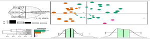
2. rész

A statisztika alapfogalmai

Az információtömörítés módszertana

Komputerstatisztika kurzus

Barczy Mátyás és Ispány Márton 2010
Informatikai Kar
Debreceni Egyetem



A 2. rész témái

1. **Statisztikai elemzés**
2. **Statisztikai minta**
3. **Leíró statisztikák**
4. **Grafikus elemzés**
5. **Függelék**

A statisztikai elemzés

A statisztikai elemzéshez ismerni kell:

- az elemzés tárgyát, a vizsgálandó objektumokat,
- az elemzés szempontjait, mely tulajdonságokat (jellemzőket) vizsgálunk,
- az elemzés módszerét, milyen statisztikai eljárást használunk.

Példa (tárgy és szempont)

- Debrecen város lakosságának testsúlya.
- A magyar népesség IQ–ja.
- Egy bank ügyfeleinek viselkedése.
- Egy mobiltelefon–hálózat kihasználtsága.
- Informatikai eszközök (processzorok, tároló kapacitások) időbeli fejlődése.

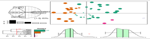
Az elemzés tárgya: a statisztikai sokaság

A statisztikai elemzés tárgyával szembeni legfontosabb követelmény, hogy elég nagy legyen ahhoz, hogy statisztikus (véletlenszerű) viselkedést mutasson.

- Egy teljes körűen megfigyelhető sokaságot tökéletesen jellemezhetünk egyedeinek teljes felsorolásával, pl. egy gyakorlati csoport.
- Nem teljes körűen megfigyelhető sokaságnál, pl. egy város lakossága, bank ügyfelei, ez már nem vagy csak extrém költségek mellett tehető meg.

Definíció (Statisztikai sokaság)

A statisztikai vizsgálat tárgyát képező objektumok (egyedek, egységek stb.) összességét statisztikai sokaságnak nevezzük. A statisztikai sokaságra használatos még a populáció elnevezés is.

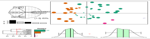


Statisztikai sokaság típusai:

- véges vagy végtelen,
- diszkrét vagy folytonos,
- valódi vagy fiktív.

Példa

- Debrecen város lakossága egy véges, diszkrét és valódi sokaság.
- Egy kísérlet sorozatban való megismétlése végtelen, fiktív sokaság.
- A havi csapadék mennyiségek valódi, folytonos sokaság.



Az elemzés szempontja: statisztikai változó

A vizsgálandó objektumok sokféle tulajdonságát elemezhetjük:

- súly, méret, életkor, szín, sebesség stb.

A vizsgált tulajdonság lehet kvantitatív (mennyiségi) vagy kvalitatív (minőségi), azonban lényeges, hogy numerikusan kódolható legyen. Másik fontos szempont, hogy a vizsgált jellemző statisztikus viselkedést mutasson.

Példa

Emberi populációnál

- statisztikus jellemzők: nem, foglalkozás (egy véges halmazból választva), életkor, IQ;
- nem statisztikus jellemzők: személyi szám (minden ilyen típusú azonosító), lakcím.

Definíció (Statisztikai változó)

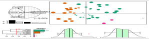
A statisztikai sokaság elemeinek egy tulajdonságát numerikusan jellemző mennyiséget statisztikai változónak nevezzük.

Statisztikai változók osztályozása:

- Értékkészlet alapján: diszkrét illetve folytonos.
- Jellemzés alapján: kvantitatív (mennyiségi) illetve kvalitatív (minőségi).

Jelölések:

- statisztikai sokaságot felsoroljuk: $\{1, \ldots, N\}$ amennyiben véges; $\mathbb{N} = \{1, 2, \ldots\}$ amennyiben végtelen,
- statisztikai változók: $X, Y, Z, \ldots$,
- $X(i)$ az X statisztikai változó értéke az i–edik objektum esetén.



Mérési (szintek) skálák (Stevens, 1946)

Mérés alatt számoknak eseményekhez, dolgokhoz való hozzárendelését értjük valamilyen szabály alapján.

Skálatípusok:

- névleges (nominális),
- sorrendi (ordinális),
- különbségi (intervallum),
- arány (hányados, ratio).

Tekintsünk két objektumot, A-t és B-t. Legyen X egy statisztikai változó, értékei az A, ill. B objektum esetén $X(A)$, ill. $X(B)$.

Névleges skála: kizárólag az egyedekhez rendelt számértékek egyező vagy különböző voltát engedi meg az egyedeket ténylegesen is jellemző tulajdonságként elfogadni. Vagyis A-ról és B-ről csak annyit tudunk mondani, hogy $X(A) = X(B)$ vagy $X(A) \neq X(B)$.

Példa

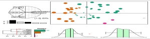
Az emberek neme névleges, kvalitatív, diszkrét változó. Pl. férfi = 1, nő = 2 vagy férfi = 101, nő = 102, az értékek között nincs semmiféle kapcsolat, csak az számít, hogy különbözőek.

Sorrendi skála: Nemcsak a skálaértékek azonos vagy nem azonos volta, hanem azok sorrendisége is értelmezhető a vizsgált egyedekre. Ebben az esetben azon kívül, hogy $X(A) = X(B)$ vagy $X(A) \neq X(B)$, azt is tudjuk mondani, hogy $X(A) > X(B)$ vagy $X(A) < X(B)$.

A skálaértékek bármilyen, az egyedek adott sorrendjét megtartó számértékek lehetnek, hiszen maguk a számértékek nem hordoznak információt, csak azok sorrendje.

Példa

Az emberek iskolai végzettsége sorrendi kvalitatív, diszkrét változó.



Különbségi skála: a skálaértékek **különbségei** is valós információt nyújtanak a sokaság egyedeiről (egymáshoz való viszonyukról). A korábbiakon túl azt is mondhatjuk, hogy A a B-től $X(A) - X(B)$ egységgel különbözik.

Az előző két skálával ellentétben e skálának általában már valamilyen mértékegység is szerves tartozéka. A skála kezdőpontjának – a 0 pontnak – a megválasztása önkényes, valamilyen konvención alapszik.

Példa

A hőmérséklet különbségi, kvantitatív, folytonos változó (Fahrenheit (f)–, Celsius (c)–skála: $f = 9c/5 + 32$).

Arány skála: a kezdőpont is egyértelműen adott, rögzített, a skálaértékek egymáshoz viszonyított aránya is hordoz információt. Ha $X(A) > X(B)$, akkor a korábbiakon túl azt is mondhatjuk, hogy A $X(A)/X(B)$-szer nagyobb, mint B.

Példa

A testsúly arány skálán mért kvantitatív, folytonos változó. A Kelvin-féle hőmérsékleti skálán (melynek 0 pontja az abszolút zérus hőmérséklet) mért hőmérséklet kvantitatív, folytonos változó.

Ez az osztályozás hierarchikus felépítésű, azaz minden skála rendelkezik az őt megelőző skála tulajdonságaival.

Megjegyzés

Ha c_1 és c_2 Celsius–skálán mért hőmérsékletek, f_1, illetve f_2 a nekik megfelelő Fahrenheit–skálán mért hőmérsékletek, akkor

$$c_2 - c_1 = \frac{5}{9} f_2 - 32 - \left(\frac{5}{9} f_1 - 32 \right) = \frac{5}{9} (f_2 - f_1),$$

$$\frac{c_2}{c_1} = \frac{\frac{5}{9} f_2 - 32}{\frac{5}{9} f_1 - 32}.$$

Így ha $c_2 - c_1$ ismert, akkor $f_2 - f_1$ egyértelműen meghatározott, míg ha c_2/c_1 ismert, akkor f_2/f_1 még nem egyértelműen meghatározott.

Ha pl. $c_1 = 10°C$ és $c_2 = 30°C$, akkor $f_1 = 50°F$, $f_2 = 86°F$, $c_2 - c_1 = 20°C$, $f_2 - f_1 = 36°F$, $c_2/c_1 = 3$ és $f_2/f_1 = 1.72$. Azt nem mondhatjuk, hogy a $30°C$-os hőmérséklet háromszor olyan magas, mint a $10°C$-os hőmérséklet, mivel a Fahrenheit–skálán ugyanennek a két hőmérsékletnek az aránya 1.72.

Azt viszont mondhatjuk, hogy a $30°C$ távolsága a $0°C$-tól háromszor akkora, mint a $10°C$-é a $0°C$-tól. Ez igaz Fahrenheit–skálán is, ugyanis $86°F - 32°F = 3(50°F - 32°F)$.

A Kelvin-féle hőmérsékleti skálán mondhatjuk azt, hogy a $30°K$ háromszor akkora, mint a $10°K$.

A statisztika módszerei

Alapvető módszerek:

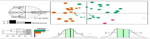
- Mintavétel: hogyan juthatunk a vizsgált statisztikai sokaságot jellemző mintához.
- Becslések: ismeretlen paraméterek, függvények meghatározása.
- Hipotézisek: döntés a minta eloszlására vonatkozó állításokról.

Feladattípusok:

- Egy változó vizsgálata.
- Két változó összehasonlítása, kapcsolatuk vizsgálata.
- Többdimenziós adatállományok elemzése.

A statisztika fejezetei:

- Paraméterbecslés,
- Statisztikai próbák (szórásanalízis),
- Regresszió és korreláció analízis,
- Minőségbiztosítás.



A statisztikai minta (első definíció)

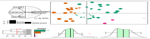
Mivel a statisztikai sokaságot teljes körűen nem tudjuk megfigyelni, ezért annak egy részét vesszük abban a reményben, hogy annak megfigyeléséből vissza tudunk következtetni a teljes sokaságra.

Definíció (Megfigyelt minta)

Sokasági minta alatt a statisztikai sokaság egy n elemű részhalmazát értjük, ahol véges sokaság esetén $1 \leq n \leq N$, végtelen sokaság esetén pedig $n \in \mathbb{N}$. Az X statisztikai változóra vett n elemű minta (megfigyelt minta) alatt a sokasági minta elemeinek az X változó szerinti értékeit értjük.

A megfigyelt minta n darab valós szám (vagy vektor)!
Jelölések:

- sokasági minta: $\{i_1, i_2, \ldots, i_n\} \subset \{1, 2, \ldots, N\}$.
  (Pl. Debrecen lakosai közül 200 ember.)
- megfigyelt minta: $x_1 = X(i_1), \ldots, x_n = X(i_n)$, röviden $x_1, x_2, \ldots, x_n$. (Pl. a kiválasztott 200 ember súlya.)



Mintatér és rendezett minta

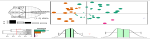
Az összes lehetséges n elemű minta az n–dimenziós euklideszi tér ($\mathbb{R}^n$) egy $\mathbb{X}^n$ részhalmazát alkotja, amelyet mintatérnek nevezünk.

Példa (Különböző mintaterek)

- $\{0, 1\}^n$: a vizsgált emberek neme,
- $\mathbb{N}^n$: a gyerekek száma a családban, ahol $\mathbb{N}$ a nemnegatív egész számok halmazát jelöli,
- $(0, \infty)^n$: a vizsgált emberek testsúlya,
- $\mathbb{R}^n$: napi átlaghőmérséklet.

Definíció (Rendezett minta)

Az $x_1, \ldots, x_n$ n elemű minta $x_1^* \leq \ldots \leq x_n^*$ nagyság szerint növekvő sorrendbe rendezését rendezett mintának nevezzük.



A statisztika fogalma

Statisztika alatt a minta tömörítését értjük, célunk az, hogy a mintában lévő információ minél nagyobb részét magyarázzuk a mintanagysághoz képest kevés számú mutatószámmal.

