Leíró statisztika

Első közelítésben a statisztikai tevékenységeket négy csoportba sorolhatjuk, de ezek között nincs éles határ:

1. adatgyűjtés,
2. az adatok áttekinthetővé tétele,

(ehhez nincs szükség a valószínűség fogalmára:

leíró statisztika)

3. az adatok elemzése,
4. következtetések

(a valószínűségszámítás alapjai nélkül nem nagyon érthető)

induktív statisztika

1. adatgyűjtés

az adatgyűjtés valamilyen **cél** eléréséhez szükséges (azonosítás, megkülönböztetés)

az adatok egy része ismert, csak meg kell kérdezni valakitől, másik részét csak meg kell figyelni, egy harmadikat meg kell mérni valahogy (orvosi vizsgálat)

2. az adatok áttekinthetővé tétele

A mindennapi életben is gyakran előfordul, hogy egy probléma kapcsán viszonylag sok adat áll rendelkezésünkre.
Ilyen esetekben szükséges, hogy az adatokról valamilyen áttekintésünk legyen.

2/a táblázat

Kórokozó	*Betegség*	abszolút gyakoriság		relatív gyakoriság		feltételes relatív gyakoriság	
baktérium	Salmonellosis (szalmonella fertőzés)	94	208	0,280	0,619	0,452	1,000
	Scarlatina (skarlát)	102		0,303		0,490	
	egyéb bakteriális eredetű	12		0,036		0,058	
vírus	Hepatitis infectiosa (fertőző májgyulladás)	22	126	0,065	0,375	0,175	1,000
	Mononucleosis infectiosa	22		0,065		0,175	
	Lyssa (veszettség)	74		0,220		0,587	
	egyéb vírusos eredetű	8		0,025		0,063	
egyéb	egyéb fertőző betegségek	2	2	0,006	0,006	1,000	1,000
összesen		336	336	1,000	1,000		

abszolút gyakoriságok
relatív gyakoriságok
hányados, ezért nemcsak azt kell tisztázni, hogy minek a relatív gyakoriságáról beszélünk, hanem azt is, hogy mihez viszonyítunk
feltételes relatív gyakoriságok
a különbség csupán annyi, hogy itt szűkebb összességhez viszonyított relatív gyakoriságról van szó

Összegzési szabályok
az abszolút gyakoriságok mindig összeadhatók

a relatív gyakoriságok is összeadhatók abban az esetben, ha ugyanahhoz az összességhez viszonyítjuk őket

Szorzási szabály

$$\frac{\text{skarlátosok száma}}{\text{baktériummal fertőzöttek száma}} \cdot \frac{\text{baktériummal fertőzöttek száma}}{\text{összes fertőzöttek száma}} = \frac{\text{skarlátosok száma}}{\text{összes fertőzöttek száma}}$$

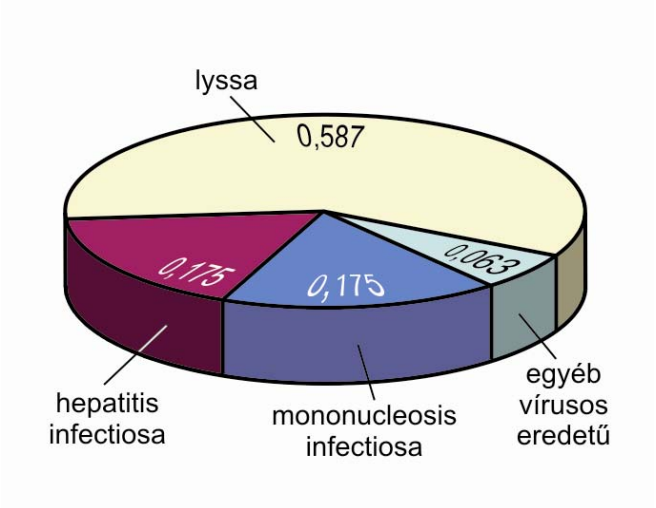
Feladat:

Egyetemünkön a tavalyi vizsgákon az elégtelen osztályzatok relatív gyakorisága 0,15, a sikeres vizsgák között a jelesek relatív gyakorisága 0,2 volt. Az összes vizsgajegy között mennyi volt a jeles relatív gyakorisága?

