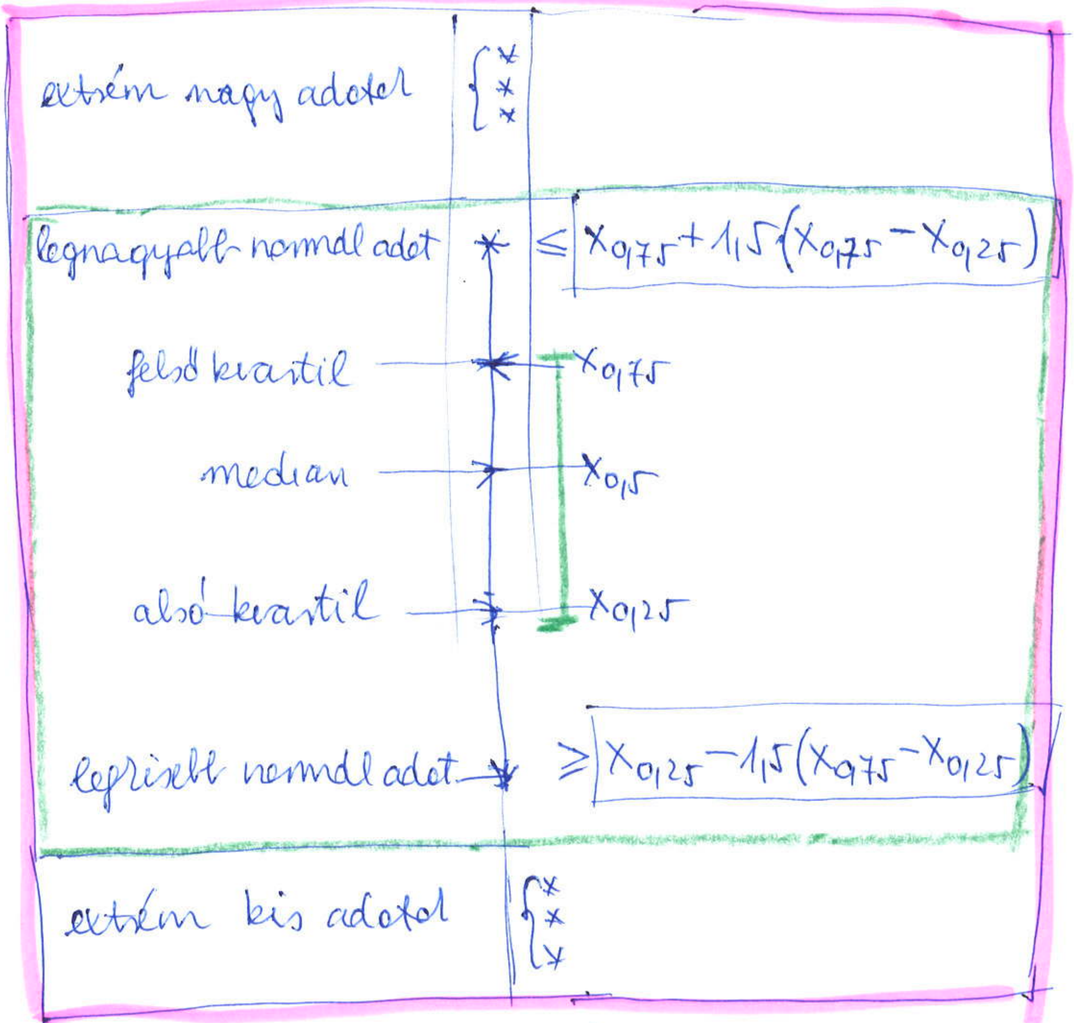
e) egy kis élelmiszerboltban naponta eladott friss literes tejek db számai
3, 4, 0, 0, 2, 3, 3, 1, 2, 2, 2, 5, 4, 5, 7, 7, 8, 8, 3, 3
6, 7, 3, 2, 1, 7, 5, 5, 4, 1, 3, 8, 7, 5, 5, 4, 3, 3, 4, 4
3, 3, 3, 2, 5, 5, 4, 3, 3, 3, 3, 4, 8, 7, 7, 5, 4, 1, 1, 3
0, 0, 1, 0, 0, 4, 3, 2, 1, 2, 1, 4, 0, 0, 1, 1, 1, 1, 5, 0, 1
0, 2, 2, 3, 1, 3, 6, 6, 7, 6, 3, 4, 3, 2, 3, 1, 5, 3, 3, 2
(akkor éri meg a friss tej árusítása, ha a napok legalább 70%-ban min
2 db-t el tud adni. Megéri-e ?)