Definíció (Statisztika)

Legyen $T : \mathbb{X}^n \to \mathbb{R}^k$ egy függvény. Ekkor a $T(x_1, \ldots, x_n)$ vektort az $x_1, \ldots, x_n$ minta egy statisztikájának nevezzük. Magára a T leképezésre szintén használjuk a statisztika elnevezést.

A definícióban T–ről szinte semmit nem követeltünk meg, szintén nem követeltük meg, de a gyakorlatban k jóval kisebb, mint n.

Példa (Egyszerű statisztikák)

- $\overline{x} := \frac{1}{n} \sum_{i=1}^{n} x_i$ mintaátlag,
- $(\min, \max) := (x_1^*, x_n^*)$ a legkisebb és a legnagyobb mintaelem.

A tapasztalati eloszlásfüggvény (első definíció)

A mintából kinyerhető összes információt tartalmazza.

Definíció (Tapasztalati eloszlásfüggvény)

Az $x_1, \ldots, x_n$ n elemű minta $F_n^* : \mathbb{R} \to [0,1]$ tapasztalati eloszlásfüggvénye:

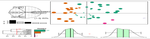
$$F_n^*(x) := \begin{cases} 0 & \text{ha } x \leq x_1^*, \\ k/n & \text{ha } x_k^* < x \leq x_{k+1}^*, \qquad 1 \leq k < n, \\ 1 & \text{ha } x > x_n^*. \end{cases}$$

Ha valamilyen k–ra $x_k^* = x_{k+1}^*$, akkor $(x_k^*, x_{k+1}^*] = \emptyset$. Ilyenkor F_n^* az x_k^* pontban (legalább) $2/n$–et ugrik.

A tapasztalati eloszlásfüggvény egy másik előállítása:

$$F_n^*(x) = \frac{1}{n}\sum_{k=1}^{n} I(x_k < x), \quad x \in \mathbb{R},$$

ahol I az indikátorfüggvényt jelöli. A tapasztalati eloszlásfüggvény annak a diszkrét valószínűségi változónak az eloszlásfüggvénye, amely minden egyes mintaelemet $1/n$ valószínűséggel vesz fel (egybeeső mintaelemek esetén $1/n$ megfelelő egész számú többszörösét).



Példa

Tekintsük a $3, 2, 2, 4, 1$ megfigyelt mintát. Ekkor a rendezett minta $1, 2, 2, 3, 4$. A tapasztalati eloszlásfüggvény:

$$F_5^*(x) = \begin{cases} 0 & \text{ha } x \leq 1, \\ 0.2 & \text{ha } 1 < x \leq 2, \\ 0.6 & \text{ha } 2 < x \leq 3, \\ 0.8 & \text{ha } 3 < x \leq 4, \\ 1 & \text{ha } x > 4. \end{cases}$$

A mintavétel módszerei

A minta osztályozása nagyság szerint:

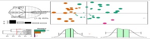
- kis minta: 50 alatt,
- átlagos mintanagyság: 50–1000,
- nagy minta: 1000 felett (adatbányászat).

A mintavétel alapproblémája:

gazdaságosság <–> pontosság.

Módszerek:

- egyszerű véletlen (visszatevés nélküli, EV–minta),
- független azonos eloszlású (visszatevéses, FAE–minta),
- szekvenciális (sorozatos/folyamatos) mintavétel. Bevezetése Wald Ábrahám (1902 – 1950) nevéhez fűződik. Pl. veszünk egy 100 elemű mintát valamilyen hipotézis eldöntése céljából, de ha kiderül, hogy a vett minta alapján nem tudunk dönteni, akkor újabb mintát veszünk,



- rétegzett (egyenletes, arányos, Neumann–féle optimális). Pl. a férfiak és nők aránya adott a társadalomban és azt szeretnénk, hogy a mintában is ez az arány legyen: arányos mintavétel.
  Az ún. reprezentatív mintavétel 4–féle módon rétegzett minta: a sokaság életkor, jövedelem, iskolai végzettség és lakóhely szerinti megoszlása tükröződik vissza a mintában,
- csoportos. A sokaságot csoportokra osztjuk és egyszerű véletlen mintavételezéssel kiválasztunk néhányat a csoportok közül majd a kiválasztott csoportokban mindenkit megfigyelünk,
- két– és többfokozatú (két– és többlépcsős).
  A kétlépcsős eljárás első lépcsője egy csoportos mintavétel, majd – ellentétben a csoportos mintavétellel – a kiválasztott csoportok elemeiből is mintát veszünk (valamilyen módszer szerint), azaz nem figyeljük meg teljeskörűen a kiválasztott csoportokat.

A gyorsrendezés algoritmusa (quick sort)

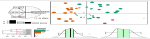
Az oszd meg és uralkodj elv alkalmazása.

$$A := \text{a mintaelemeket tartalmazó tömb,}$$
$$p := \text{az első elem indexe,}$$
$$r := \text{az utolsó elem indexe.}$$

Felosztás: Az $A[p..r]$ tömböt két, nemüres $A[p..q]$ és $A[q+1..r]$ résztömbre osztjuk úgy, hogy az $A[p..q]$ minden eleme kisebb vagy egyenlő $A[q+1..r]$ minden eleménél. A q index kiszámítása része ennek a felosztó eljárásnak.

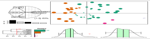
Uralkodás: Az $A[p..q]$ és $A[q+1..r]$ résztömböket a gyorsrendezés rekurzív hívásával rendezzük.

Egyesítés: Mivel a két résztömböt helyben rendeztük, nincs szükség egyesítésre: az $A[p..r]$ tömb rendezett.



Pszeudokód

```
RENDEZ(A,p,r)
   if p<r
     then q <- FELOSZT(A,p,r)
   RENDEZ(A,p,q)
   RENDEZ(A,q+1,r)
FELOSZT(A,p,r)
   x <- A[p]
   i <- p-1
   j <- r+1
   while i<j
       do repeat j <- j-1
            until A[j] <= x
          repeat i <- i+1
            until A[i] >= x
          if i < j
            then A[i] <-> A[j]
            else return j
```

Megjegyzés

A

```
repeat utasítás
       until feltétel
```

vezérlési szerkezet jelentése: egyszer mindenképpen végrehajtódik az `utasítás`, utána ellenőrződik a `feltétel`. Ha a `feltétel` igaz, akkor kilépünk, ha nem, akkor újra végrehajtódik az `utasítás`, majd a `feltétel` ellenőrzésre kerül, stb. Az `utasítás` addig ismétlődik, amíg a `feltétel` nem lesz igaz, amikor igaz, akkor kilépünk. Felhívjuk a figyelmet, hogy az `utasítás` legalább egyszer végrehajtódik (akkor is, ha a `feltétel` az első alkalommal nem igaz). Ilyen vezérlési szerkezet van a Pascal–ban, de C–ben nincs ilyen.

Megjegyzés

A `FELOSZT(A,p,r)` tulajdonképpen egyetlen műveletet végez: az x-nél kisebb elemeket a tömb elejére, míg az x-nél nagyobb elemeket a tömb végére helyezi. Pontosabban: az $A[q]$ elem és az előtte levők nem nagyobbak, mint x, az $A[q]$ utáni elemek pedig nem kisebbek, mint x.

1. Példa, 1. rész

Minta: A = [5, 1, 3, 6, 6, 7, 2, 5, 1].

Ekkor $p = 1$ és $r = 9$.

FELOSZT(A,1,9) működése:

$x = 5, \; i = 0, j = 10,$

$j = 9, A[9] = 1 < 5,$

$i = 1, A[1] = 5 >= 5,$

$i = 1 < j = 9 \implies A[1] \leftrightarrow A[9], \quad [1,1,3,6,6,7,2,5,5],$

$j = 8, A[8] = 5 <= 5,$

$i = 2, A[2] = 1 < 5,$

$i = 3, A[3] = 3 < 5,$

$i = 4, A[4] = 6 > 5,$

$i = 4 < j = 8 \implies A[4] \leftrightarrow A[8], \quad [1,1,3,5,6,7,2,6,5],$

$j = 7, A[7] = 2 < 5,$

$i = 5, A[5] = 6 > 5,$

$i = 5 < j = 7 \implies A[5] \leftrightarrow A[7], \quad [1,1,3,5,2,7,6,6,5].$

1. Példa, 2. rész

$j = 6, A[6] = 7 > 5,$
$j = 5, A[5] = 2 < 5,$
$i = 6, A[6] = 7 > 5,$
$i = 6 > j = 5 \implies \text{ return } \; j = 5.$

Tehát FELOSZT(A,1,9)=5, így
RENDEZ(A,1,5) és RENDEZ(A,6,9) kerül végrehajtásra.

FELOSZT(A,1,5) működése: A=[1,1,3,5,2,7,6,6,5].

$x = 1, \; i = 0, j = 6,$
$j = 5, A[5] = 2 > 1,$
$j = 4, A[4] = 5 > 1,$
$j = 3, A[3] = 3 > 1,$
$j = 2, A[2] = 1 <= 1,$
$i = 1, A[1] = 1 >= 1,$
$i = 1 < j = 2 \implies A[1] \leftrightarrow A[2], \quad$ [1,1,3,5,2,7,6,6,5].

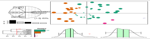
1. Példa, 3. rész

$j = 1, A[1] = 1 <= 1,$
$i = 2, A[2] = 1 >= 1,$
$i = 2 > j = 1 \implies \text{return } j = 1.$

Tehát FELOSZT(A,1,5)=1, így
RENDEZ(A,1,1) és RENDEZ(A,2,5) kerül végrehajtásra.

FELOSZT(A,6,9) működése: A=[1,1,3,5,2,7,6,6,5].
$x = 7, \ i = 5, j = 10,$
$j = 9, A[9] = 5 < 7,$
$i = 6, A[6] = 7 >= 7,$
$i = 6 < j = 9 \implies A[6] \leftrightarrow A[9],$ [1,1,3,5,2,5,6,6,7],
$j = 8, A[8] = 6 < 7,$
$i = 7, A[7] = 6 < 7,$
$i = 8, A[8] = 6 < 7,$
$i = 9, A[9] = 7 >= 7,$
$i = 9 > j = 8 \implies \text{return } j = 8.$

Tehát FELOSZT(A,6,9)=8, így
RENDEZ(A,6,8) és RENDEZ(A,9,9) kerül végrehajtásra.



1. Példa, 4. rész

FELOSZT(A,2,5) működése: A=[1,1,3,5,2,5,6,6,7].

$x = 1,\ \ i = 1, j = 6,$

$j = 5, A[5] = 2 > 1,$

$j = 4, A[4] = 5 > 1,$

$j = 3, A[3] = 3 > 1,$

$j = 2, A[2] = 1 <= 1,$

$i = 2, A[2] = 1 >= 1,$

$i = 2 = j = 2 \Longrightarrow$ return $j = 2.$

Tehát FELOSZT(A,2,5)=2, így
RENDEZ(A,2,2) és RENDEZ(A,3,5) kerül végrehajtásra.

Átgondolható, hogy mivel az $5, 6, 6, 7$ minta már rendezett, FELOSZT(A,6,8)=6, így
RENDEZ(A,6,6) és RENDEZ(A,7,8) kerül végrehajtásra.
Hasonlóan, FELOSZT(A,7,8)=7, így
RENDEZ(A,7,7) és RENDEZ(A,8,8) kerül végrehajtásra.

1. Példa, 5. rész

FELOSZT(A,3,5) működése: A=[1,1,3,5,2,5,6,6,7].