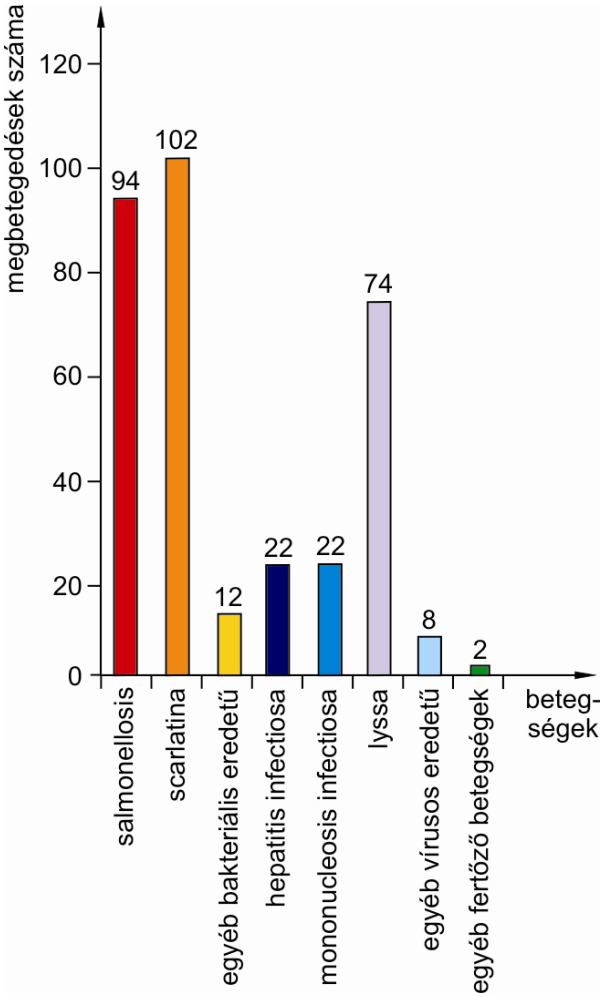
$$\frac{\text{elégtelenek száma}}{\text{összes hallgató száma}} + \frac{\text{sikeres vizsgák száma}}{\text{összes hallgató száma}} = 1$$

$$\frac{\text{sikeres vizsgák száma}}{\text{összes hallgató száma}} \cdot \frac{\text{jelesek száma}}{\text{sikeres vizsgák száma}} = \frac{\text{jelesek száma}}{\text{összes hallgató száma}}$$

