A MATEMATIKAI STATISZTIKA ELEMEI

Az Eötvös Lóránd Tudományegyetem Természettudományi Karán a Fizikai Kémiai Tanszék évek óta kémia-szakos tanárhallgatóknak matematikai bevezetõ elõadásokat tart. Az elõadások célja az, hogy a hallgatóságnak ne legyenek idegenek azok a matematikai levezetések, állítások, amelyekkel késõbbi, fõleg fizikai-kémiai tanulmányai során találkozik.

A valószínûségszámítási fejezetek ismertetése után a matematikai statisztika elemei következnek. Ennek a tanfolyamnak lényeges részeit tartalmazza jelen vázlatos ismertetés, elsõsorban azért, hogy a hallgatóság a tanulás elõsegítésére azokat a számítógépes hálózatról magának letölthesse. Az anyag a félév végéig fokozatosan kerül fel a hálóra, tartalomjegyzéke ennek megfelelõen folyamatosan bõvül.

A tárgy elõadója Szepesváry Pál.

Budapest 2002 április

TARTALOM

1. A matematikai statisztika jellemzése

A matematikai statisztika a *véletlen* (valószínûségi) változókkal jellemezhetõ (továbbiakban *véletlen*) *rendszerek* leíró adatainak feldolgozásáról, értelmezésérõl és felhasználásáról szóló tudományos módszertan.

Amíg a valószínûségszámítás fogalmai axiomákkal összhangban definiált vagy azokból levezetett absztrakt fogalmak, amelyek tulajdonságai ílymódon adottak, a matematikai statisztika megfigyelt, leszámlált vagy mért sajátságokat feleltet meg a valószínûségszámítás absztrak fogalmainak, sokszor megállapodásszerû módon. Szokásos mondás: "amíg a valószínûségszámítás megtanit valószínûségekkel számolni, addig a statisztika megtanit valószínûséget mérni".

Miután a véletlen által befolyásolt jelenségek nem biztos kimenetelûek, a matematikai statisztikában nincsenek biztos ítéletek. A matematikai statisztika *becsül,* megbecsülhetõ valószínûségû ítéleteket hoz.

Igen ritka az az eset, amelynél egy véletlen rendszer viselkedését minden elképzelhetõ kimenetelnél meg lehet figyelni. A matematikai statisztika következésképpen csak a rendszer valamely szemügyre vett részletébõl, valamely folyamat pillanatnyi állapotából, tehát a rendszer egy *mintájából* következtet magára a rendszerre. Ez a statisztikus megállapítások bizonytalanságának további oka.

A matematikai statisztika feladata tehát (1) jellemzõ számadatok, megállapítások levezetése, bemutatása megfigyelt adatokból, (2) valószínûség hozzárendelése a kapott vagy levont következtetésekhez, (3) döntés valamely fent alapon megfogalmazott állítás (*hipotézis*) elfogadásáról vagy elvetésérõl, végül, (4) olyan kisérleti feltételek meghatározása (olyan *kisérletek tervezése*), amelyek számunkra az állítások megbízhatósága szempontjából legkedvezõbbek.

4. Leíró és felderítõ statisztika

Vizsgált rendszereink vagy teljesen ismeretlenek vagy vannak róla elõzetes (*a priori*) ismereteink. Ha vannak, képesek vagyunk többé-kevésbé alkalmas (adekvát) matematikai *modellt* alkotni, és ez esetben a statisztikai adatgyüjtés célja a modell *paramétereinek* megbecslése. Ha nincsenek elõzetes ismereteink, a leíró és felderítõ statisztika módszereit alkalmazzuk, amelyekre persze a modell alapú vizsgálatoknál is szükség van. A felderítõ statisztika az adatok, a minta kezelésére, jellemzésére, ábrázolására vonatkozóan ad útmutatásokat, több változó esetén pedig számos további feladatot old meg (alakfelismerés, csoportosítás, osztályozás).

5. Sokaság és minta

Viszgálatunk tárgya egy *rendszer*. Egy rendszernek elemei (*objektumai*) vannak, az objektumoknak *tulajdonságai*.

(Objektumok például: emberek, társadalmak, folyók, biotópok, oldatok, spektrumok, tulajdonságok az emberek testméretei, emberek, társadalmak, folyók, biotópok, oldatok, spektrumok, tulajdonságok az emberek testméretei, a társadalmak lakosságszáma, nemzeti jövedelme, a folyók vízhozama adott idõben, helyen, biotópok fajainak száma, egyedsûrüsége, oldatok koncentrációi, spektrumok csúcsmagasságai adott hullámhosszon stb.)