$x = 3,\ \ i = 2, j = 6,$

$j = 5, A[5] = 2 < 3,$

$i = 3, A[3] = 3 >= 3,$

$i = 3 < j = 5 \implies A[3] \leftrightarrow A[5], \quad [1,1,2,5,3,5,6,6,7],$

$j = 4, A[4] = 5 > 3,$

$j = 3, A[3] = 2 < 3,$

$i = 4, A[4] = 5 > 3,$

$i = 4 > j = 3 \implies \text{return } \ j = 3.$

Tehát FELOSZT(A,3,5)=3, így
RENDEZ(A,3,3) és RENDEZ(A,4,5) kerül végrehajtásra.

1. Példa, 6. rész

FELOSZT(A,4,5) működése: A=[1,1,2,5,3,5,6,6,7].

$x = 5,\ i = 3, j = 6,$

$j = 5, A[5] = 3 < 5,$

$i = 4, A[4] = 5 >= 5,$

$i = 4 < j = 5 \implies A[4] \leftrightarrow A[5],$ [1,1,2,3,5,5,6,6,7],

$j = 4, A[4] = 3 < 5,$

$i = 5, A[5] = 5 >= 5,$

$i = 5 > j = 4 \implies$ return $j = 4$.

Tehát FELOSZT(A,4,5)=4, így
RENDEZ(A,4,4) és RENDEZ(A,5,5) kerül végrehajtásra.

A rendezett minta tehát: [1,1,2,3,5,5,6,6,7].

Módosított gyorsrendezés

A FELOSZT(A,p,r) algoritmus más egy kicsit.

Input: (A,p,r).

Output: olyan j index és olyan átrendezett A tömb, hogy $A[i] <= A[j]$, ha $i < j$, és $A[k] >= A[j]$, ha $k > j$.

MODFELOSZT(A,p,r) pszeudokódja:

```
pivot <- A[r]
left  <- p
right <- r-1
while left<=right do
   while (left<=right and A[left]<=pivot)
             do left <- left+1
   while (left<=right and A[right]>=pivot)
             do right <- right-1
   if (left<right)
       then exchange A[left] <-> A[right]
exchange A[r] <-> A[left]
return left
```

Megjegyzés

A MODFELOSZT(A,p,r) abban tér el a FELOSZT(A,p,r)-től, hogy

- a minta végéről választ pivot elemet (strázsát),
- a pivot elemet csak a legvégén cseréli be. Emiatt ugyanarra a mintára nem feltétlenül ugyanazt adja outputként, mint a FELOSZT(A,p,r).

Azonban, ha a RENDEZ(A,p,r)-t a MODFELOSZT(A,p,r)-el használjuk ugyanazt az output-ot kapjuk, mintha a FELOSZT(A,p,r)-el használjuk, azaz a rendezett mintát.

2. Példa, 1. rész

Minta: A = [6, 3, 7, 3, 2, 5, 7, 5].

Ekkor $p = 1$ és $r = 8$.

MODFELOSZT(A,1,8) működése:

pivot=5, left=1, right=7, A = [6, 3, 7, 3, 2, 5, 7, 5].

$1 < 7,\ A[1] = 6 > 5 \Rightarrow$ left=1, A = [6, 3, 7, 3, 2, 5, 7, 5],

$1 < 7,\ A[7] = 7 > 5 \Rightarrow$ right=6, A = [6, 3, 7, 3, 2, 5, 7, 5],

$1 < 6,\ A[6] = 5 = 5 \Rightarrow$ right=5, A = [6, 3, 7, 3, 2, 5, 7, 5],

$1 < 5,\ A[5] = 2 < 5 \Rightarrow$ right=5, A = [6, 3, 7, 3, 2, 5, 7, 5],

$1 < 5 \Rightarrow A[1] \leftrightarrow A[5] \Rightarrow$ A = [2, 3, 7, 3, 6, 5, 7, 5],

$1 < 5,\ A[1] = 2 < 5 \Rightarrow$ right=2, A = [2, 3, 7, 3, 6, 5, 7, 5],

2. Példa, 2. rész

$2 < 5$, $A[2] = 3 < 5 \Rightarrow$ left=3, A = [2, 3, 7, 3, 6, 5, 7, 5],
$3 < 5$, $A[3] = 7 > 5 \Rightarrow$ left=3, A = [2, 3, 7, 3, 6, 5, 7, 5],
$3 < 5$, $A[5] = 6 > 5 \Rightarrow$ right=4, A = [2, 3, 7, 3, 6, 5, 7, 5],
$3 < 4$, $A[4] = 3 < 5 \Rightarrow$ right=4, A = [2, 3, 7, 3, 6, 5, 7, 5],
$3 < 4$, $A[3] \leftrightarrow A[4] \Rightarrow$ A = [2, 3, 3, 7, 6, 5, 7, 5],
$3 < 4$, $A[3] = 3 < 5 \Rightarrow$ left=4, A = [2, 3, 3, 7=7, 6, 5, 7, 5],
$4 = 4$, $A[4] = 7 > 5 \Rightarrow$ left=4, A = [2, 3, 3, 7=7, 6, 5, 7, 5],
$4 = 4$, $A[4] = 7 > 5 \Rightarrow$ right=3, A = [2, 3, 3, 7, 6, 5, 7, 5],
$4 > 3 \;\Rightarrow$ right=3, A = [2, 3, 3, 7, 6, 5, 7, 5],
$4 > 3 \;\Rightarrow$ nincs csere és a „nagy” while ciklus folytatódik tovább,
$A[8] \leftrightarrow A[4]$, A = [2, 3, 3, 7, 6, 5, 7, 5],
return left=4.

Tehát MODFELOSZT(A,1,8) outputja:
$j = 4$ és A = [2, 3, 3, 7, 6, 5, 7, 5].

2. Példa, 3. rész

FELOSZT(A,1,8) működése:

$$x = 6,\ \ i = 0, j = 9,$$
$$j = 8, A[8] = 5 < 6,$$
$$i = 1, A[1] = 6 >= 6,$$
$$i = 1 < j = 8 \implies A[1] \leftrightarrow A[8], \quad [5,3,7,3,2,5,7,6],$$
$$j = 7, A[7] = 7 > 6,$$
$$j = 6, A[6] = 5 < 6,$$
$$i = 2, A[2] = 3 < 6,$$
$$i = 3, A[3] = 7 > 6,$$
$$i = 3 < j = 6 \implies A[3] \leftrightarrow A[6], \quad [5,3,5,3,2,7,7,6],$$
$$j = 5, A[5] = 2 < 6,$$
$$i = 4, A[4] = 3 < 6,$$
$$i = 5, A[5] = 2 < 6,$$
$$i = 6, A[6] = 7 > 6,$$
$$i = 6 > j = 5 \implies \text{return} \ \ j = 5.$$

Tehát FELOSZT(A,1,8) outputja:

$j = 5$ és A=[5,3,5,3,2,7,7,6].

Gyorsrendezés egy funkcionális nyelven

A Haskell funkcionális programozási nyelvben:

```
sort[] = []

sort(p:q) = sort [ y | y <- q, y < p]
            ++ [p]
            ++ sort [ y | y <- q, y >= p]
```

`[]` = üres lista,
++ = két lista összefűzése (sorozatok egymás után írása).

Ez az algoritmus listákon működik („véges sorozatok”), a korábbiak pedig tömbökön, ez a legfőbb különbség közöttük.

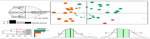
A kód olvasata:

- ha egy üres listát rendezünk, akkor üres listát kapunk,
- egy nem üres listát, aminek az első eleme `p`, az összes többi eleme pedig a `q` lista, a következőképpen kell rendezni:
  - egy listába összegyűjtjük a q-ban szereplő, p-től kisebb elemeket: `[ y | y <- q, y < p]`,
  - egy másik listába összegyűjtjük a q-ban szereplő, p-től nagyobb vagy egyenlő elemeket: `[ y | y <- q, y >= p]`,
  - ezt a két listát rekurzívan rendezzük: `sort [ y | y <- q, y < p]` és `sort [ y | y <- q, y >= p]`,
  - az eredmény a három lista összefűzése (ahol `[p]` a p-t tartalmazó egyelemű lista).

A gyorsrendezés véletlen változatai

1. verzió. Feltételezzük, hogy mielőtt a `RENDEZ(A,p,r)` megkezdené az átrendezést, véletlenszerűen átrendezi a bemeneti mintát, hogy ezzel biztosítsa a különböző permutációk azonos valószínűségét.

2/a. verzió. Feltételezzük, hogy rendelkezésünkre áll egy `RANDOM` nevű véletlenszám generátor. Ha $a, b \in \mathbb{Z}$, akkor a `RANDOM(a,b)` meghívásakor egy, az $[a, b]$ intervallumba eső egész számot kapunk vissza oly módon, hogy minden $[a, b]$-beli egész szám ugyanolyan valószínűséggel fordul elő. A `RENDEZ(A,p,r)` minden lépésénél, a tömb felosztása előtt kicseréljük az $A[p]$ elemet az A tömb egy véletlenszerűen választott elemével. Ez a módosítás biztosítja, hogy a strázsa elem, azaz $x = A[p]$, egyenlő eséllyel lehet a tömb bármelyik $r - p + 1$ eleme.



A `FELOSZT(A,p,r)` helyett:

```
RANDOM-FELOSZT(A,p,r)
   i <- RANDOM(p,r)
   A[p] <-> A[i]
   return FELOSZT(A,p,r),
```

illetve a `RENDEZ(A,p,r)` helyett:

```
RANDOM-RENDEZ(A,p,r)
  if p<r
   then q <- RANDOM-FELOSZT(A,p,r)
             RANDOM-RENDEZ(A,p,q)
             RANDOM-RENDEZ(A,q+1,r)
```

2/b. verzió. A Haskell-ben megírt gyorsrendezés véletlenített változata: minden egyes lépésben a `(p:q)` listában az első elemet (p-t) felcseréljük a lista egy véletlenszerűen választott elemével, majd az így kapott listát rendezzük a `sort` függvénnyel.

A véletlenített gyorsrendezés átlagos futási ideje

A Haskell-ben megírt gyorsrendezés véletlenített változatát (2/b. verzió) elemezzük.

Művelet: két mintaelem összehasonlítása.

Felhívjuk a figyelmet, hogy az elemzések során két tömbindex összehasonlítását nem szokás műveletnek hívni, mely azzal kapcsolatos, hogy a tömbbeli számok valós számok, míg a tömbindexek mindig egészek.

M_n : a műveletek átlagos száma n szám rendezésénél.

Jelölje a továbbiakban $X_1, \ldots, X_n$ az inputként kapott n számot.

A véletlenített algoritmus definíciójából adódóan tetsző-leges $i \in \{1, \ldots, n\}$ esetén annak a valószínűsé-ge, hogy az első csere után a rendezett minta i–edik tagja lesz a lista első eleme $1/n$, azaz

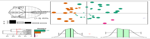
$$P(X_1 = X_i^*) = \frac{1}{n}, \qquad i = 1, \ldots, n.$$

Rekurzió M_n-re:

A teljes várható érték tétele alapján, ha ξ_n jelöli egy n elemű minta esetén a műveletek számát:

$$
\begin{aligned}
M_n = \mathbb{E}(\xi_n) &= \sum_{i=1}^{n} \mathbb{E}(\xi_n \mid X_1 = X_i^*) P(X_1 = X_i^*) \\
&= \sum_{i=1}^{n} (n - 1 + M_{i-1} + M_{n-i}) \frac{1}{n} \\
&= n - 1 + \frac{1}{n} \sum_{i=1}^{n} (M_{i-1} + M_{n-i}) \\
&= n - 1 + \frac{2}{n} \sum_{i=1}^{n} M_{i-1},
\end{aligned}
$$

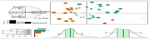
ahol a kezdő érték $M_0 := 0$.