2/b grafikon



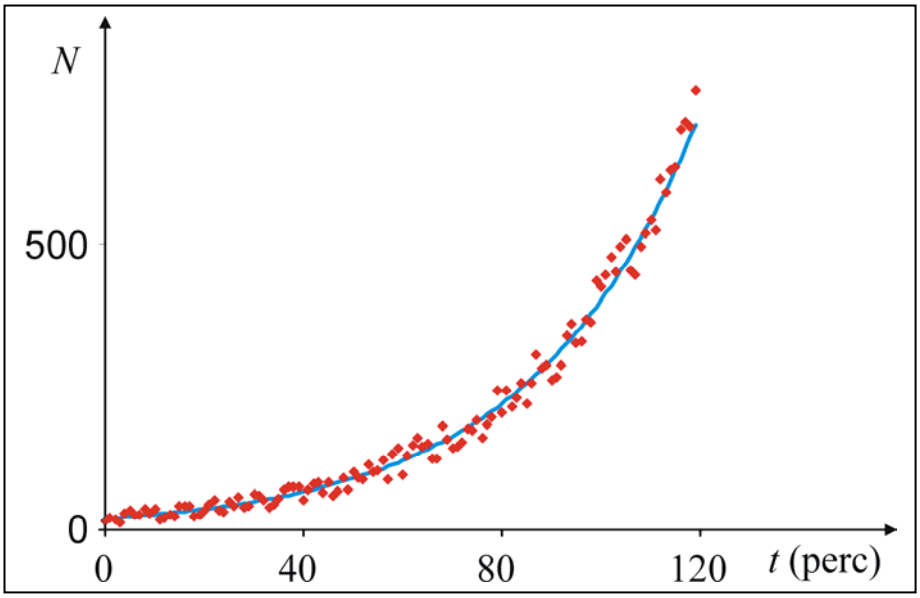
torta diagram



oszlop diagram

Számszerű adatok jellemzői

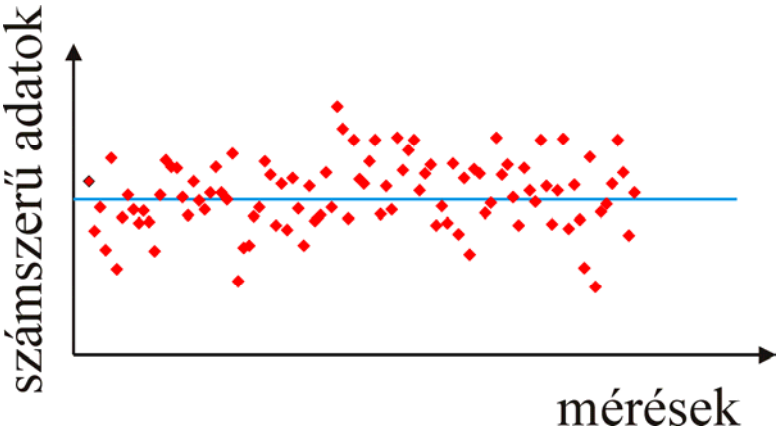
Emlékeztető: baktérium populáció szaporodása



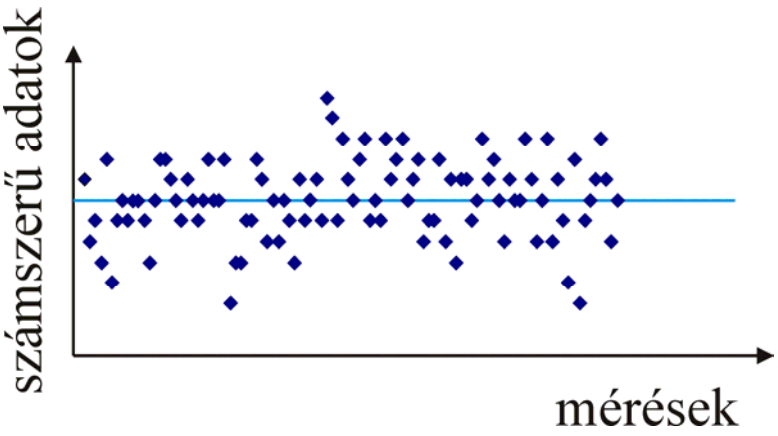
A változások determinisztikus és sztochasztikus része mindig **egyszerre** fordul elő.

Van-e mód a szétválasztásukra?

Kiindulásképpen tegyük fel, hogy egyelőre csak olyan adatokat vizsgálunk, ahol nincs determinisztikus változás.