Egy rendszernek általában sok objektuma, azoknak sok, számos esetben végtelen sok értékû tulajdonsága van. A rendszert alkotó objektumok, pontosabban azok tulajdonságait leíró (végtelen) sok jellemzõ változó adat alkotja az adatok *sokaságát*. A sokaság elemei tehát lehetnek fizikai létezõk, de elméletiek is. A sokaság szabatos meghatározása fontos feltétele a statisztikai munkának, hiszen ez jelenti a feldolgozásra váró adatok pontos meghatározását.

(Egy folyó vizállása április 16-án és november 1-én például két statisztikai sokaság).

Általában csak arra van módunk, hogy a rendszer egy részletét, vagy egy bizonyos állapotát figyeljük meg, azaz annak leíró adataiból *mintát* vegyünk. Szokás mondani: a sokaság az összes elképzelhetõ minta halmaza.

A minta vizsgálatának eredményébõl következtetünk a sokaságra, a minta vétele tehát az eredmények értéke szempontjából elsõrendûen fontos. A minta legyen

(a) reprezentatív, összetételében képviselje helyesen a sokaságot, amelybõl vették,

(b) véletlen, a mintaelemek kerüljenek egymástól függetlenül, egyenlõ valószínûséggel a mintába,

(c) elégséges méretû, elegendõen nagy ahhoz, hogy a minta alapján levont következtetések kellõen valószínûek legyenek.

3.1 Az adatok

3.1.1 Az adatok fajtái

Az adatokat *kategorikus* és *nem kategorikus* (kvantitatív) jellegûekre szokás felosztani. A kategorikus adatok alapján az objektumokat *osztályozni lehet*. A kategorikus adatok lehetnek *nevesítõek* (nominálisak) és *rendezõek* (ordinálisak). A nevesitõ adat egy-egy objektumot valamely (esetleg egyelemû) osztályba osztályba sorol, a rendezõ adat már sorrendet is definiál. (3.1/a táblázat)

3.1/a táblázat. Kategorikus adatok

Adatfajta	Az adatokon értelmezhetõ müvelet	Példa
Nevesítõ (nominális)	= , ≠	Nem, név, állampolgárság, foglalkozás, telefonszám
Rendezõ (ordinális)	= , ≠ , < , >	Iskolai osztályzat, rang, betegség foka, IQ

Azokat a kategorikus adatokat, amelyek csak két osztály valamelyikébe sorolhatnak, *dichotómikus* vagy *bináris* adatoknak nevezik.

(Dichotómikus adatok: férfi-nõ, igaz-hamis, kicsi-nagy, beteg-egészséges)

A *kvantitatív* adatok lehetnek *folytonos* vagy *diszkrét* (mérhetõ vagy leszámlálható, gyakran *metrikusnak* nevezettek) adatok. Szokásosan megkülönböztetik azokat adatokat, amelyek skálájának önkényes a 0-pontja,.lényegében különbségük értelmes (intervallum skála) azoktól, amelyekre multiplikatív aritmetikai mûveletek is alkalmazhatók (arányos skála).(3.1/b táblázat).

3.1/b táblázat. Metrikus adatok példái

Adatskála	Folytonos	Diszkrét
Intervallum	Potenciál, Celsius fokban mért hõmérséklet	Naptári napok
Arányos	Tömeg, Abszolut hõmérséklet	Részecskeszám

Vegyészi gyakorlatunkban az esetek túlnyomó részében metrikus adatokkal (tömeg, anyagmennyiség, térfogat, koncentráció, nyomás, hõmérséklet, energiák sebességek) van dolgunk.

3.1.2 Az adatok kezelése, a skálázás

A sokaságból vett *n* elemû minta *i*-edik adata egy:

mintaelem	x_i	$i = 1,2 \ldots n$	(3.1)

A mintaelemek sorozata a

minta	$\mathbf{x} = x_1, x_2, \ldots, x_n$	(3.2)

ahol i index az adat mérési sorszáma.