2.

(A) eset, (sok adat, de ebből kevés a kül. érték (≤ 10))

Box-Plot ábra (kiugró vagy hibás adatok kiszűrése, kiemelése)

Mire jó még: a minta vizuális megjelenítése
kül. minták vizuális összehasonlítása



elnev.:
kvartilis terjedelem:
$[X_{0,75}-X_{0,25}] =: IQT$

(lehet vízszintesen is ábrázolni!)

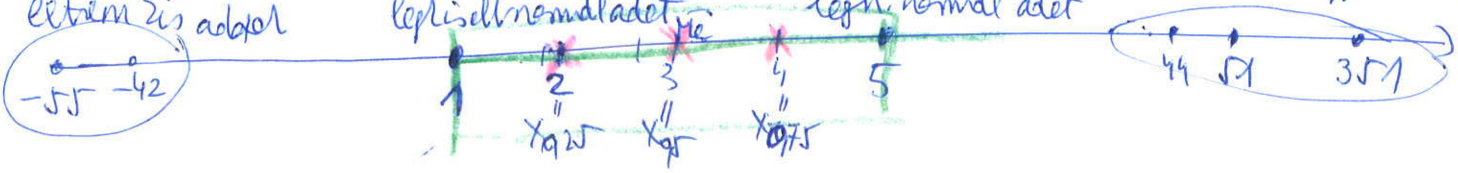
1. Pl. /b (1. szem.) készítsük el a Box-Plot ábrát!

legnagyobb normál adat $\le x_{0,75}+1,5(x_{0,75}-x_{0,25}) = 4+1,5\cdot(\overbrace{4-2}^{2}) = 4+3 = 7$

a legn. normál adat, amely ≤ 7 az 5

legkisebb normál adat: $\ge 2-1,5\cdot 2 = 2-3 = -1$

a legkisebb normál adat, amely ≥ -1 az 1



(B) eset: sok kül. érték, x_{min}, x_{max} megállapítása

az adatokat **K db osztályba soroljuk** ($K \le 5\lg n$ vagy $6 \le K \le 20$, $K \le \sqrt{n}$)

(db. szabály ∄ az osztályok db számára)

Az osztályok szélességei lehetnek $\left<\begin{array}{l}\boxed{\text{egyenlőek}}\\ \text{különbözőek}\end{array}\right)$ (feladattól függ)

Ifh.: az osztályok egyenlő hosszúak

1. lépés osztályok meghat.-a, K megadása

$a \le x_{min}$, $b \ge x_{max}$, a, b értékeit úgy választjuk meg, hogy $\left(\frac{b-a}{K}\right)$

"egyszerű" érték legyen

(K=4 esetén)

az osztályok osztópontjai: $a_0 := a \le x_{min}$

$a_K := b \ge x_{max}$

$[a_{i-1}, a_i)$ ~ i. osztály
$i = 1, \dots, K$

Minden adatot, amely az i. osztályba esik az ú.n. osztályközéppel: u_i értékkel helyettesítünk. Így visszavezettük a (B) esetet az (A) esetre.

Tehát a továbbiakban az u_i osztályközepekkel számolunk.

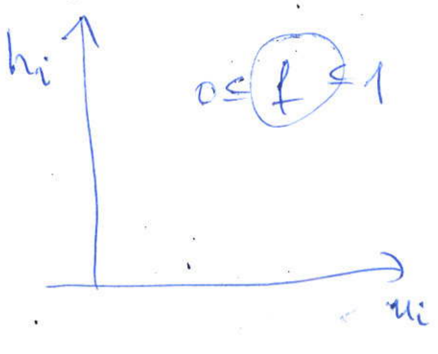
(kevés kül. érték) Elkészítjük az osztályközepektől a rendezett mintától.

