BÁNHALMI ÁRPÁD*

Statisztikai mutatók elemi úton

Statistical Indicators in Elementary Way

From the academic year 2004/2005 statistics is a part of the high-school graduation, so that's why it's important that the primary methods of statistics can be demonstrated by elementary means. First we are going into the Jensen's inequality, the special case of which can be shown by comparing of convex ranges very well. The examination is based on the simple statement that a convex polygon bounded by chords of a convex function is a subset of the convex range „over" the graph of the function. At the statistical examinations the „notable" means: the Harmonic, the Geometric, the Arithmetic and the Quadratic mean get an important rule. It can be stated about each of them that they are special cases of the so-called Generalized mean. To summarize the relations among the „notable" means the examination of the Generalized mean is very convenient. With the help of the Generalized mean such functions can be defined, the examination of which makes elementary properties of more statistical indicators outcrop and makes their relations to each other clear. The examination of two functions – that are interesting because of their geometric properties – the δ(x) and the σ(x) gives help for analyzing the median, the population variance and Gini's Coefficient.

A 2004/2005-ös tanévtől kezdve a statisztika az érettségi anyag része, ezért fontos, hogy elemi eszközökkel – kapcsolva a középiskolás matematika anyaghoz – szemléltethetőek legyenek a statisztika főbb módszerei.

Először a JENSEN-egyenlőtlenségre térünk ki, aminek a speciális esete igen jól szemléltethető konvex tartományok összehasonlításával. A vizsgálat azon az egyszerű megállapításon alapul, hogy egy konvex függvény húrjai által határolt konvex sokszöglap részhalmaza a függvény grafikonja „feletti" konvex tartománynak.

Definíció: Legyen adva $f : R \to R$ konvex függvény.

$$H_f = \left\{ \begin{bmatrix} x \\ y \end{bmatrix} \middle| f(x) \le y \right\}.$$

A H_f ponthalmaz épp az $f : R \to R$ függvény grafikonja „feletti" tartományt jelenti. Ismeretes, ha $f : R \to R$ konvex, akkor a hozzá tartozó H_f tartomány konvex. Tehát a H_f tartomány tetszőleges két pontját összekötő szakaszt is tartalmazza, ami az jelenti, hogy az $f : R \to R$ függvény grafikonjának tetszőleges húrjai által határolt sokszöglap is részhalmaza.

Definíció: Legyenek adva a $0 < x_1 \le x_2 \le x_3 \le \ldots \le x_k$ valós számok. Legyen adva $f : R \to R$, konvex függvény. Jelölje $\underline{x} = [x_1, x_2, x_3, \ldots, x_k]^*$! Ekkor az

$$\begin{bmatrix} x_i \\ f(x_i) \end{bmatrix}$$

pontok által kifeszített tartomány:

* BGF Külkereskedelmi Főiskolai Kar, Matematika–Statisztika Tanszék, főiskolai tanársegéd.

$$T_{\underline{x},f} = \left\{ \begin{bmatrix} x \\ y \end{bmatrix} \middle| \begin{bmatrix} x \\ y \end{bmatrix} = \sum_{i=1}^{k} \lambda_i \cdot \begin{bmatrix} x_i \\ f(x_i) \end{bmatrix}, 0 \le \lambda_i, \sum_{i=1}^{k} \lambda_i = 1 \right\}.$$

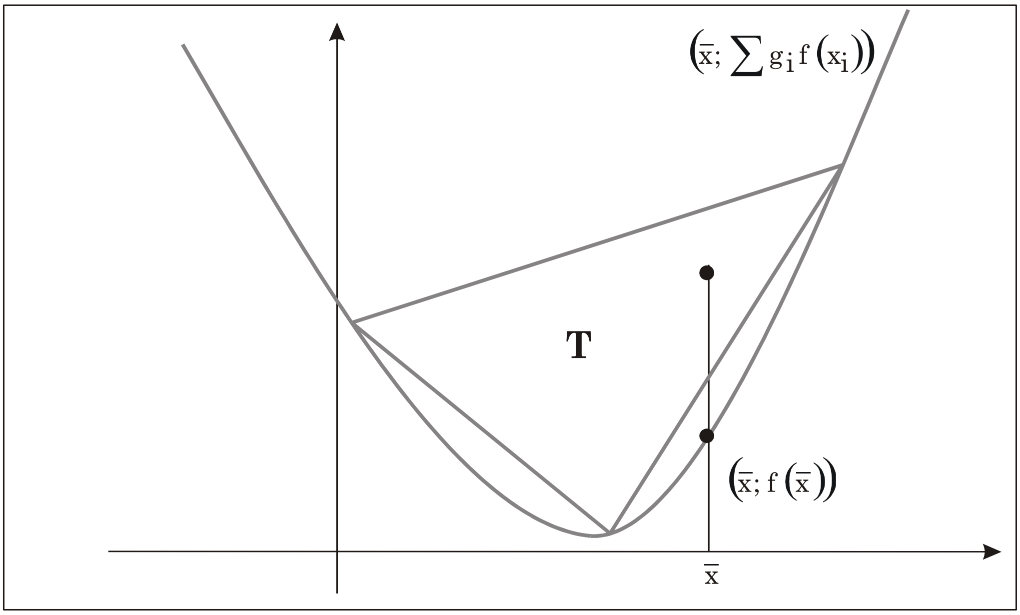
A most definiált tartomány konvex sokszög, hisz *k* darab vektor konvex lineáris kombinációja. Mivel az

$$\begin{bmatrix} x_i \\ f(x_i) \end{bmatrix}$$

pontok éppen az $f: R \to R$ függvény grafikonjának pontjai, az is elmondható, hogy a szóban forgó sokszög minden oldala a grafikon egy húrja, tehát a fent elmondottak alapján $T_{\underline{x},f} \subset H_f$. Legyen a szokásos jelöléssel