Valóban, feltételezve, hogy az X_1 mintaelem az i–edik legnagyobb, $n-1$ db összehasonlítás (az n-hosszúságú lista első elemének a lista utolsó $n-1$ tagjával való összehasonlítása) szükséges az $(i-1)$-hosszúságú `[ y | y <- q, y < p]` és az $(n-i)$-hosszúságú `[ y | y <- q, y >= p]` listák előállításához, továbbá ezen listák rendezéséhez szükséges műveletek átlagos száma M_{i-1} és M_{n-i}.

Vegyük az $\{nM_n\}_{n=1}^{\infty}$ sorozat elsőrendű differenciáját:

$$
\begin{aligned}
& nM_n - (n-1)M_{n-1} \\
& = n(n-1) - (n-1)(n-2) + 2\left(\sum_{i=1}^{n} M_{i-1} - \sum_{i=1}^{n-1} M_{i-1}\right) \\
& = 2(n-1) + 2M_{n-1}, \qquad n \geq 1.
\end{aligned}
$$



Ez alapján kapjuk az

$$\frac{M_n}{n+1} = \frac{M_{n-1}}{n} + \frac{2(n-1)}{n(n+1)}, \qquad n \geq 1$$

rekurziót. Valóban,

$$\begin{aligned}\frac{M_n}{n+1} - \frac{M_{n-1}}{n} &= \frac{nM_n - (n+1)M_{n-1}}{n(n+1)} \\ &= \frac{nM_n - (n-1)M_{n-1} - 2M_{n-1}}{n(n+1)} \\ &= \frac{2(n-1) + 2M_{n-1} - 2M_{n-1}}{n(n+1)} \\ &= \frac{2(n-1)}{n(n+1)}, \qquad n \geq 1.\end{aligned}$$

Ezért

$$
\begin{aligned}
\frac{M_n}{n+1} &= 2\sum_{k=1}^{n} \frac{k-1}{k(k+1)} = 2\sum_{k=1}^{n}\left(\frac{2}{k+1} - \frac{1}{k}\right) \\
&= 4\left(\frac{1}{2} + \cdots + \frac{1}{n+1}\right) - 2\left(\frac{1}{1} + \cdots + \frac{1}{n}\right) \\
&= 4\left(\frac{1}{1} + \cdots + \frac{1}{n}\right) + \frac{4}{n+1} - 4 - 2\left(\frac{1}{1} + \cdots + \frac{1}{n}\right) \\
&= 2\sum_{k=1}^{n} \frac{1}{k} - \frac{4n}{n+1}.
\end{aligned}
$$

Mivel

$$\ln(n+1) < \sum_{k=1}^{n} \frac{1}{k} < 1 + \ln(n), \qquad n > 1,$$

kapjuk, hogy

$$2(n+1)\ln(n+1) - 4n < M_n = (n+1)2\sum_{k=1}^{n}\frac{1}{k} - 4n$$

$$< 2(n+1)(1+\ln(n)) - 4n,$$

$$\Downarrow$$

$$\left(1+\frac{1}{n}\right)\frac{\ln(n+1)}{\ln(n)} - \frac{2}{\ln(n)} < \frac{M_n}{2n\ln(n)}$$

$$< \left(1+\frac{1}{n}\right)\left(\frac{1}{\ln(n)}+1\right) - \frac{2}{\ln(n)}.$$

Mivel a $\mathcal{L}$'Hospital szabály alapján

$$\lim_{n\to\infty}\frac{\ln(n+1)}{\ln(n)} = \lim_{n\to\infty}\frac{1/(n+1)}{1/n} = 1,$$

$n \to \infty$ határátmenettel kapjuk, hogy

$$\lim_{n \to \infty} \frac{M_n}{2n \ln(n)} = 1.$$

Azaz nagy n-re

$$M_n = (n+1)2\sum_{k=1}^{n} \frac{1}{k} - 4n$$

$$\approx 2n\ln(n) = 2n\frac{\log_2(n)}{\log_2(e)} = 2n\ln(2)\log_2(n)$$

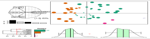
$$\approx 1.386 n \log_2(n).$$

Leíró statisztikák

Adatelemzés fajtái:

Feltáró (exploratív): célja az adatokkal való első ismerkedés, helyes kép kialakítása a mintáról. Paraméteres eloszláscsaládnál kiinduló becslés adása a paraméterekre.

Megerősítő (konfirmatív): van valamilyen előfeltételezésünk az adatokról és ezt akarjuk megerősíteni vagy megcáfolni.



A leíró statisztikákat általában a feltáró adatelemzés során használjuk.

Típusai:

- gyakorisági tábla,
- középértékek vagy helyzet (location) mutatók,
- szóródási vagy skála (scale) mutatók,
- alakmutatók (shape).

Az utóbbi három mutatóval számos paraméteres eloszláscsaládot jellemezhetünk.

Jól használhatóak:

- adattömörítésre (és később grafikus megjelenítésre),
- kilógó (irreguláris, outlier) értékek megállapítására (adatok tisztítása),
- szükséges változó és adat transzformációk meghatározására.

Az alábbiakban egy változót (egydimenziós eloszlást) jellemző mutatókkal foglalkozunk.

Gyakorisági tábla

Egy diszkrét változó értékeiről és eloszlásáról kaphatunk teljes képet általa. Alkalmazható folytonos változóra is diszkretizáció (ún. dobozolás) után.

A gyakorisági tábla az alábbi információkat tartalmazza:

- a változó neve és lehetséges értékei (címkével is megadva),
- a változó értékeinek gyakorisága, azaz előfordulásuk száma a mintában,
- a változó értékeinek százalékos aránya, azaz előfordulási valószínűsége a mintában,
- a változó értékeinek felfelé kumulált (összegzett) gyakorisága,
- a változó értékeinek felfelé kumulált százalékos aránya.

Példa:

Egy társaságban 20 férfi és 30 nő van.

Nem	Gyako-riság	Százalékos gyakoriság	Kumulált gyakoriság	Kumulált százalékos gyakoriság
Férfi	20	40	20	40
Nő	30	60	50	100

Mutatók jellemzői

Jelöljön S egy statisztikát, legyen $\underline{x} := (x_1, \ldots, x_n)$ egy tetszőleges minta és $\underline{1} := (1, \ldots, 1)$.

Helyzetmutatók: $S(\underline{x} + c\underline{1}) = S(\underline{x}) + c, \forall\, c \in \mathbb{R}$ (eltolást mérnek).

Szóródási mutatók:

- $S(\underline{x} + c\underline{1}) = S(\underline{x}), \forall\, c \in \mathbb{R}$ (eltolás–invariancia),
- $S(d\underline{x}) = dS(\underline{x}), \forall\, d \in \mathbb{R}_+$ (skálát mérnek).

Alakmutatók: $S(d\underline{x} + c\underline{1}) = S(\underline{x}), \forall\, d \in \mathbb{R}_+, c \in \mathbb{R}$ (eltolás– és skála–invariancia, csak alakot jellemeznek).

Helyzetmutatók (középértékek)

A. Kvantilisek

Definíció (Elméleti kvantilis)

Legyen ξ egy valószínűségi változó, $\alpha \in (0,1)$. Egy $Q(\alpha)$ valós számot ξ α–kvantilisének nevezünk, ha

$$P(\xi \leq Q(\alpha)) \geq \alpha \quad \text{és} \quad P(\xi \geq Q(\alpha)) \geq 1-\alpha.$$

Ha F jelöli ξ eloszlásfüggvényét, akkor

$$F(Q(\alpha)) \leq \alpha \quad \text{és} \quad F(Q(\alpha)+0) \geq \alpha,$$

ahol

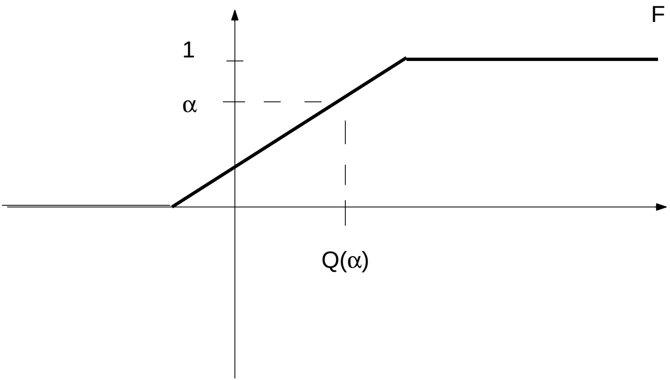
$$F(Q(\alpha)+0) := \lim_{x \downarrow Q(\alpha)} F(x).$$

1. eset: Ha az

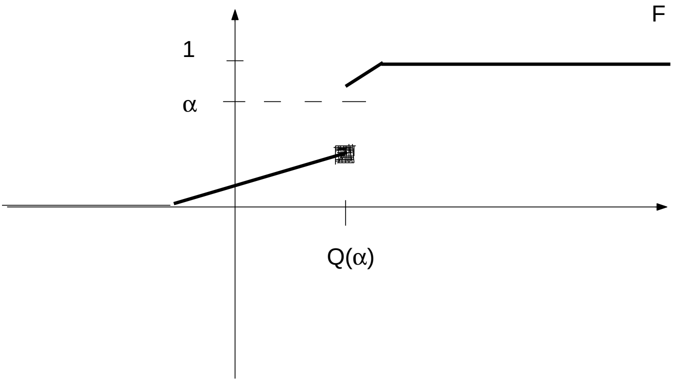
$$F(x) = \alpha$$

egyenletnek van megoldása, de csak egy, akkor ez az $F^{-1}(\alpha)$ megoldás az egyetlen α-kvantilis.

Speciálisan, ha ξ–nek az F eloszlásfüggvénye folytonos és szigorúan monoton növekvő (pl. normális eloszlás), akkor a α–kvantilis a fenti egyenlet egyértelmű megoldása.



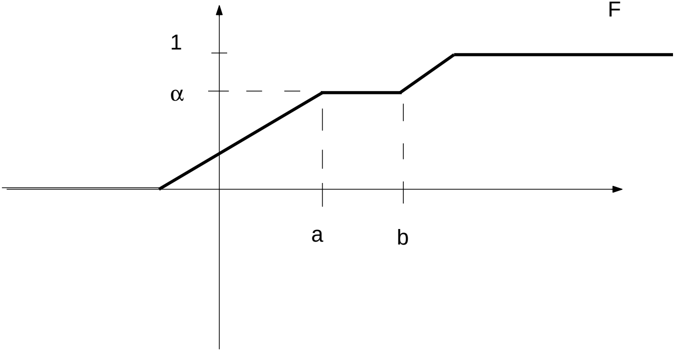
2. eset: Ha az $F(x) = \alpha$ egyenletnek nincs megoldása, akkor egyetlen α–kvantilis van, mégpedig az a szám, ahol az F függvény átugorja a α értéket.



3. eset: Ha az $F(x) = \alpha$ egyenletnek több megoldása van, akkor a megoldáshalmaz az $(a, b]$ vagy az $[a, b]$ intervallum, és a α–kvantilisek éppen az $[a, b]$ intervallum pontjai, ahol

$$a := \inf\left\{x \in \mathbb{R} : F(x+0) \geq \alpha\right\},$$

$$b := \sup\left\{x \in \mathbb{R} : F(x) \leq \alpha\right\}.$$



Megjegyzés

- A kvantilis nem feltétlenül egyértelmű (lásd 3. eset).
- Ha létezik az f sűrűségfüggvény, akkor

$$\int_{-\infty}^{Q(\alpha)} f(x)\,dx = \alpha \qquad \text{és} \qquad \int_{Q(\alpha)}^{\infty} f(x)\,dx = 1 - \alpha.$$

Elnevezések:

- $\alpha = 0.5$: medián,
- $\alpha = 0.25, 0.75$: alsó (Q_1) és felső (Q_3) kvartilis,
- $\alpha = 0.1, \ldots, 0.9$: decilisek ($D_1, \ldots, D_9$),
- $\alpha = 0.01, \ldots, 0.99$: percentilisek.
- $h = Q_3 - Q_1$: interkvartilis terjedelem.