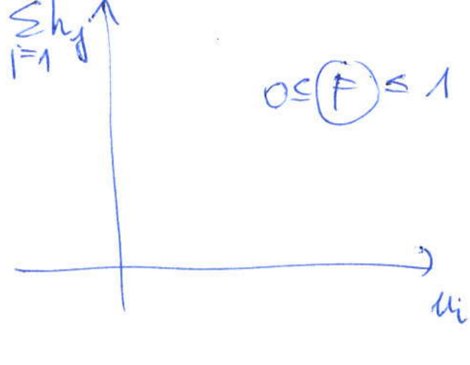
Az **1. táblázat** ebben az esetben:

i	osztályok	osztályközép	gyakoriság (H_i)	rel. gyak. (h_i)	összegzett rel. gyak. ($\sum_{j=1}^{i} h_j$)
	$[a_i, a_{i-1})$	u_i			

A táblázatból meghatározható: alsó-, felső kvartilis, medián, módusz, átlag értékeit az osztályközepekből számoljuk

A tapasztalati sűrűség- és eloszlásfv. képeit is az osztályközepekkel készítjük el.





3. ea. Középértékek és az ingadozás mérőszámai

Középértékek (helyzetmutatók)

átlagok | **helyzeti középértékek**: Me, D (extrém adatokra érzéketlenek)

átlagok	
1. $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ számtani közép	
2. $\bar{x}_g$ ~ geometriai közép; $\bar{x}_g = \sqrt[n]{x_1 x_2 \cdots x_n}$	érzékeny a nagyon kis értékekre; akkor használjuk, ha az értékek szorzata értelmezhető ill. ha az átlagolandó értékek exponenciálisan nőnek vagy csökkennek (pl. relatív változások átlagolásakor)
3. $\bar{x}_h$ ~ harmonikus közép; $\bar{x}_h := \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$	érzékeny a nagyon kis értékekre; akkor használjuk, ha az értékek reciprokainak összege értelmezhető (főleg indexszámítás esetén)
4. $\bar{x}_q$ ~ négyzetes közép; $\bar{x}_q := \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2}$	érzékeny a kiugróan magas értékekre; akkor használjuk, ha lehetnek poz. és neg. értékek is, de az értékek abszolút nagysága érdekes

Egy adott mintánál az átlagok közötti nagyságrendi reláció:

$$x_{min} < \bar{x}_h < \bar{x}_g < \bar{x} < \bar{x}_q < x_{max}$$

(feltéve, hogy nem $\forall$ adat azonos)

1. Pl. (ea-n) Egy kerékpárgyártó cég a forgalmát évről-évre növelni tudta az alábbiak szerint: (%-ban mindig az előző évhez viszonyítva)

1990→1991 107%-ra
1991→1992 102%-ra
1992→1993 108%-ra
1993→1994 111%-ra
1994→1995 105%-ra

Mekkora az átlagos évi növekedés és az átlagos évi növekedési ütem?

2. Egy kereszteződésben mérik a sebességet, az alábbi eredményeket kapjuk: $N = 64$

x_i	55	65	75	85	95	105	115	125
H_i	3	6	15	21	12	4	2	1
h_i	0,047	0,094	0,234	0,328	0,188	0,062	0,031	0,016

$\bar{x}, \bar{x}_g, \bar{x}_h, \bar{x}_q = ?$

Me, D = ? ábrázoljál

(ea-n)