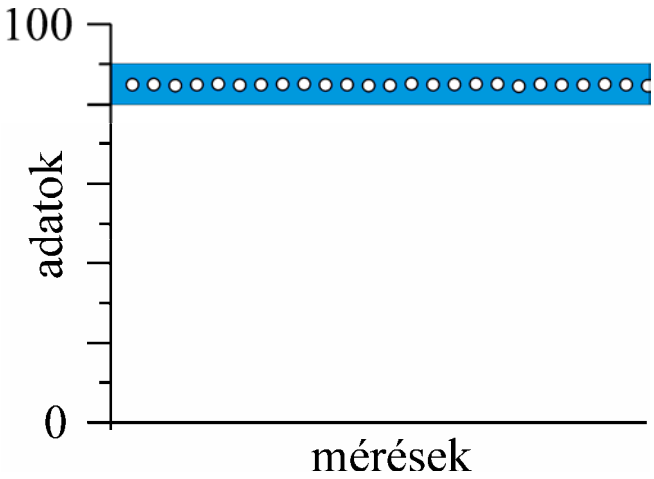
„folytonos” eset

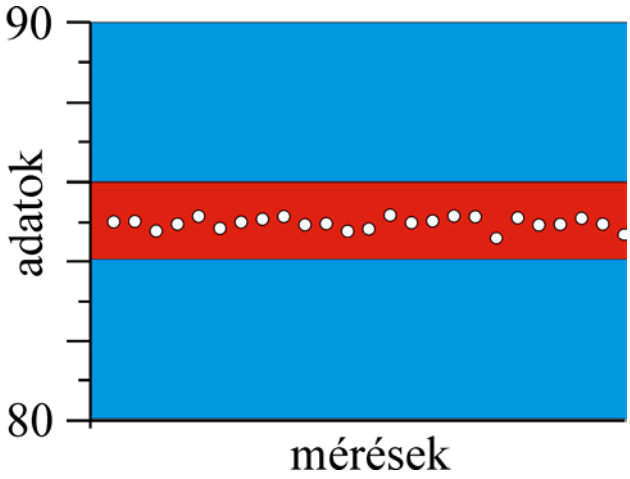


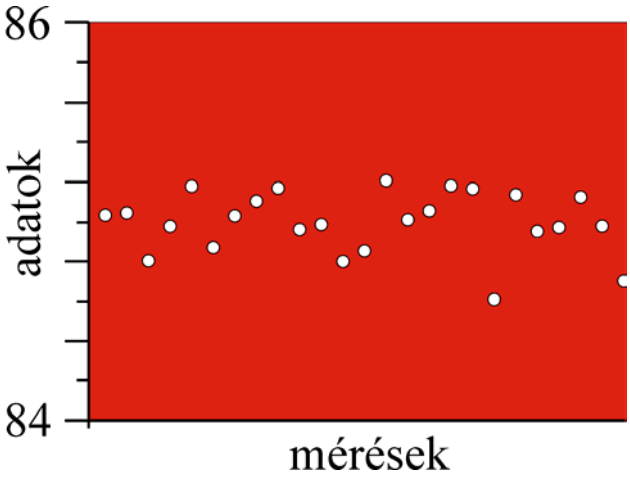
diszkrét eset

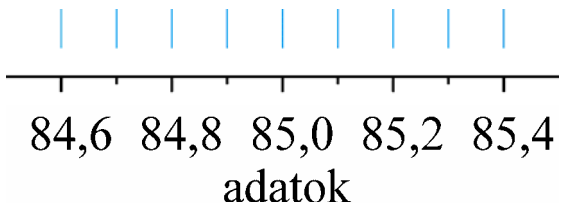
Vegyünk konkrét számokat.

85,0361847182	85,0
85,0472573819	85,0
84,8050250975	84,8
84,9817905059	85,0
85,1795262649	85,2
+ 95 számadat	