Ha a minta adatait nagyságuk szerint állítjuk sorba, a *rendezett mintához* jutunk:

A rendezett minta	$x_1^*, x_2^*, \ldots, x_n^*$	$x_1^* \le x_2^* \le \ldots \le x_n^*$	(3.3)

3.1.3 Az adatok skálázása

Egy minta természetes terjedelmét a számegyenesen a legkisebb és legnagyobb értékû mintaelem határozza meg. Különbozõ okokból szükség lehet arra, hogy ezt a terjedelmet módosítsuk, hogy az adatokat más egységben, más skálán tekintsük. Ezt *skálázással* lehet elérni, amelynek során az eredeti mintaelemekhez valamely számot hozzáadunk, vagy/és azokat valamely azonos számmal osztjuk. A számos skálázási lehetõség közül a vegyészi gyakorlatban a *mértékegységváltás*, a minta *normálása* 0 és 1 érték közé (móltört, tömegtört megadás), a minta *centrálása*, és a minta *standardizálása* leggyakoribbak.

Normált mintához jutunk, ha az eredeti minta minden elemét az elemek összegével osztjuk. Ennek egy eleme:

$$z_i = \frac{x_i}{\sum_{k=1}^{n} x_k} \qquad (3.4)$$

Felhívjuk a figyelmet arra, hogy az így normált adatok között egy már független a többitõl, az adatok összegébõl és a *n* - 1 adatból a függõ már kiszámítható.

Centrált minta keletkezik, ha minden elembõl kivonjuk az elemek átlagát (l. 3.7 képlet):

$$x_i^{(c)} = x_i - \bar{x} \qquad (3.5)$$

A centrált mintában szükségképpen pozitiv és negatív értékek lépnek fel, az elemek összeg 0. Ebbõl következik, hogy a centrált adatok közül is csak *n* - 1 darab független.

A standardizált lesz a minta akkor, ha az eredeti mintaelemekbõl kivonjuk azok átlagát és a különbségeket a minta empirikus szórásával (l. 3.10 képlet) osztjuk:

$$u_i = \frac{x_i - \bar{x}}{s} \qquad (3.6)$$

A standardizált minta 0-közepû, szórása 1

3.1.4 Az adatok ábrázolása

Mintákról szemléletes képet ad a pontsor, azaz a mintaelemek ábrázolása a számegyenesen, az (egyváltozós) szóródási kép (*univariate scatter plot*).

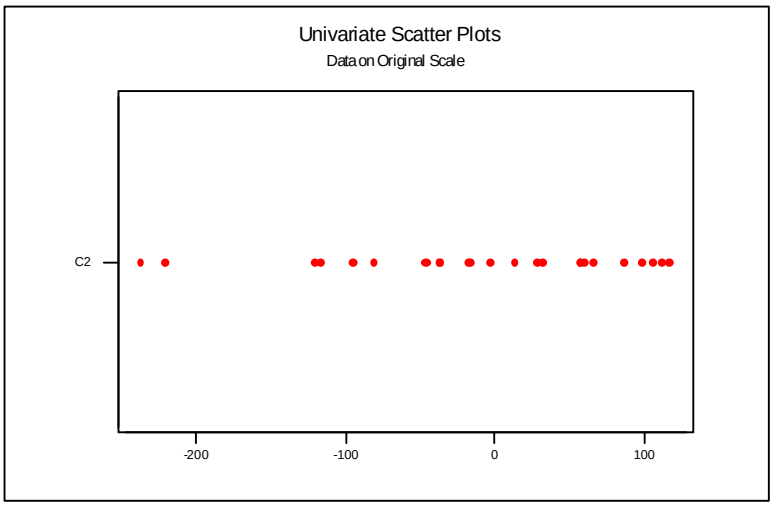
3.1. példa: Tekintsünk egy 24 elemû mintát:

-37,-46, 67,-81,17, 107, 33, -120, 113, -236, -2, -220
99, 117, 57, -35, 60, -117, -95, -16, 14, 29. 58, 87

Rendezve:

-236,-220,-120,-117,-95,-81,-46,-37,-25,-17,-16,-2
14, 29, 33, 57, 58, 60, 67, 87, 99, 107, 113, 117

Pontsorral ábrázolva:



3.1 ábra Pontsoros ábrázolás

3.2 Az adatok eloszlása, a minták jellemzói

Bár az adatok sorozatának és képének megtekintése bizonyos fokig tájékoztat az adatok elhelyezkedésérõl, szükség van olyan számadatokra, amelyek tömören jellemzik a minta (a) közepét, (b) terjedelmét és (c) eloszlását. Egy-egy célra több jellemzõ közül lehet választani.