3. Egy hallgatói csop. matematika dolgozat jegyei

x_i	2	3	4	5
a) H_i	12	28	21	3
b) H_i	6	14	10	2

$\bar{x}, \bar{x}_g, \bar{x}_h, \bar{x}_q = ?$

$Me, D = ?$

Ingadozás mérőszámai (szóródás mérőszámai) szóródási mutatók

1. terjedelem: $\boxed{R := x_{max} - x_{min}}$ (adatsávszélesség)
2. kvartilis távolság = interkvartilis terjedelem: $\boxed{x_{0,75} - x_{0,25} = IQT}$
3. átlagos abszolút eltérés: $\boxed{\Delta := \frac{1}{n}\sum_{i=1}^{n} |x_i - \bar{x}|}$ (középeltérés)
4. empirikus szórás: $\boxed{\sigma := \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}}$ $\left(= \sqrt{Var(X)}\right)$

$Var \overset{jel}{=} \sigma^2 := \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ empirikus variancia)

5. korrigált empirikus szórás: $\boxed{\sigma^* := \sqrt{\frac{n}{n-1}} \cdot \sigma}$ (kisebb minta esetén)

$((\sigma^*)^2 \sim$ korrigált empirikus variancia)

6. relatív szórás: $\boxed{v := \frac{\sigma}{\bar{x}}}$ (vagy %-ban: $v = \frac{\sigma}{\bar{x}} \cdot 100\%$)

// (tapasztalati variancia együttható)

Megj.: – a szórás érzékeny a kiugró elemekre
– Δ kevésbé érzékeny a -||- -||-
– $\boxed{\sigma^2 = \left(\frac{1}{n}\sum_{i=1}^{n} x_i^2\right) - \bar{x}^2} = \bar{x}_q^2 - \bar{x}^2$

R	
IQT	
Δ	
σ_n	
σ_n^*	
v	

Az ingadozás mérőszámait a $\boxed{\text{2. táblázat}}$ segítségével könnyű kiszámolni

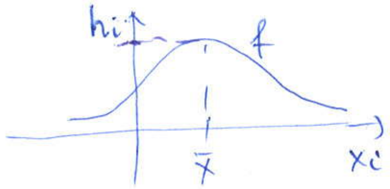
$x_i^* H_i$	H_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})^2 \cdot H_i$
	$\Sigma = n$	Δ		$\sum_{i=1}^{k}(x_i - \bar{x})^2 \cdot H_i = n \cdot \sigma^2$

ebből számoljuk σ-t és σ^*-t, és v-t

$\bar{x} = \frac{1}{n}\sum x_i$

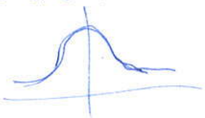
4. ea. Alakmutatószámok: asszimetria / lapultság jellemző mutatói

Probléma: Adott adatsor/minta esetén elkészíthető az empirikus sűrűségfv. képe.
Az lenne a "kívánatos", h. az $\bar{x}$ körül szimm. legyen, azaz "normális".
Az u.n. normális eloszlás sűrűségfv.-e: a Gauss-féle harranggörbe (egymóduszú esetén).
Ehhez hasonlítjuk az adatokból kapott empirikus sűrűségfv.-t.

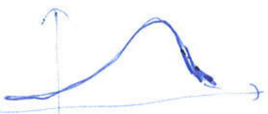


I. Milyen lehet az empirikus sűrűségfv. szimmetria szempontjából?

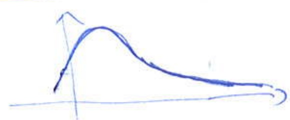
- egymóduszú
  - szimmetrikus



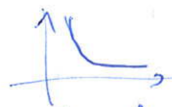
  - aszimmetrikus
    - mérsékelten
      - balra elhúzódó



      - jobbra elhúzódó

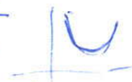


    - erősen
      - J-alakú

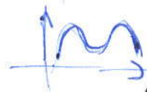


      - fordított J alakú

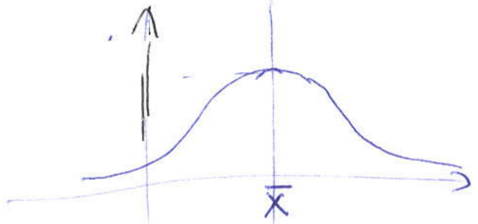
- (több móduszú
  - U-alakú



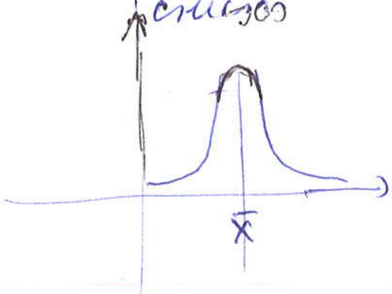
  - M-alakú)