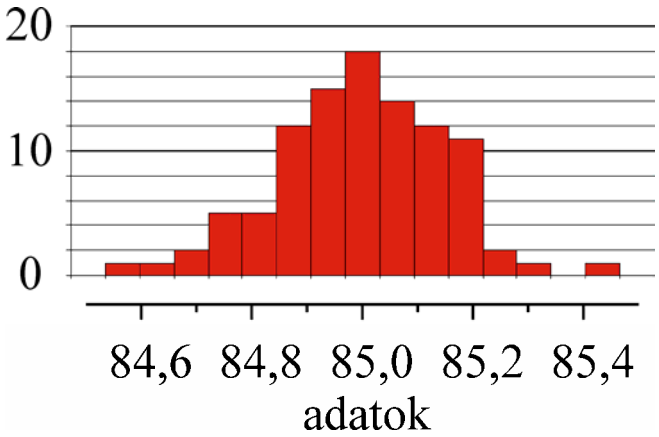


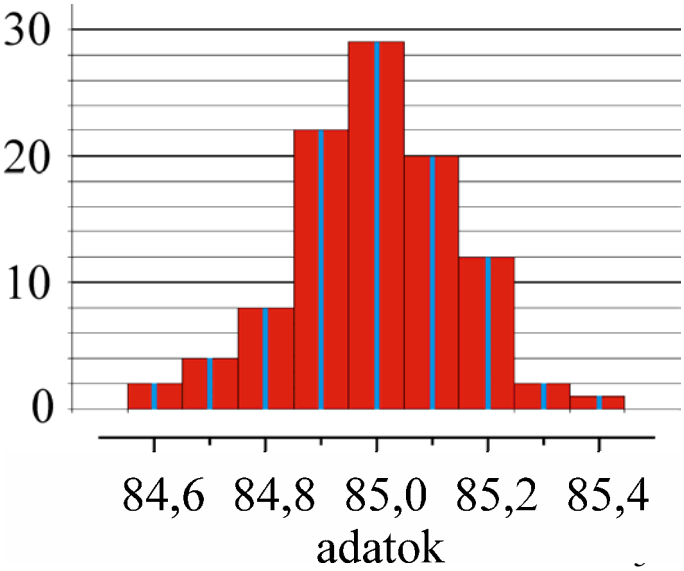




A folytonos esetben „sohasem” kapunk egyforma adatokat.

A diszkrét esetben meg kell adnunk a gyakoriságokat is.





Gyakorisági eloszlás

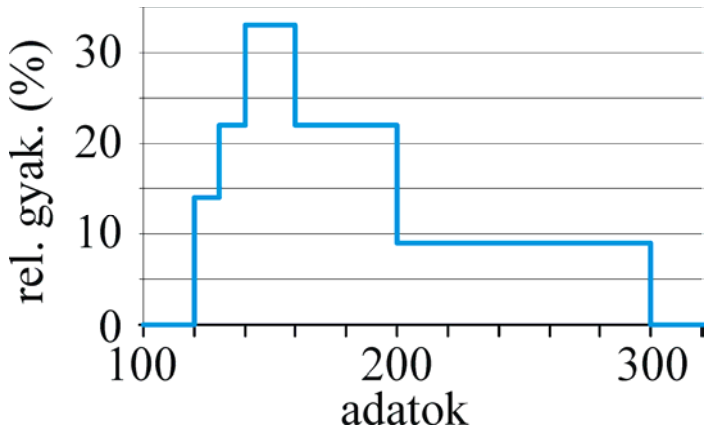
A diszkrét esetben egyértelmű.
A folytonos esetben alakja függ az intervallumok, más néven az **osztályok** megválasztásától (de általában nem nagyon).

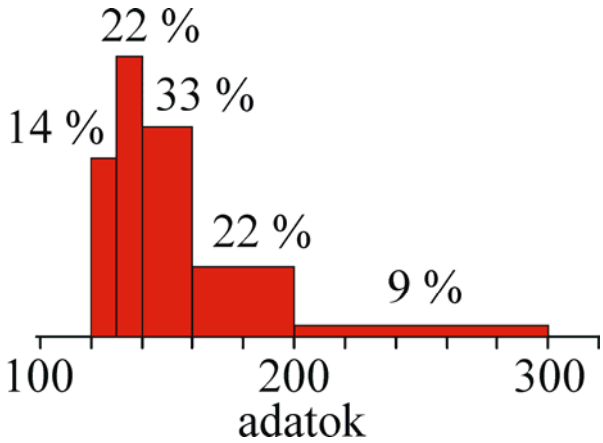
A jobb összehasonlíthatóság kedvéért sokszor a relatív gyakoriságokat adjuk meg.