Tapasztalati kvantilisek

Definíció (Tapasztalati kvantilis)

Legyen $x_1, \ldots, x_n$ n–elemű minta, $\alpha \in (0,1)$. Az

$$\widehat{Q}(\alpha) := x^*_{[\alpha(n+1)]} + \{\alpha(n+1)\}(x^*_{[\alpha(n+1)]+1} - x^*_{[\alpha(n+1)]})$$

valós számot tapasztalati α–kvantilisnek nevezzük, ahol $[\cdot]$ az egész, $\{\cdot\}$ a törtrészt jelöli, továbbá $x^*_0 := x^*_1$ és $x^*_{n+1} := x^*_n$.

Megjegyzés

- Ha $1 \leq [\alpha(n+1)] \leq n$, akkor nincs szükség az $x^*_0 := x^*_1$ és $x^*_{n+1} := x^*_n$ definíciókra.
- Vannak más definíciók is a tapasztalati kvantilisre: pl.

$$\widehat{Q}(\alpha) := x^*_{[\alpha n]} + \{\alpha n\}(x^*_{[\alpha n]+1} - x^*_{[\alpha n]}).$$

Tapasztalati medián

Példa (A medián számolása)

$$\widehat{Me} := \widehat{Q}(1/2) = \begin{cases} x^*_{k+1} & \text{ha } n = 2k+1, \\ (x^*_k + x^*_{k+1})/2 & \text{ha } n = 2k. \end{cases}$$

Valóban, ha $n = 2k+1$, akkor $\alpha(n+1) = 0.5(2k+2) = k+1$ és így $[\alpha(n+1)] = k+1$ és $\{\alpha(n+1)\} = 0$. Ezért $\widehat{Me} = x^*_{k+1}$.

Ha $n = 2k$, akkor $\alpha(n+1) = 0.5(2k+1) = k+0.5$ és így $[\alpha(n+1)] = k$ és $\{\alpha(n+1)\} = 0.5$. Ezért

$$\widehat{Me} = x^*_k + 0.5(x^*_{k+1} - x^*_k) = \frac{x^*_k + x^*_{k+1}}{2}.$$

B. Közepek

Definíció (Mintaátlag)

Az $x_1, \ldots, x_n$ minta mintaátlaga (számtani közepe) alatt az

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

valós számot értjük.

A mintaátlag iteratív számolása Spicer átmeneti átlagok algoritmusával (1972):

$$\overline{x}_n = \overline{x}_{n-1} + \frac{1}{n}(x_n - \overline{x}_{n-1}) = \frac{1}{n}x_n + \frac{n-1}{n}\overline{x}_{n-1}, \quad n > 1,$$

ahol a kezdőérték $\overline{x}_1 := x_1$.

Előnye: nincs „túlcsordulás”, abban az értelemben, hogy a számolás során nem lépünk ki az $[x_1^*, x_n^*]$ intervallumból.

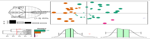
További közepek:

- geometriai: $\left(\prod_{i=1}^{n} x_i\right)^{1/n}$,
- harmonikus: $n\left(\sum_{i=1}^{n} x_i^{-1}\right)^{-1}$,
- négyzetes: $\sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2}$.

A közepek kapcsolata:

$$\text{harmonikus} \leq \text{geometriai} \leq \text{számtani} \leq \text{négyzetes}.$$

Felhívjuk a figyelmet, hogy a geometriai, harmonikus és négyzetes középre nem teljesül az a tulajdonság, hogy $S(\underline{x} + c\underline{1}) = S(\underline{x}) + c, \forall\, c \in \mathbb{R}$, azaz nem „mérik” az eltolást.



C. Módusz

Definíció

Az $x_1, \ldots, x_n$ minta módusza alatt a leggyakrabban előforduló értéket értjük. Ha több ilyen is van, akkor azok közül a legkisebbet értjük móduszon.

Ha pl. a mintaelemek páronként különbözőek, akkor a módusz a legkisebb mintaelem: x_1^*.

Szóródási (terjedelem) mutatók

A. Terjedelemek

Azt mérik, hogy a minta mennyire koncentrálódik egy bizonyos tartományra.

Mintaterjedelem:

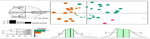
$$\widehat{R} := x_n^* - x_1^*.$$

Interkvartilis terjedelem:

$$\widehat{h} := \widehat{Q}_3 - \widehat{Q}_1.$$

B. Mintaátlagtól való abszolút eltérés

$$\widehat{\delta} := \frac{1}{n}\sum_{i=1}^{n} |x_i - \overline{x}|.$$



C. Mintaátlagtól való négyzetes eltérés

Tapasztalati szórás:

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

Korrigált tapasztalati szórás:

$$s_n^* := \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}.$$

Kis mintáknál s_n^{*2} ad jobb becslést a szórásnégyzetre.

Steiner–formula

Definiáljuk egy tetszőleges $a \in \mathbb{R}$ szám körüli általánosított tapasztalati szórásnégyzetet (a-tól való négyzetes eltérés) az alábbi módon:

$$s_n^2(a) := \frac{1}{n} \sum_{i=1}^{n} (x_i - a)^2.$$

Állítás (A mintaátlag jellemzése)

Minden $a \in \mathbb{R}$ számra $s_n^2 = s_n^2(\overline{x}) \leq s_n^2(a)$, azaz az általánosított tapasztalati szórásnégyzet a mintaátlagban veszi fel a minimumát és ez a minimum a tapasztalati szórásnégyzet.

Steiner–formula

Minden $a \in \mathbb{R}$ számra

$$ns_n^2 = \sum_{i=1}^{n} (x_i - a)^2 - n(\overline{x} - a)^2.$$

Bizonyítás. A teve–szabály alapján:

$$
\begin{aligned}
ns_n^2 = \sum_{i=1}^{n} (x_i - \overline{x})^2 &= \sum_{i=1}^{n} ((x_i - a) + (a - \overline{x}))^2 \\
&= \sum_{i=1}^{n} (x_i - a)^2 + 2\sum_{i=1}^{n} (x_i - a)(a - \overline{x}) + \sum_{i=1}^{n} (a - \overline{x})^2.
\end{aligned}
$$

Mivel

$$
\begin{aligned}
2\sum_{i=1}^{n} (x_i - a)(a - \overline{x}) &= 2(a - \overline{x}) \sum_{i=1}^{n} (x_i - a) \\
&= 2(a - \overline{x})(n\overline{x} - na) = -2n(a - \overline{x})^2
\end{aligned}
$$

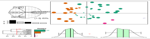
kapjuk a formulát. □

Az mintaátlagra vonatkozó állítás a Steiner–formula közvetlen következménye, hiszen minden $a \in \mathbb{R}$ esetén

$$s_n^2 = s_n^2(a) - (\overline{x} - a)^2.$$

Megj.: A Steiner–formulából $a = 0$ választással:

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - (\overline{x})^2.$$



D. Relatív szórás

$$\widehat{V} := \frac{s_n}{\overline{x}}.$$

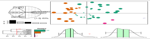
Felhívjuk a figyelmet, hogy a relatív szórásra nem teljesül az eltolás–invariancia és az a tulajdonság sem, hogy $S(d\underline{x}) = dS(\underline{x}), \forall\, d \in \mathbb{R}_+$.

E. Az átlag standard hibája

$$\frac{s_n^*}{\sqrt{n}},$$

mely azt méri, hogy mennyire ingadozhat a várható érték a mintaátlag körül.

Gondoljunk például normális eloszlás várható értékére vonatkozó konfidencia intervallum szerkesztésére ismeretlen szórásnégyzet esetén, lásd később.



F. Medián abszolút eltérés

Az

$$|x_1 - \widehat{Me}|, \ldots, |x_n - \widehat{Me}|$$

sorozat mediánja, azaz a mediántól való abszolút eltérések mediánja.

G. Átlagos abszolút eltérés a mediántól

$$\frac{1}{n}\sum_{i=1}^{n} |x_i - \widehat{Me}|.$$

Feladat

Adjuk meg a medián mintaátlaghoz hasonló jellemzését az átlagos abszolút eltérés segítségével.

H. Tapasztalati momentumok

Legyen $r \in \mathbb{N}$.

Tapasztalati r-edik momentum:

$$\widehat{M_r} := \frac{1}{n} \sum_{i=1}^{n} x_i^r .$$

Tapasztalati r-edik centrált momentum:

$$\widehat{M_r^c} := \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^r .$$

Felhívjuk a figyelmet, hogy az $\widehat{M_r}$ és $\widehat{M_r^c}$ mutatókra csak $r = 1$ esetén teljesül az $S(d\underline{x}) = dS(\underline{x}), \forall\, d \in \mathbb{R}_+$ tulajdonság. Általában az igaz, hogy $S(d\underline{x}) = d^r S(\underline{x}), \forall$ $d \in \mathbb{R}_+$. Az $(\widehat{M_r})^{1/r}$ és $(\widehat{M_r^c})^{1/r}$ mutatókra viszont már teljesül a $S(d\underline{x}) = dS(\underline{x}), \forall\, d \in \mathbb{R}_+$ tulajdonság.

Megj.: $\widehat{M_2^c} = s_n^2$.

I. Gini–mutató (átlagos abszolút különbség)

$$\widehat{G} := \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} |x_i - x_j|$$

Megj.: Azért $n(n-1) = n^2 - n$–el osztunk és nem n^2–tel, mert $i = j, i = 1, \ldots, n$ esetén az $x_i - x_j$ különbségek mindegyike 0.

Feladat

Mutassuk meg, hogy

- $$0 \leq \widehat{G} \leq \frac{2}{n-1} \sum_{i=1}^{n} |x_i|.$$
- ha $x_i \geq 0, i = 1, \ldots, n$, akkor
  $$0 \leq \widehat{G} \leq 2\overline{x}.$$

J. Lorenz–görbe, koncentrációs együttható

Akkor beszélünk koncentrációról, hogy ha a sokasághoz tartozó teljes értékösszeg $(\sum_{i=1}^{n} x_i)$ vagy annak jelentős része a sokaság kis részére összpontosul.

Legyen $x_1, \ldots, x_n$ egy olyan minta, melyre $x_i \geq 0$, $i = 1, \ldots, n$ és nem mindegyik mintaelem 0 (mely azzal ekvivalens, hogy $\overline{x} \neq 0$).

Koncentrációs együttható:

$$L := \frac{\widehat{G}}{2\overline{x}}.$$

Megj.: A korábbiak alapján $0 \leq L \leq 1$.

Állítás

$$L = \frac{2}{n(n-1)\overline{x}} \sum_{i=1}^{n} i x_i^* - \frac{n+1}{n-1}.$$

Lorenz–görbe: a $(0,0)$ és a

$$\left(\frac{k}{n}, \frac{\sum_{i=1}^{k} x_i^*}{\sum_{i=1}^{n} x_i^*}\right), \quad k = 1, \ldots, n,$$

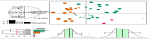
pontokat összekötő töröttvonal.

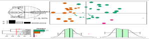
Állítás

- *a Lorenz-görbe monoton növekvő, konvex függvény.*
- $$L = \frac{n}{n-1} 2t_c,$$

  ahol t_c az ún. koncentrációs terület: a $(0,0)$ és $(1,1)$ végpontú szakasz és a Lorenz–görbe által határolt poligon területe.

Minél kisebb a koncentrációs terület, annál kisebb a koncentráció, minél nagyobb a koncentrációs terület, annál nagyobb a koncentráció.