A valószínûségszámítás sokaságok eloszlásának jellemzésére pontosan definiált mennyiségeket, mint várható érték, szórás, ferdeség, lapultság, korrelációs együttható stb. Levezette ezeknek a mennyiségeknek tulajdonságait is. Az alábbiakban ismertetett tapasztalati (kisérleti, gyakran *statisztikáknak* nevezett) jellemzõk ezeknek az elméleti mennyiségeknek *becslései.* A becslések között különösen értékeljük azokat, amelyek *torzítatlanok.* Torzítatlan az a becslés, amelynek várható értéke megegyezik azzal a mennyiséggel, amelyiket becsül.

3.2.1 Mintaközép jellemzõk

a) *számtani közép,* mintaátlag, (*mean*)

A számtani közép $$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$ (3.7)

A számtani közép a hagyományos legkisebb négyzetek elvének megfelelõ jellemzõ, a várható érték torzítatlan becslése. Hátránya, hogy érzékeny a szélsõségesen eltérõ ("kilógó") adatokra.

Az 3.1 példában szereplõ adatok számtani közete: - 7.542

b) *medián,* (*median*)

A medián az *x* változó azon értéke, amelynél a minta elemek fele kisebb, fele nagyobb.

$$\tilde{x} = x_{\frac{n+1}{2}} = x_{m+1}$$ ha a minta páratlan elemõ, *n = 2m+1* (3.8/a)

$$\tilde{x} = \frac{x_m + x_{m+1}}{2}$$ ha a minta páros elemû, *n = 2m* (3.8/b)

A medián nem érzékeny szélsõséges értékekre, u.n. *robusztus* becslõ.

Az 3.1 példában szereplõ adatok medánja: 6

c) *módusz* (*mode*)

A módusz a leggyakrabban elõforduló mintaelem értéke*

$$d = x_{leggyakoribb}$$ (3.9)

* több maximumos eloszlásoknál a leggyakoribb, majd a második leggyakoribb...

A módusz a valószínûségi változó sûrûségfüggvényének maximumhelye. Kisérleti meghatározása nagy mintákból lehetséges, ahol beszélhetünk azonos értékû mintaelemekrõl-rõl.

További, adott esetben hasznos, de gyakorlatunkban ritkábban elõforduló mintaközép jellemzõk még a *mértani közép:*

$$\bar{x}^{(g)} = \sqrt[n]{x_1 . x_2 \ldots x_n}$$

és a *harmonikus közép*:

$$\bar{x}^{(h)} = \left(\frac{\frac{1}{x_1} + \frac{1}{x_2} + \ldots + \frac{1}{x_n}}{n} \right)^{-1}$$

3.2.2 Kiterjedés jellemzõk

a) *standard deviáció* (tapasztalati szórás, korrigált empirikus szórás (*standard error, standard deviation*):

Standard deviáció $$s = \sqrt{\frac{\sum \left(x_i - \bar{x}\right)^2}{n-1}} = \sqrt{\frac{\sum x_i^2 - n\bar{x}^2}{n-1}}$$ (3.10)

Ez a jellemzõ az elméleti szórás becslése. Nevezõjében a kézenfekvõ *n* helyett azért szerepel *n* - 1, mert azt csak *n* - 1 független mért adatból számíthattuk ki. A számtani közép ugyanis egy adatot az *n* közül a többibõl kiszámíthatóvá tesz. Ha a nevezõben *n* állna, a standard deviáció torzítottan becsülné a szórást.

Fontos megjegyzés: Ha *n* adat között *m* darab megkötés létezik, az *n* adat között csak *n - m* darab független. A független adatoknak ezt a számát szabadsági foknak (degree of freedom, DF) is nevezik.