Pl. Fizetési kimutatás (Ft):

120 és < 130 ezer között	124	14%
130 és < 140 ezer között	195	22%
140 és < 160 ezer között	293	33%
160 és < 200 ezer között	195	22%
200 és 300 ezer között	80	9%
összesen	887	100%

Hogyan szemléltessük?





Hisztogram

az oszlopok területe adja meg a relatív gyakoriságokat.
A teljes terület 100 %, vagyis 1.

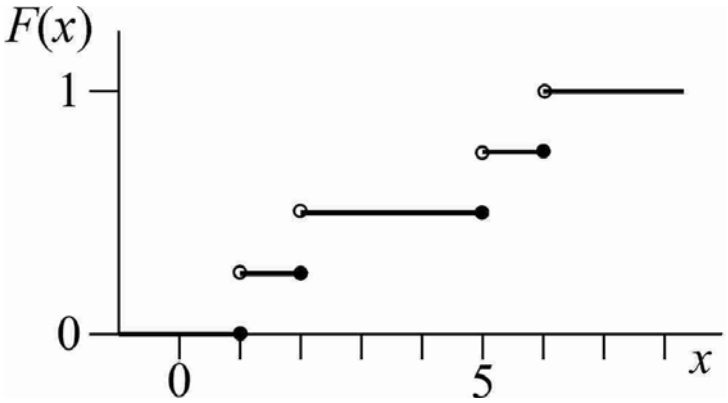
Amennyiben az oszlopok egyforma szélesek, – tehát egyenközű osztályszélességek esetén – a két ábrázolás azonos.

Eloszlásfüggvény

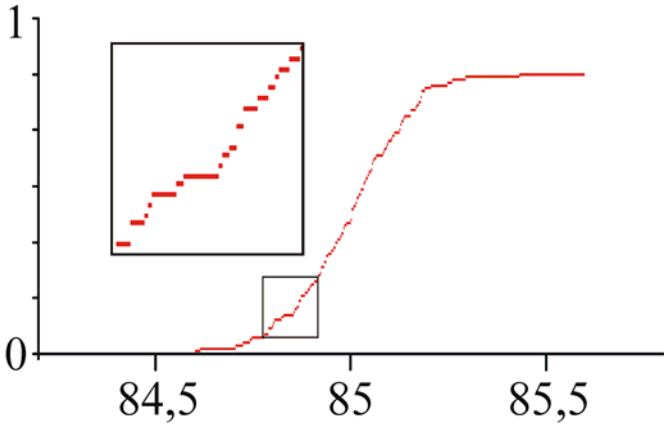
Az $x_1, x_2, x_3, \ldots, x_n$ adatrendszer F eloszlásfüggvénye:

$$F(x) = \frac{\text{az } x - \text{nél kisebb adatok darabszáma}}{n}$$

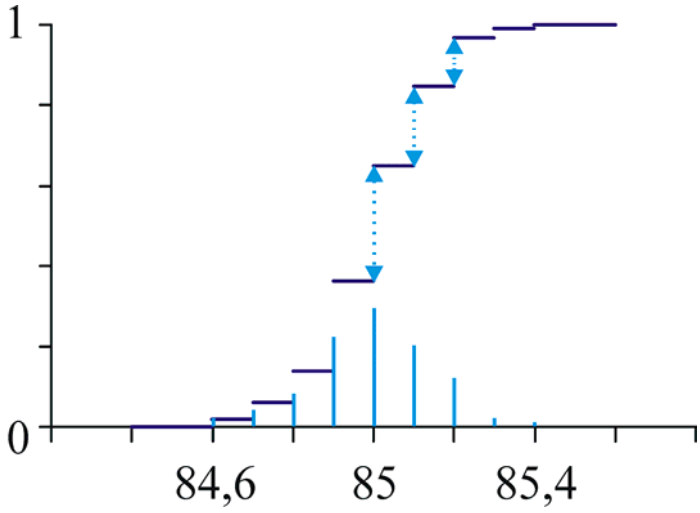
Pl. (1) [1, 2, 5, 6]



Pl. (2) A korábbi 100 adat folytonos esetben.