Alakmutatók

Aszimmetria mutatók

A. Pearson–féle (1895)

$$\widehat{P} := \frac{3(\overline{x} - \widehat{Me})}{s_n}.$$

B. Ferdeség (Skewness)

Az X valószínűségi változó (elméleti) ferdesége:

$$S := \frac{E(X - E(X))^3}{D^3(X)}.$$

Az elméleti ferdeségre használatos még az α_3 jelölés is.

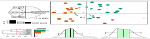
Normális eloszláscsalád, illetve szimmetrikus eloszlások esetén (mikor is X és $-X$ eloszlása megegyezik) $S = 0$.

Tapasztalati ferdeség:

$$\widehat{S_1} := \frac{\widehat{M_3^c}}{(\widehat{M_2^c})^{3/2}} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^3}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2\right)^{3/2}},$$

$$\widehat{S_2} := \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{s_n^*}\right)^3 .$$

Kis minták esetén $\widehat{S_2}$ jobban közelíti az elméleti ferdeséget, mint $\widehat{S_1}$.



C. Kvantilis-ferdeségek

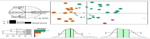
Legyen $\alpha \in (0,1)$.

$$\widehat{F}_\alpha := \frac{(\widehat{Q}(1-\alpha) - \widehat{Me}) - (\widehat{Me} - \widehat{Q}(\alpha))}{(\widehat{Q}(1-\alpha) - \widehat{Me}) + (\widehat{Me} - \widehat{Q}(\alpha))},$$

$$\widehat{F}_{0.25} := \frac{(\widehat{Q}_3 - \widehat{Me}) - (\widehat{Me} - \widehat{Q}_1)}{(\widehat{Q}_3 - \widehat{Me}) + (\widehat{Me} - \widehat{Q}_1)},$$

$$\widehat{F}_{0.1} := \frac{(\widehat{D}_9 - \widehat{Me}) - (\widehat{Me} - \widehat{D}_1)}{(\widehat{D}_9 - \widehat{Me}) + (\widehat{Me} - \widehat{D}_1)}.$$

Nyilván, $\widehat{F}_\alpha \in [-1,1]$.



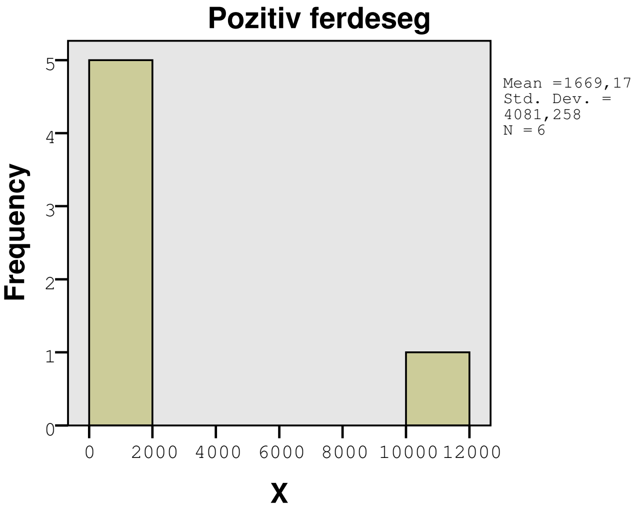
Aszimmetria mutatók értelmezése

Baloldali aszimmetriát, mikor is az eloszlás jobboldali farka vastagabb és az eloszlás nagyobb tömege koncentrálódik balra a mutatók pozitív értékei jelzik.

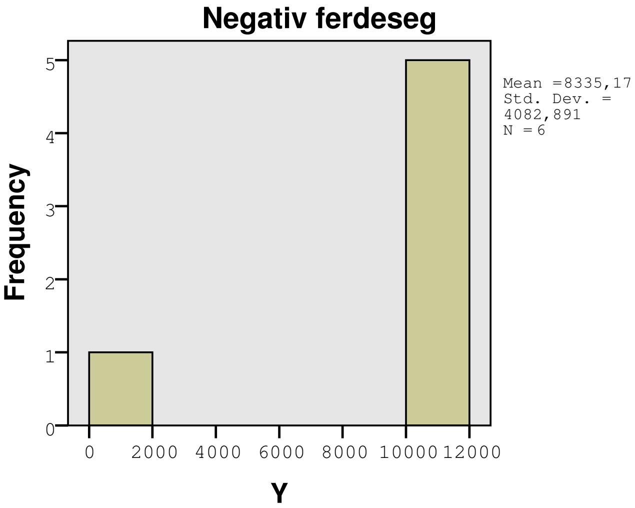
Például: az $X = (1, 2, 3, 4, 5, 1000)$ minta esetén $\widehat{S_2} \approx 2.449$.

Jobboldali aszimmetriát, mikor is az eloszlás baloldali farka vastagabb és az eloszlás nagyobb tömege koncentrálódik jobbra a mutatók negatív értékei jelzik.

Például: az $Y = (1, 10000, 10001, 10002, 10003, 10004)$ minta esetén $\widehat{S_2} \approx -2.449$.



Előállította az SPSS rendszer



Előállította az SPSS rendszer

Csúcsossági mutatók

A. Lapultság (Kurtosis)

Az X valószínűségi változó (elméleti) lapultsága:

$$K := \frac{E(X - E(X))^4}{D^4(X)} - 3.$$

Az elméleti lapultságra használatos még az α_4 jelölés is. Belátható, hogy $K \geq -2$, továbbá normális eloszlás-család esetén $K = 0$.

Tapasztalati lapultság:

$$\widehat{K_1} := \frac{\widehat{M_4^c}}{(\widehat{M_2^c})^2} - 3 = \frac{\frac{1}{n}\sum_{i=1}^n (x_i - \overline{x})^4}{\left(\frac{1}{n}\sum_{i=1}^n (x_i - \overline{x})^2\right)^2} - 3,$$

$$\widehat{K_2} := \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum_{i=1}^n \left(\frac{x_i - \overline{x}}{s_n^*}\right)^4 - \frac{3(n-1)^2}{(n-2)(n-3)}.$$

Kis minták esetén $\widehat{K_2}$ jobban közelíti az elméleti ferdeséget, mint $\widehat{K_1}$.

B. $\widehat{K_3}$–mutató

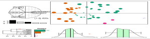
$$\widehat{K_3} := \frac{\widehat{Q}_3 - \widehat{Q}_1}{2(\widehat{D}_9 - \widehat{D}_1)}.$$

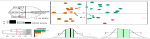
Nyilván, $\widehat{K_3} \in [0, 1/2]$.

Ha X egy normális eloszlású v. v., akkor

$$\frac{Q_3 - Q_1}{2(D_9 - D_1)} = \frac{\Phi^{-1}(3/4) - \Phi^{-1}(1/4)}{2(\Phi^{-1}(9/10) - \Phi^{-1}(1/10))}$$

$$\approx \frac{0.675 - (0.675)}{2(1.285 - (-1.285))} \approx 0.263.$$





Csúcsossági mutatók értelmezése

Ha $\widehat{K_2} > 0$, akkor a vizsgált minta sűrűséghisztogramja $\overline{x}$ körül csúcsosabb (hegyesebb), mint az $\mathcal{N}(\overline{x}, (s_n^*)^2)$ sűrűségfüggvénye és vastagabb farkakkal is rendelkezik. (Elnevezés: leptokurtic.)

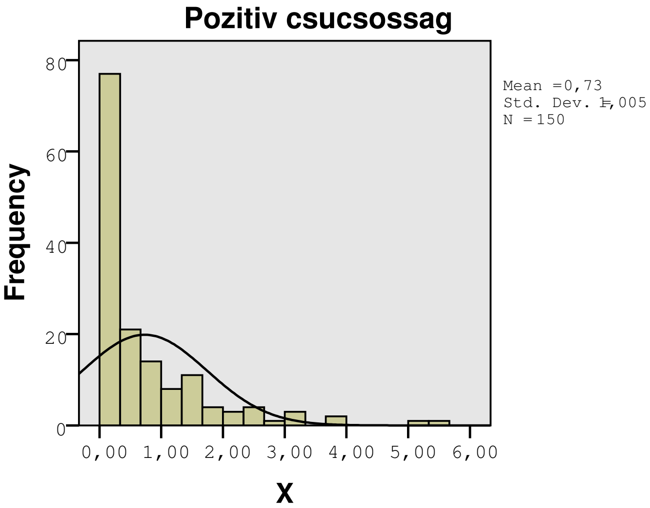
Ha $\widehat{K_2} < 0$, akkor a vizsgált minta sűrűséghisztogramja $\overline{x}$ körül lapultabb, mint az $\mathcal{N}(\overline{x}, (s_n^*)^2)$ sűrűségfüggvénye és vékonyabb farkakkal is rendelkezik. (Elnevezés: platykurtic.)

Ha $\widehat{K_3} > 0.263$, akkor ugyanaz az értelmezés, mint $\widehat{K_2} < 0$ esetén.

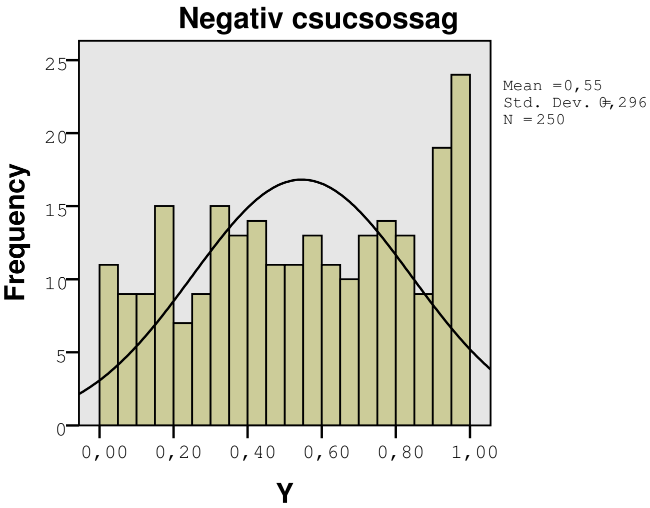
Ha $\widehat{K_3} < 0.263$, akkor ugyanaz az értelmezés, mint $\widehat{K_2} > 0$ esetén.

A $\widehat{K_1}$ mutató értelmezése analóg a $\widehat{K_2}$ mutatóéhoz.

Statisztikai elemzés
statisztikai sokaság
statisztikai változó
mérési skálák
statisztikai módszerek
Statisztikai minta
minta, mintatér, rendezett minta, a statisztika fogalma
tapasztalati eloszlásfüggvény
mintavételi módszerek
gyorsrendezés
Leíró statisztikák
Gyakorisági tábla
Helyzetmutatók
Szóródási mutatók
Alakmutatók
Grafikus elemzés
Hisztogramok
Decimális grafikon
Boxdiagram
Gauss-papír
QQ-plot és PP-plot
Függelék
Irodalomjegyzék
Összefoglalás



Előállította az SPSS rendszer



Előállította az SPSS rendszer

Statisztikai elemzés
statisztikai sokaság
statisztikai változó
mérési skálák
statisztikai módszerek
Statisztikai minta
minta, mintatér, rendezett minta, a statisztika fogalma
tapasztalati eloszlásfüggvény
mintavételi módszerek
gyorsrendezés
Leíró statisztikák
Gyakorisági tábla
Helyzetmutatók
Szóródási mutatók
Alakmutatók
Grafikus elemzés
Hisztogramok
Decimális grafikon
Boxdiagram
Gauss-papír
QQ-plot és PP-plot
Függelék
Irodalomjegyzék
Összefoglalás

Megjegyzés

A homogén populációra vett minta eloszlása általában unimodális. Multimodális eloszlás keverékre, inhomogén populációra utal. Ekkor haladottabb statisztikai módszereket kell használni.