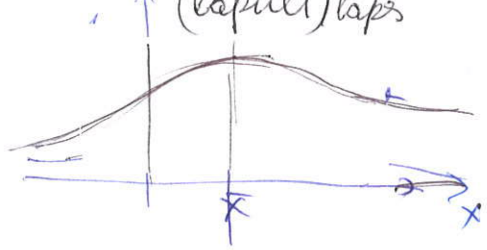


II. Milyen lehet az empirikus sűrűségfv. lapultság/csúcsosság szempontjából?



"normális"





A fenti esetek leírásában segítenek az u.n. alakmutató számok

I. aszimmetria-mérőszámok egymóduszú sűrűségfv. esetén

a, Ha f szimmetrikus $\Rightarrow \bar{x} = D = Me$

b, Milyen a jó mérőszám? < dimenzió nélküli / szimmetrikus esetben 0

1, $$\tilde{F} := \frac{(x_{0,75}-Me)-(Me-x_{0,25})}{(x_{0,75}-Me)+(Me-x_{0,25})}$$

a felső és alsó kvartilis mediántól mért távolságai különbségének és összegének a hányadosa

Áll.: $-1 \le \tilde{F} \le 1$	ha f szimm. $\Rightarrow \tilde{F} = 0$
ha $\lvert\tilde{F}\rvert > 0,3 \Rightarrow$	ha f balra elnyúló $\Rightarrow \tilde{F} < 0$
számottevő a ferdeség	ha f jobbra " $\Rightarrow \tilde{F} > 0$

2) Pearson-index: $P := \dfrac{\bar{x} - Me}{\sigma}$

Áll: ha f szimm. $\Rightarrow P = 0$

ha f balra elnyúló $\Rightarrow P < 0$

ha f jobbra — " — $\Rightarrow P > 0$

P-nek $\nexists$ korlátja, de ha $|P| > 1 \Rightarrow$ erős az asszimetria

3) ferdeség: $\gamma := \dfrac{\beta_3}{\sigma^3}$, ahol $\beta_3 := \dfrac{\left(\sum_{i=1}^{n}(x_i - \bar{x})^3\right)\frac{1}{n}}{\sqrt{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^3}}$

$\left[\beta_3 := \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3\right.$

ún. centrális harmadik momentum$\left.\right]$

(ahol $\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$)

Áll: ha f szimm. $\Rightarrow \gamma = 0$

ha f balra elnyúló $\Rightarrow \gamma < 0$

ha f jobbra — " — $\Rightarrow \gamma > 0$

II. A sűrűségfv. meredekségi mérőszáma: Excess,

$Exz := \dfrac{\beta_4}{\sigma^4} - 3$ $\left(\beta_4 := \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4\right.$ centrális 4. momentum$\left.\right)$

Áll: ha $Exz = 0 \Rightarrow$ "normális" v. közepes meredekségű

ha $Exz < 0 \Rightarrow$ lapultabb (a normálisnál)

ha $Exz > 0 \Rightarrow$ csúcsosabb

Az alakmutató számolást a $\boxed{\text{3. tábl.}}$ segével számoljuk:

i	x_i	H_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2 \cdot H_i$	$(x_i - \bar{x})^3 \cdot H_i$	$(x_i - \bar{x})^4 \cdot H_i$
	$\bar{x}$			σ	β_3	β_4
			P		γ	Exz

Pl.