Az 3.1 példában szereplõ adatok standard deviációja: 98.64

b) *variációs együttható (coefficient of variation)*

$$"c.v."= V = s / \bar{x} \qquad (3.11)$$

A variációs együttható azt mutatja meg, hányadrésze, hány százaléka a tapasztalati szórás a középértéknek. Bizonyos esetekben (pl 0 várható értékû sokaságoknál) értelmetlen.

c) *terjedelem* (*range*)

A terjedelem a legnagyobb és legkisebb mintaelem különbsége

$$d = x_{max} - x_{min} \qquad (3.12)$$

Az 3.1 példában szereplõ adatok terjedelme: 117 - (-236) = 353

d) *kvantilisek* (*quantiles*)

p-s kvantilis az *x* változó azon értéke, amelynél kisebb mintaelemek hányada *p*)

0.1-es kvantilis	= decilis		= 10. percentilis
0.25-ös kvantilis	= elsõ kvartilis	(Q_1)	= 25. percentilis
0.5-ös kvantilis	= második kvartilis	(Q_2)	= 50. percentilis = medián
0.75-ös kvantilis	= harmadik kvartilis	(Q_3)	= 75. percentilis
0.90-es kvantilis			= 90. percentilis

Az 3.1 példában szereplõ adatok elsõ kvartilisa -63.5, mediánja 6, harmadik kvartilisa 63.5

3.2.3 Egyéb eloszlásjellemzõk

3.2.3 Egyéb eloszlásjellemzõk

a) *ferdeség (skewness)*

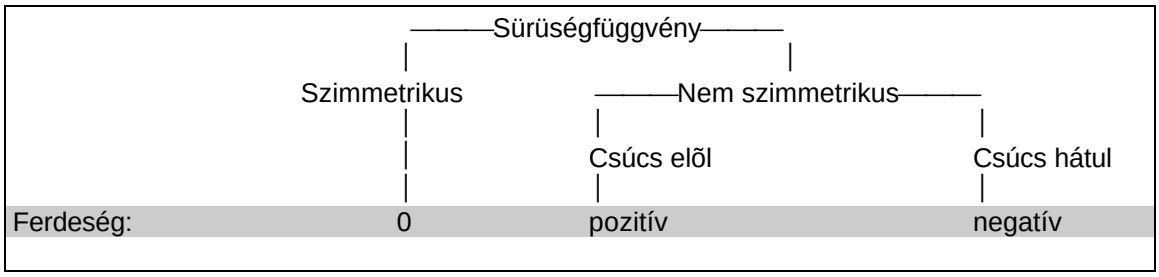
A ferdeség

$$g_1 = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s^*} \right)^3 \qquad (3.13)$$

Ez a mennyiség a harmadik centrális momentum/szórás^3 módon, a

$$\gamma_1 = \frac{E\left[(\xi - E(\xi))^3\right]}{\sigma^3}$$

képlettel definiált mennyiség becslése. A ferdeség valószínûségi változóknak különbözõ sûrûségfüggvényei esetén az alábbiak szerint alakul:



Az 3.1 példában szereplõ adatok ferdesége: -0.7285

b) *lapultság (kurtosis)*

A lapultság: $$g_2 = \frac{n(n-1)}{(n-1)(n-2)(n-3)} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s^*} \right)^4 - 3\frac{(n-1)^2}{(n-2)(n-3)} \quad (3.14)$$

A lapultság a

$$\gamma_2 = \frac{E\left[(\xi - E(\xi))^4\right]}{\sigma^4} - 3$$

képlettel, γ_2 = negyedik centrális momentum / szórás^4 -3 módon definiált mennyiség becslése.

Ha a lapultság pozitív, akkor a sokaság eloszlásának sûrûségfüggvénye csúcsosabb, mint a normális eloszlás haranggörbéjéé, ha negatív, akkor laposabb, ha 0, akkor egyezõ.

Az 3.1 példában szereplõ adatok lapultsága : - 0.3232

3.2.4 Megjegyzések a középértékrõl és a szórásról.

a) A középérték lineáris funkcionál:

$$E(\alpha X + \beta Y) = \alpha E(X) + \beta E(Y)$$

b) Néhány fontos tétel a szórásról és a szórásnégyzetrõl (varianciáról):

$D^2(X \pm Y) = D^2(X) + D^2(Y) = \sigma^2_x + \sigma^2_y$	$D(X+Y) = (\sigma^2_x + \sigma^2_y)^{1/2}$
$D^2(\alpha X) = \alpha^2 D^2(X)$	$D(\alpha X) = \alpha D(X)$
$D^2(X \pm \alpha) = D^2(X)$	$D(X \pm \alpha) = D(X)$

Fentiekbõl következik:

A középérték szórásának becslése

$$s_m = \frac{s}{\sqrt{n}} = \sqrt{\frac{\sum\left(x_i - \bar{x}\right)^2}{n(n-1)}} \qquad (3.15)$$