Pl. (3) A korábbi 100 adat diszkrét esetben.



Az oszlopok a relatív gyakoriságokat mutatják. Ezeket egymás után összeadva megkapjuk az eloszlásfüggvény megfelelő értékeit. Az állítás megfordítása is igaz. Az eloszlásfüggvény különbségei a relatív gyakoriságokat adják meg.

Kevés adat esetén nem látszanak a gyakorisági eloszlás jellegzetességei.

egy csúcsú			több csúcsú
szimmetrikus	jobbra ferde	balra ferde	
csúcsos	lapult		

Számszerű jellemzők (mindig meghatározhatók):

Középértékek (*n* elemű adatrendszer)

1. az adatrendszer **átlaga** (számtani közép)

1.Pl. 1,2,2,7; 2.Pl. 1,1,3,7

$$x_{\text{átlag}} = \bar{x} = \frac{1+2+2+7}{4} = 3 = \frac{1}{4}\sum_{i=1}^{4} x_i = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Érzékeny a kiugró értékekre!

Helyzeti középértékek (helyzetüknél fogva azok)

2. az adatrendszer **mediánja** ($x_{\text{medián}}$)
nagyság szerint sorba rendezzük az adatokat és megkeressük a középsőt vagy középsőket

1.Pl. 1,2,2,7 $x_{\text{medián}} = 2$; 2.Pl. 1,1,3,7 $1 \le x_{\text{medián}} \le 3$

3. ha az adatrendszerben vannak azonosak, akkor azt, amelyikből a legtöbb van az adatrendszer **móduszának** ($x_{\text{módusz}}$) nevezzük (ebből lehet több is)

(„mode” divatos)

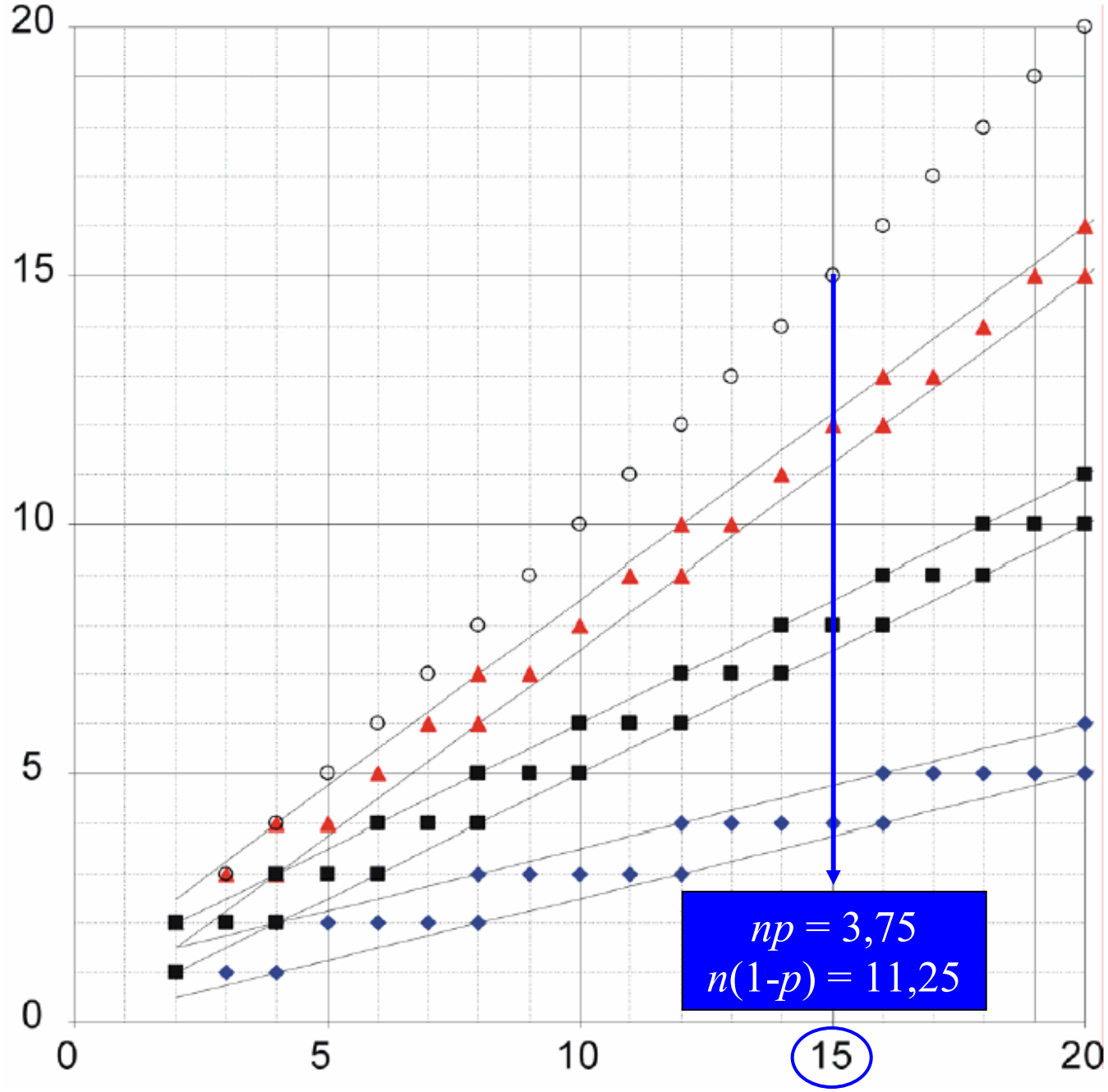
1.Pl. 1,2,2,7 $x_{\text{módusz}} = 2$; 2.Pl. 1,1,3,7 $x_{\text{módusz}} = 1$