A legismertebb grafikus elemzési módok

- empirikus eloszlásfüggvény,
- gyakorisági- és sűrűséghisztogram,
- kördiagram,
- oszlopdiagram,
- decimális grafikon (stem-and-leaf plot, leveles ág ábra),
- boxdiagram (boxplot),
- Gauss-papír,
- QQ-plot, PP-plot.

Hisztogram

Tekintsünk egy $x_1, \ldots, x_n$ mintát.

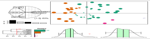
Osszuk fel a számegyenest $y_1 < \ldots < y_r$ osztópontokkal.

Tegyük fel, hogy minden mintaelem az (y_1, y_r) intervallumba esik.

Jelölje ν_i az $[y_i, y_{i+1})$ intervallumba eső mintaelemek számát, $i = 1, \ldots, r-1$.

Rajzoljunk az $[y_i, y_{i+1})$ intervallum fölé egy ν_i-vel arányos területű téglalapot, $i = 1, \ldots, r-1$.

Így kapunk egy hisztogramot.



Gyakorisági hisztogram: a téglalapok összterülete n.

Sűrűséghisztogram: a téglalapok összterülete 1.

Definíció

Gyakorisági hisztogram: $f_n : \mathbb{R} \to [0, \infty)$,

$$f_n(x) := \begin{cases} \frac{\nu_i}{(y_{i+1} - y_i)} & \text{ha } x \in [y_i, y_{i+1}),\ i = 1, \ldots, r-1, \\ 0 & \text{ha } x \notin [y_1, y_r). \end{cases}$$

Sűrűséghisztogram: $f_n : \mathbb{R} \to [0, \infty)$,

$$f_n(x) := \begin{cases} \frac{\nu_i}{n(y_{i+1} - y_i)} & \text{ha } x \in [y_i, y_{i+1}),\ i = 1, \ldots, r-1, \\ 0 & \text{ha } x \notin [y_1, y_r). \end{cases}$$

Az osztópontok sűrítésével, illetve ritkításával érhetjük el, hogy a hisztogram ne legyen se túl durva, se ne ugráljon.

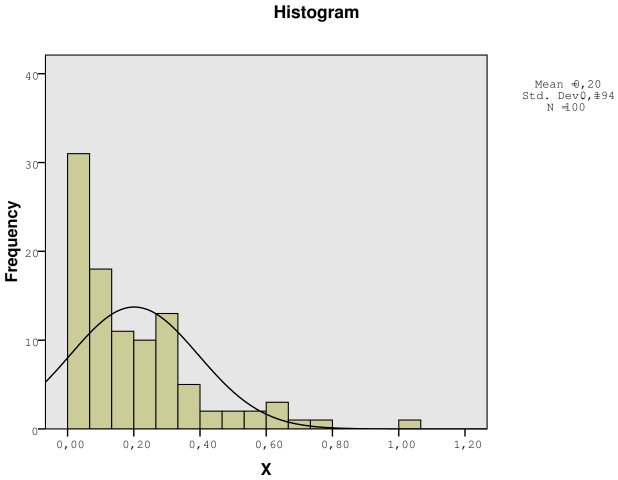
Különböző ajánlások vannak a beosztás sűrűségének megállapítására a mintaelemszám függvényében. Egyenlő osztályközű beosztást feltételezve:

- Scott–féle módszer:

$$\frac{3.5 s_n}{\sqrt[3]{n}},$$

- Freedman–Diaconis módszer:

$$\frac{2\widehat{h}}{\sqrt[3]{n}} = \frac{2(\widehat{Q}_3 - \widehat{Q}_1)}{\sqrt[3]{n}}.$$



Előállította az SPSS rendszer

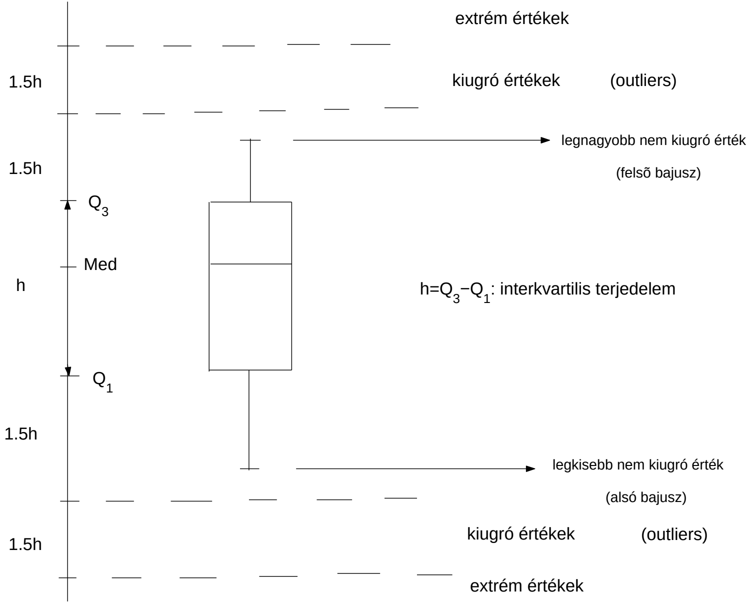
Decimális grafikon

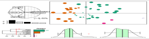
A stem=ág és leaf=levél jelentésnek megfelelően úgy bontjuk fel a számértékű adatokat, hogy a fő rész tizedes jegyei adják a „stem”-et (ágat), az utolsó jegy pedig a „leaf”-et (levelet).

Pl. a 34 szám esetén a „stem” 3, a „leaf” 4.

A decimális grafikon segítségével könnyen elkészíthető a rendezett minta.

Boxdiagram (nem méretarányos ábra)





Gauss-papír

Legyen $x_1, \ldots, x_n$ egy minta(realizáció).

Legyen X egy olyan valószínűségi változó, melynek F_X eloszlásfüggvénye szigorúan monoton növekvő.

Legyen $y_1 < y_2 < \ldots < y_r$ a számegyenes egy felosztása.

Jelölje μ_i az y_i-nél kisebb mintalemek számát.

Tekintsük az

$$\left(y_i, F_X^{-1}\left(\frac{\mu_i}{n}\right)\right), \qquad i = 1, \ldots, r,$$

pontokat összekötő töröttvonalat.

Ha az $x_1, \ldots, x_n$ minta(realizáció) az X valószínűségi változóra vett minta(realizáció), akkor ez a töröttvonal közelítőleg az $y = x$ egyenes.

Valóban, nagy n esetén

$$\frac{\mu_i}{n} \approx P(X < y_i) = F_X(y_i),$$

amiből

$$F_X^{-1}\left(\frac{\mu_i}{n}\right) \approx y_i, \qquad i = 1, \ldots, n.$$

Az eddigiek tetszőleges (szigorúan monoton növekvő eloszlásfüggvényű) alapeloszlás esetén érvényesek. Azonban elsősorban normalitás vizsgálatára használják.

A Gauss-papír elnevezés onnan adódik, hogy a töröttvonalat eredetileg speciális beosztású, ún. Gauss-papírra rajzolták.

Ha $X \sim \mathcal{N}(m, \sigma^2)$, akkor az

$$\left(y_i, \Phi^{-1}\left(\frac{\mu_i}{n}\right)\right) \approx \left(y_i, \frac{y_i - m}{\sigma}\right), \qquad i = 1, \ldots, r,$$

pontokat összekötő töröttvonal közelítőleg $1/\sigma$ meredekségű, az y-tengelyt $-m/\sigma$-ban metsző egyenes.

A normalitást akkor kell elvetni, ha a töröttvonal az egyenestől „szisztematikus” eltérést mutat (pl. „S-alakú”).

QQ-plot és PP-plot

Legyen $x_1, \ldots, x_n$ egy olyan minta, hogy nincsenek egybeeső mintaelemek, azaz $x_1^* < x_2^* < \ldots < x_n^*$.

Normal QQ-plot: az

$$\left(x_i^*, s_n^* \Phi^{-1}\left(\frac{i-1/2}{n}\right) + \overline{x}\right), \qquad i = 1, \ldots, n,$$

pontokat összekötő töröttvonal.

Megj.: $\Phi^{-1}((i-1/2)/n)$ nemmás, mint a standard normális eloszlás $(i-1/2)/n$ kvantilise, korábbi jelöléseinkkel $Q_{\mathcal{N}(0,1)}((i-1/2)/n)$. Továbbá, a fenti képletekben a -1 helyetti $-1/2$ egyfajta folytonossági korrekciónak tekinthető.

Ha a minta(realizáció) $\mathcal{N}(m, \sigma^2)$ eloszlásból származik, akkor elég nagy n-re az ábrázolt pontok közelítőleg az $y = x$ egyenesre illeszkednek.

Valóban, a Glivenko–Cantelli–tétel alapján

$$P\left(\lim_{n\to\infty}\sup_{x\in\mathbb{R}}|F_n^*(x)-F_{\mathcal{N}(m,\sigma^2)}(x)|=0\right)=1.$$

Mivel

$$F_{\mathcal{N}(m,\sigma^2)}(x)=\Phi\left(\frac{x-m}{\sigma}\right),\qquad x\in\mathbb{R},$$

elég nagy n-re

$$\Phi^{-1}\left(F_n^*(x)\right)\approx\frac{x-m}{\sigma},\qquad x\in\mathbb{R}.$$

Az $x=x_i^*$ választással:

$$F_n^*(x_i^*)=\frac{i-1}{n}\quad\Longrightarrow\quad\Phi^{-1}\left(\frac{i-1}{n}\right)\approx\frac{x_i^*-m}{\sigma},$$

ami alapján

$$\left(x_i^*,s_n^*\Phi^{-1}\left(\frac{i-1/2}{n}\right)+\overline{x}\right)\approx\left(x_i^*,s_n^*\frac{x_i^*-m}{\sigma}+\overline{x}\right)$$
$$\approx(x_i^*,x_i^*).$$

Detrended Normal QQ-plot: az

$$\left(x_i^*, x_i^* - \left(s_n^* \Phi^{-1}\left(\frac{i-1/2}{n}\right) + \overline{x}\right)\right), \qquad i = 1, \ldots, n,$$

pontokat összekötő töröttvonal.

Ha a minta(realizáció) $\mathcal{N}(m, \sigma^2)$ eloszlásból származik, akkor elég nagy n-re

$$\left(x_i^*, x_i^* - \left(s_n^* \Phi^{-1}\left(\frac{i-1/2}{n}\right) + \overline{x}\right)\right) \approx (x_i^*, 0),$$

azaz normalitás esetén az ábrázolt pontok közelítőleg az $y = 0$ egyenesre illeszkednek.

Normal PP-plot: az

$$\left(\frac{i-1/2}{n}, \Phi\left(\frac{x_i^*-\overline{x}}{s_n^*}\right)\right), \qquad i=1,\ldots,n,$$

pontokat összekötő töröttvonal.

Ez a töröttvonal a $[0,1]\times[0,1]$ egységnégyzeten belül helyezkedik el.

Ha a minta(realizáció) $\mathcal{N}(m,\sigma^2)$ eloszlásból származik, akkor elég nagy n-re az ábrázolt pontok közelítőleg az $y=x$ egyenesre illeszkednek.

Valóban, elég nagy n-re

$$\Phi\left(\frac{x_i^*-\overline{x}}{s_n^*}\right) \approx \Phi\left(\frac{x_i^*-m}{\sigma}\right) \approx F_n^*(x_i^*) = \frac{i-1}{n}.$$

Detrended Normal PP-plot: az

$$\left(\frac{i-1/2}{n}, \frac{i-1/2}{n} - \Phi\left(\frac{x_i^* - \overline{x}}{s_n^*}\right)\right), \qquad i = 1, \ldots, n,$$

pontokat összekötő töröttvonal.