Levezetés:

$$D^2(\bar{x}) = D^2\left(\sum x / n\right) = \frac{1}{n^2} D^2\left(\sum x\right) = \frac{1}{n^2} \sum D^2(x) = \frac{\sum s_x^2}{n^2} = \frac{n s_x^2}{n^2} = \frac{s_x^2}{n}$$

Fontos összefüggés:

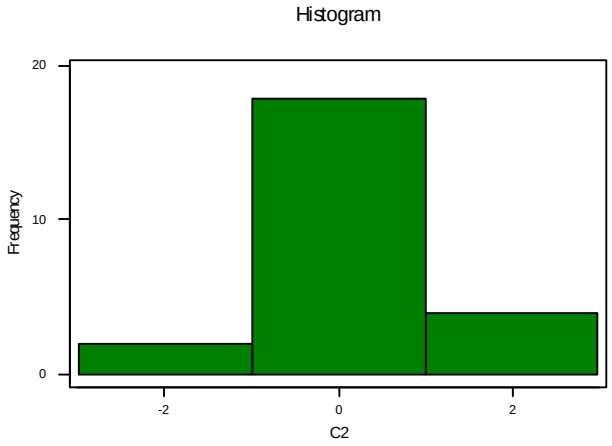
$$\sum (x_i - \bar{x})^2 = \sum x_i^2 - \bar{x} \sum x_i = \sum x_i^2 - (\sum x_i)^2 / n = \sum x_i^2 - n\bar{x}^2$$

$$n\sum\left(x_i - \bar{x}\right)^2 = n\sum x_i^2 - (\sum x_i)^2$$

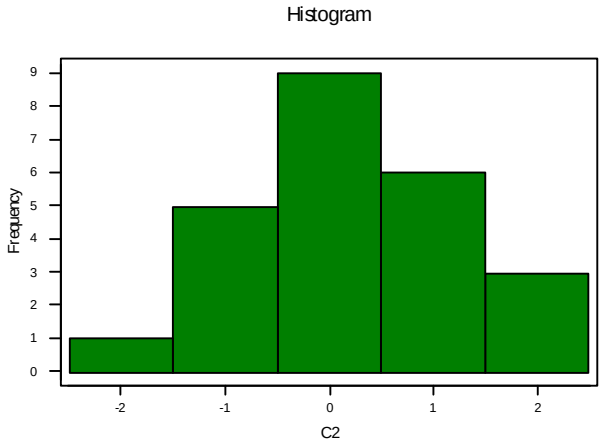
3.2.5 A minta eloszlásának grafikus szemléltetése

a) *a hisztogram*

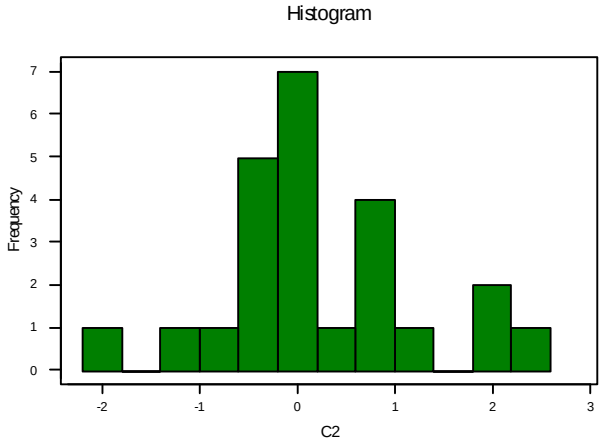
A hisztogram egy *rendezett* minta elõre kitûzött *változó-tartományaiba* esõ elemek *számát* vagy *gyakoriságát* ábrázolja. A hisztogram hasábjainak szélessége a változó-tartományt, magassága az (abszolut vagy relatív) gyakoriságot ábrázolja. Túl kevés tartomány kitûzésekor az információ szegényes (3.2/a ábra), túl sok esetén a kapott kép áttekinthetelen. (3.2/c ábra)