Nem érzékenyek a kiugró értékekre!

Kvantilisek (osztóértékek)

Mekkora jövedelem esetén tartozik valaki a felső „tízezer”-be.

Az adatokat először itt is nagyság szerint sorba rendezzük.
Pl. alsó kvartilis, középső kvartilis = medián, felső kvartilis



Legyen p 0 és 1 közötti szám az adatrendszer p-kvantilisének nevezzük azt a számot, amelynél kisebb adatok darabszáma legfeljebb np és amelynél nagyobb adatok darabszáma legfeljebb $n(1-p)$

kvintilis ($p = 1/5$), decilis ($p = 1/10$), percentilis ($p = 1/100$)

A szóródás jellemzői

1. az adatrendszer legnagyobb és legkisebb elemének eltérése az adatrendszer **terjedelme**

Mekkora az adatok (átlagos) eltérése az átlagtól?

1.Pl. 1,2,2,7 $(3-1)+(3-2)+(3-2)+(3-7)=0$

2.Pl. 1,1,3,7 $(3-1)+(3-1)+(3-3)+(3-7)=0$

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

Mekkora az adatok (átlagos) abszolút eltérése a mediántól?

Belátható, hogy minimális, azaz

legyen x^* egy rögzített érték, akkor az

$\frac{1}{n}\sum_{i=1}^{n}|x_i-x^*|$ kifejezés akkor minimális, ha $x^* = x_{\text{medián}}$

Mekkora az adatok (átlagos) négyzetes eltérése az átlagtól?

Belátható, hogy minimális, azaz

legyen x^* egy rögzített érték, akkor az

$\frac{1}{n}\sum_{i=1}^{n}(x_i-x^*)^2$ kifejezés akkor minimális, ha $x^* = x_{\text{átlag}}$

2. az adatrendszer átlagától vett négyzetes eltérések átlagát az adatrendszer **szórásnégyzetének** vagy **varianciájának** nevezzük,

az adatrendszer **szórását** pedig az

$$s_x=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}$$ kifejezés adja meg

További jellemzők is megadhatók (ferdeségre, csúcsosságra).