Ha a minta(realizáció) $\mathcal{N}(m, \sigma^2)$ eloszlásból származik, akkor elég nagy n-re az ábrázolt pontok közelítőleg az $y = 0$ egyenesre illeszkednek.

Jelölje $F_{Exp(1)}$ az 1-paraméterű exponenciális eloszlás eloszlásfüggvényét.

Exponential QQ-plot: az

$$\left(x_i^*, \overline{x} F_{Exp(1)}^{-1}\left(\frac{i-1/2}{n}\right)\right), \qquad i = 1, \dots, n,$$

pontokat összekötő töröttvonal.

Ha a minta(realizáció) $Exp(\lambda)$ eloszlásból származik, ahol $\lambda > 0$, akkor elég nagy n-re az ábrázolt pontok közelítőleg az $y = x$ egyenesre illeszkednek.

Valóban, a Glivenko–Cantelli–tétel alapján

$$P\left(\lim_{n\to\infty}\sup_{x\in\mathbb{R}}|F_n^*(x)-F_{Exp(\lambda)}(x)|=0\right)=1.$$

Mivel

$$F_{Exp(\lambda)}(x)=F_{Exp(1)}(\lambda x),\qquad x\in\mathbb{R},$$

elég nagy n-re

$$F_{Exp(1)}^{-1}\left(F_n^*(x)\right)\approx\lambda x,\qquad x\in\mathbb{R}.$$

Az $x=x_i^*$ választással:

$$F_n^*(x_i^*)=\frac{i-1}{n}\quad\Longrightarrow\quad F_{Exp(1)}^{-1}\left(\frac{i-1}{n}\right)\approx\lambda x_i^*,$$

ami alapján

$$\left(x_i^*,\overline{x}F_{Exp(1)}^{-1}\left(\frac{i-1/2}{n}\right)\right)\approx(x_i^*,\overline{x}\lambda x_i^*)\approx(x_i^*,x_i^*),$$

ahol az utolsó lépésben felhasználtuk, hogy egy λ paraméterű exponenciális eloszlású valószínűségi változó várható értéke $1/\lambda$.

Detrended Exponential QQ-plot: az

$$\left(x_i^*, x_i^* - \overline{x}F^{-1}_{Exp(1)}\left(\frac{i-1/2}{n}\right)\right), \qquad i = 1, \dots, n,$$

pontokat összekötő töröttvonal.

Ha a minta(realizáció) $Exp(\lambda)$ eloszlásból származik, ahol $\lambda > 0$, akkor elég nagy n-re az ábrázolt pontok közelítőleg az $y = 0$ egyenesre illeszkednek.

Exponential PP-plot: az

$$\left(\frac{i-1/2}{n}, F_{Exp(1)}\left(\frac{x_i^*}{\overline{x}}\right)\right), \qquad i = 1, \dots, n,$$

pontokat összekötő töröttvonal.

Ha a minta(realizáció) $Exp(\lambda)$ eloszlásból származik, ahol $\lambda > 0$, akkor elég nagy n-re az ábrázolt pontok közelítőleg az $y = x$ egyenesre illeszkednek.

Valóban, elég nagy n-re

$$F_{Exp(1)}\left(\frac{x_i^*}{\overline{x}}\right) \approx F_{Exp(1)}\left(\lambda x_i^*\right) \approx F_{Exp(\lambda)}\left(x_i^*\right) \approx \frac{i-1}{n}.$$

Detrended Exponential PP-plot: az

$$\left(\frac{i - 1/2}{n}, \frac{i - 1/2}{n} - F_{Exp(1)}\left(\frac{x_i^*}{\overline{x}} \right) \right), \qquad i = 1, \ldots, n,$$

pontokat összekötő töröttvonal.

Ha a minta(realizáció) $Exp(\lambda)$ eloszlásból származik, ahol $\lambda > 0$, akkor elég nagy n-re az ábrázolt pontok közelítőleg az $y = 0$ egyenesre illeszkednek.

Függelék

Állítás

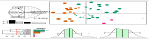
Legyen $X_1, \ldots, X_n$ *minta* $\mathcal{N}(m, \sigma^2)$ *eloszlásra. Ekkor*

$$\mathbb{E}(\widehat{S_2}) = \mathbb{E}\left(\frac{n}{(n-1)(n-2)}\sum_{i=1}^{n}\left(\frac{X_i - \overline{X}}{s_n^*}\right)^3\right) = 0,$$

és

$$\mathbb{E}(\widehat{K_2}) = \mathbb{E}\left(\frac{n(n+1)}{(n-1)(n-2)(n-3)}\sum_{i=1}^{n}\left(\frac{X_i - \overline{X}}{s_n^*}\right)^4 - \frac{3(n-1)^2}{(n-2)(n-3)}\right) = 0.$$

Azaz normális eloszlásból származó minta esetén a tapasztalati ferdeség és lapultság várható értéke 0, így megegyezik az elméleti ferdeséggel, illetve lapultsággal.



Bizonyítás. Minden $r \in \mathbb{N}$ esetén legyen

$$M_r^c := \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^r .$$

Ekkor tetszőleges eloszlásból származó minta esetén

$$\mathbb{E}M_1^c = 0, \qquad \mathbb{E}M_2^c = \frac{n-1}{n}\mathbb{E}X_1^2,$$

$$\mathbb{E}M_3^c = \frac{(n-1)(n-2)}{n^2}\mathbb{E}X_1^3,$$

$$\mathbb{E}M_4^c = \frac{(n-1)(n^2-3n+3)}{n^3}\mathbb{E}X_1^4 + \frac{3(n-1)(2n-3)}{n^3}(\mathbb{E}X_1^2)^2.$$

Továbbá, normális eloszlásból származó minta esetén (Fisher, 1930):

$$\mathbb{E}\left(\frac{M_3^c}{(M_2^c)^{3/2}}\right) = 0, \qquad \mathbb{E}\left(\frac{M_4^c}{(M_2^c)^2}\right) = \frac{3(n-1)}{n+1}. \qquad (1)$$

Így

$$\mathbb{E}(\widehat{S_2}) = \frac{n}{(n-1)(n-2)}\mathbb{E}\left(\frac{n \cdot M_3^c}{\frac{n^{3/2}}{(n-1)^{3/2}}(M_2^c)^{3/2}}\right)$$

$$= \frac{n^{1/2}(n-1)}{n-2}\mathbb{E}\left(\frac{M_3^c}{(M_2^c)^{3/2}}\right) = 0,$$

és

$$\mathbb{E}(\widehat{K_2}) = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \mathbb{E}\left(\frac{n \cdot M_4^c}{\frac{n^2}{(n-1)^2}(M_2^c)^2}\right)$$

$$- \frac{3(n-1)^2}{(n-2)(n-3)}$$

$$= \frac{(n-1)(n+1)}{(n-2)(n-3)} \cdot \frac{3(n-1)}{n+1} - \frac{3(n-1)^2}{(n-2)(n-3)} = 0.$$

A következőkben megadjuk Fisher (1) eredményének egy bizonyításvázlatát.

Fisher (1) eredményének bizonyítás vázlata:

Legyen $X_1, \ldots, X_n$ minta $\mathcal{N}(m, \sigma^2)$-re.

- $(\overline{X}, s_n^2)$ teljes és elégséges statisztika, illetve

$$\left(\sum_{i=1}^{n} \left(\frac{X_i - \overline{X}}{s_n^*} \right)^3, \sum_{i=1}^{n} \left(\frac{X_i - \overline{X}}{s_n^*} \right)^4 \right) \qquad (2)$$

  kiegészítő statisztika (azaz eloszlása nem függ (m, σ^2)-től).
- Így feltehetjük, hogy $m = 0$ és $\sigma^2 = 1$.
- Basu 2. tétele szerint $(\overline{X}, s_n^2)$ és (2) függetlenek. Így speciálisan $M_3^c / (M_2^c)^{3/2}$ és $(M_2^c)^{3/2}$ függetlenek.

- Ezért

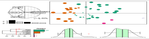
$$\mathbb{E}M_3^c = \mathbb{E}\left(\frac{M_3^c}{(M_2^c)^{3/2}} \cdot (M_2^c)^{3/2}\right)$$
$$= \mathbb{E}\left(\frac{M_3^c}{(M_2^c)^{3/2}}\right)\mathbb{E}((M_2^c)^{3/2}),$$

amiből

$$\mathbb{E}\left(\frac{M_3^c}{(M_2^c)^{3/2}}\right) = \frac{\mathbb{E}M_3^c}{\mathbb{E}(M_2^c)^{3/2}}.$$

- Hasonlóan,

$$\mathbb{E}\left(\frac{M_4^c}{(M_2^c)^2}\right) = \frac{\mathbb{E}M_4^c}{\mathbb{E}(M_2^c)^2}.$$



- $$\mathbb{E}M_3^c = \frac{(n-1)(n-2)}{n^2}\mathbb{E}X_1^3 = 0.$$

- $$\begin{aligned}\mathbb{E}M_4^c &= \frac{(n-1)(n^2-3n+3)}{n^3}\mathbb{E}X_1^4 \\ &\quad + \frac{3(n-1)(2n-3)}{n^3}(\mathbb{E}X_1^2)^2 \\ &= \frac{3(n-1)(n^2-3n+3)}{n^3} + \frac{3(n-1)(2n-3)}{n^3}.\end{aligned}$$

- Fazekas [50. old., 8. Feladat megoldása] alapján

$$\mathbb{E}(M_2^c)^2 = \frac{(n-1)^2}{n^2}\mathbb{E}(s_n^*)^4$$
$$= \frac{(n-1)^2}{n^2}\left(\frac{1}{n}\mathbb{E}X_1^4 + \frac{n^2-2n+3}{n(n-1)}(\mathbb{E}X_1^2)^2\right)$$
$$= \frac{(n-1)^2}{n^2}\left(\frac{3}{n} + \frac{n^2-2n+3}{n(n-1)}\right).$$

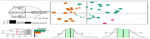
- Így

$$\mathbb{E}\left(\frac{M_4^c}{(M_2^c)^2}\right) = \frac{3(n^2-3n+3)+3(2n-3)}{3(n-1)+(n^2-2n+3)} = \frac{3(n-1)}{n+1}.$$

$\square$

Irodalomjegyzék

1. T. H. Cormen, C. E. Leiserson, R. L. Rivest: Algoritmusok. Műszaki Könyvkiadó, Budapest, 2003.
2. Fazekas I.: (szerk.), Bevezetés a matematikai statisztikába. Kossuth Egyetemi Kiadó. Debrecen, 2003.
3. Füstös L., Kovács E., Meszéna Gy., Simonné Mosolygó N.: Alakfelismerés (Sokváltozós statisztikai módszerek). Új Mandátum Könyvkiadó, Budapest, 2004.
4. Hunyadi L., Vita L.: Statisztika közgazdászoknak. KSH, Budapest, 2002.
5. Lange, K.: Numerical Analysis for Statisticians. Springer. New York, 1998.



Összefoglalás

1. Az elemzés kiinduló pontjai: a vizsgálandó statisztikai sokaság és az őt jellemző változók.
2. Az elemzés módjai, a változók típusai.
3. A megfigyelt (n-elemű) minta a sokaságra tett konkrét megfigyelések eredménye: n darab valós szám (vagy vektor). Mintatér: a lehetséges minták összessége.
4. A rendezett minta: gyorsrendezés.
5. A tapasztalati eloszlásfüggvény.
6. A minta tömörítése, jellemzése statisztikák útján történik.
7. A leíró statisztikák szerepe.
8. A grafikus elemzés elemei.