3.2/a ábra Elnagyolt hisztogram



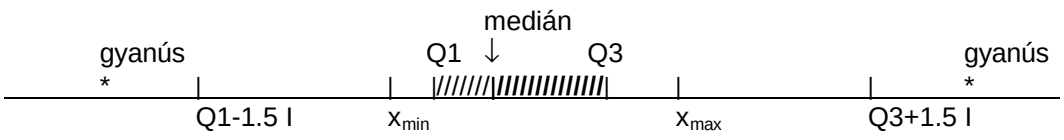
3.2/b ábra Jól méretezett hisztogram



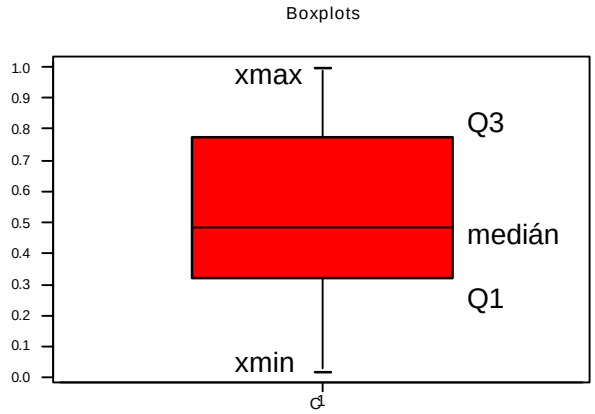
3.2/c ábra Túlrészletezett hisztogram

b) *a “box" vagy "szakállas" (box and whiskers) ábra*

A box vagy “szakállas (box and whiskers) ábra az eloszlás szemléltetésének célszerû módja, amely a változó számegyenesén különbözõ, jellemzõ kritikus pontokat tartalmaz:

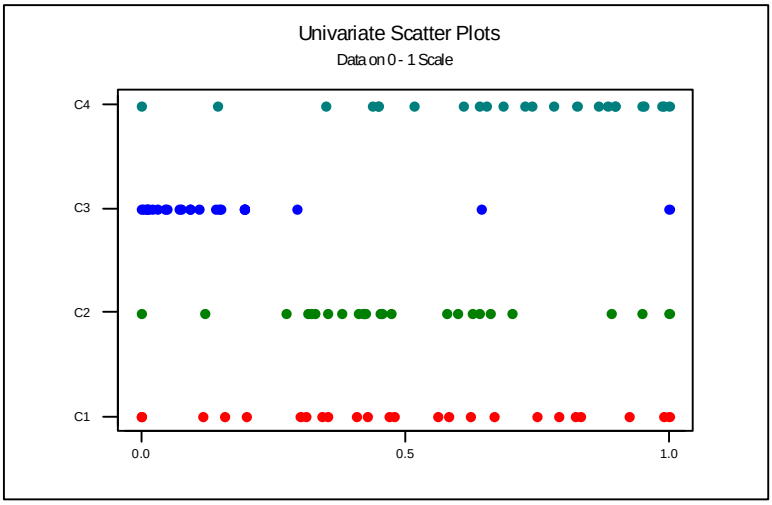


(I = Q3 - Q1 = interkvartilis távolság)

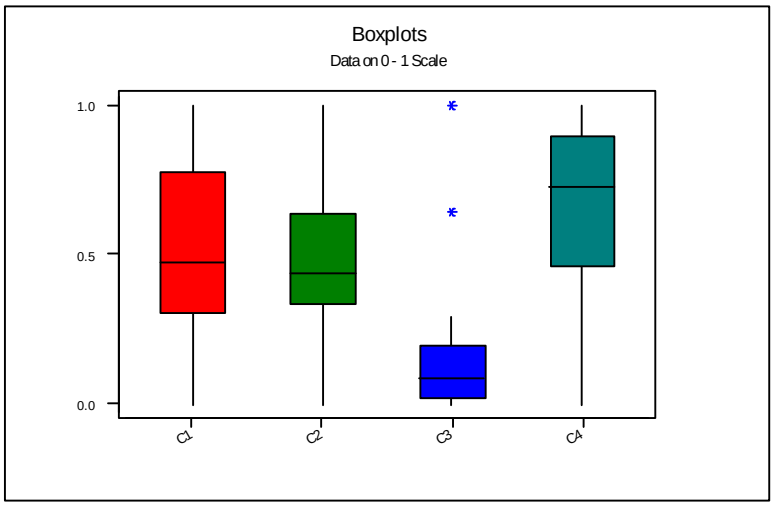


3.3 ábra A box ábra

2.1 példa. Négy mintát hasonlítunk össze. C4 és C3 adatok aszimmetrikus eloszlású sokaságokból származnak, C2 normális és C1 egyenletes eloszlásúak. A pontsor ábrák az alábbiak:



3.4 ábra A 2.1 példa mintáinak pontsor ábrái



3.5 ábra. A 2.1 példa mintáinak box ábrái