Adatok rendezése és elsődleges jellemzése **1. HF** **Lev.**

(leadási határidő: 2. előadás

1. adatsor

1, 3, 3, 3, 3, 4, 2, 3, 4, 4, 3, 4, 2, 3, 1, 4, 1, 3, 2, 2, 3, 2, 2, 5, 3, 1, 2, 4, 2, 2, 4, 1, 4, 1, 2, 2
3, 3, 2, 4, 3, 3, 4, 5, 1, 1, 5, 4, 2, 5 $(n=50)$

a, $x_{min}=?$, $x_{max}=?$, $R=?$

b, írja fel a rendezett mintát! $x_{0,25}=?$, $x_{0,5}=?$, $x_{0,75}=?$, kvartilis-távolság

c, készítse el az 1. táblázatot:

i	érték: $x_{(i)}$	abszolút gyak. H_i	relatív gyak. h_i	összegzett rel. gyak. $\sum_{j=1}^{i} h_j$

d, rajzolja fel a rel. gyak.-i poligont (=tapasztalati sűrűségfv.-t) hisztogramot!

e, rajzolja fel az összegzett rel. gyakoriság-i poligont (=tapasztalati eloszlásfv.-t) hisztogramot!

f, $\bar{x}=?$, $Me=?$, $D=?$

2. adatsor

(Mat. vizsgán elért összpontszám: [0,60])

18, 16, 25, 21, 18, 46, 24, 27, 23, 10
11, 10, 29, 17, 21, 19, 17, 18, 28, 31
14, 46, 21, 18, 23, 11, 42, 39, 19, 30
43, 14, 29, 16, 35, 21, 12, 28, 14, 32
51, 7, 42, 24, 21, 18, 37, 25, 33, 25

a) $n=?$, $R=?$, $x_{min}=?$, $x_{max}=?$

b, adjon meg egy $[a,b]$ int.-t, úgy, hogy $\forall$ adat benne legyen és $\left(\frac{b-a}{5}\right)$ könnyen számolható legyen!

c) ossza $[a,b]$-t 5 egyenlő részre, ezek lesznek az osztályok. Adja meg az osztályközepeket és készítse el az 1. táblázatot

d) $u_{min}=?$, $u_{max}=?$, $R_u=?$
$u_{0,25}=?$, $u_{0,75}=?$, $Me_u=u_{0,5}=?$

e, Rajzolja fel a tapasztalati sűrűség- és eloszlásfv.-t!

3. adatsor

4, 1, 44, 1, 2, 2, 3 51, 4, 5, 5, 4, 5, 2, 4, 2, 1, 3, 2, 2, 2, −55, 1, 1, 4, 3, 2, 4, 1, 5, 2, 3, 4, 4, 1,
2, 2, 3, 4, 4, 1, −42, 3, 51, 4, 2, 2, 3, 5, 2, 3

a, x_{min}, x_{max}, R, $\bar{x}$, D, Me

b, készítse el a Box-Plot ábrát!

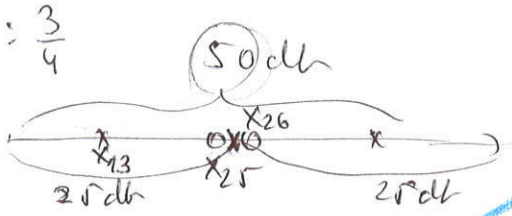
minden példában adja meg Δ, σ, δ és ε_{KV} értékeit is!

$x_{0,25} = x^*_{13}$

$Me = x_{0,5} = \frac{x^*_{25} + x^*_{26}}{2}$

$x_{0,75} = x^*_{37}$

$\frac{1}{4} : \frac{3}{4}$



$x_{0,25} = 1 \cdot \frac{(n+1)}{4}$

$x_{0,5} = \frac{n+1}{4} \cdot 2$

$x_{0,75} = \frac{n+1}{4} \cdot 3$

$\frac{51}{4} = 12,8$

$x_{0,25} = \frac{x^*_{12} + x^*_{13}}{2}$

$p : (1-p)$

$x_{0,25}$ -nél kisebb adatok és nagyobb adatok számának aránya : 1 : 3

$x_{0,75}$ 〃 〃 〃 〃 〃 : 3 : 1

$x_{0,5}$ -nél 〃 〃 〃 〃 1 : 1