$$\bar{x} = \sum_{i=1}^{k} g_i \cdot x_i ,$$

illetve adott $0 < x_1 \le x_2 \le x_3 \le ... \le x_k$ értékek, és $g_1, g_2, g_3, ..., g_k$ nem negatív, $\sum_{i=1}^{k} g_i = 1$ súlyok esetén $\bar{f} = \sum_{i=1}^{k} g_i \cdot f(x_i)$! Figyelembe véve, hogy a $\begin{bmatrix} \bar{x} \\ f(\bar{x}) \end{bmatrix}$ az $f: R \to R$ függvény grafikonjának egy pontja, az $\begin{bmatrix} \bar{x} \\ \bar{f} \end{bmatrix}$ pedig a $T_{\underline{x},f}$ tartomány egy pontja, az $f(\bar{x}) \le \bar{f}$ egyenlőtlenség áll fenn.



1. ábra

Ennek alapján kimondható a JENSEN-egyenlőtlenség: Legyenek adva a $0 < x_1 \leq x_2 \leq x_3 \leq \ldots \leq x_k$ valós számok, és a $g_1, g_2, g_3, \ldots, g_k$ nem negatív súlyok, amikre $\sum_{i=1}^{k} g_i = 1$ teljesül, továbbá $f: R \to R$, konvex függvény. Ekkor

$$f\left(\sum_{i=1}^{k} g_i \cdot x_i\right) \leq \sum_{i=1}^{k} g_i \cdot f(x_i).$$

Speciális függvényekre érdemes kipróbálni, hogy a JENSEN-egyenlőtlenség konkrétan milyen összefüggésre vezet. Ha $f(x) = ax + b$, akkor a függvényre teljesül a megkövetelt feltétel: a konvexitás. Erre a függvényre az egyenlőség teljesül, hiszen az $a \cdot \bar{x} + b = \sum_{i=1}^{k} g_i \cdot (ax_i + b)$ összefüggés tetszőleges *a*-ra és *b*-re teljesül (ez a számtani átlag ismert tulajdonsága). Ha $f(x) = x^2$, akkor a konvexitás feltétele szintén teljesül. Erre a függvényre alkalmazva az egyenlőtlenséget a

$$\left(\sum_{i=1}^{k} g_i \cdot x_i\right)^2 \leq \sum_{i=1}^{k} g_i \cdot x_i^2$$

összefüggést kapjuk. Gyökvonás után nyilvánvalóvá válik, hogy a számtani és négyzetes közepek közti összefüggéshez jutottunk:

$$\sum_{i=1}^{k} g_i \cdot x_i \leq \sqrt{\sum_{i=1}^{k} g_i \cdot x_i^2}.$$

A statisztikai vizsgálatoknál fontos szerepet kapnak a *nevezetes közepek*, a harmonikus, mértani, számtani és négyzetes közepek. Mindegyikről állítható, hogy az úgynevezett hatványközepek speciális esetei. A nevezetes közepek közti relációk összefoglalására meglehetősen alkalmas a hatványközepek vizsgálata.

Definíció: Legyenek adva a $0 < x_1 \leq x_2 \leq x_3 \leq \ldots \leq x_k$ nem mind egyenlő valós számok, és a $g_1, g_2, g_3, \ldots, g_k$ pozitív súlyok, amikre $\sum_{i=1}^{k} g_i = 1$ teljesül. Jelölje $\underline{x} = [x_1, x_2, x_3, \ldots, x_k]^*$ és $\underline{g} = [g_1, g_2, g_3, \ldots, g_k]^*$. Ekkor tetszőleges valós *x*-re

$$K_{\underline{x},\underline{g}}(x) = \begin{cases} \left(\sum_{i=1}^{k} x_i^x \cdot g_i\right)^{\frac{1}{x}} & \text{, ha } x \neq 0 \\ \prod_{i=1}^{k} x_i^{g_i} & \text{, ha } x = 0 \end{cases}$$

az $0 < x_1 \leq x_2 \leq x_3 \leq \ldots \leq x_k$ értékeknek, a $g_1, g_2, g_3, \ldots, g_k$ súlyokhoz tartozó *x*-edik hatványközepe.

Ha az $0 < x_1 \leq x_2 \leq x_3 \leq \ldots \leq x_k$ értékek és a $g_1, g_2, g_3, \ldots, g_k$ súlyok adottak, akkor egyszerűen $K(x)$-szel jelöljük a függvényt.

A K(x) függvény meghatározásából nem látszik azonnal, hogy a függvény meglehetősen „szépen” viselkedik.

I. A K(x) függvény folytonos

Az $x \neq 0$ esetben ez magától értetődő állítás, hiszen a $\left(\sum_{i=1}^{k} x_i^{x} \cdot g_i\right)^{\frac{1}{x}}$ függvényt ekvivalensen átalakítva $\left(\sum_{i=1}^{k} x_i^{x} \cdot g_i\right)^{\frac{1}{x}} = e^{\frac{1}{x}\cdot\ln\left(\sum_{i=1}^{k} x_i^{x}\cdot g_i\right)}$ összetett függvényhez jutunk, ami folytonos. Az $x = 0$ eset okoz némi gondot. Azt kell csak belátni, hogy a K(x) függvény *0*-ban vett határértéke létezik, és ez megegyezik a függvény helyettesítési értékével. Ez pedig a következő határérték kiszámításával beláthatóvá válik: $\lim_{x\to 0} K(x) =$ a függvény fenti átalakítását figyelembe véve $= \lim_{x\to 0} e^{\frac{\ln\left(\sum_{i=1}^{k} x_i^{x}\cdot g_i\right)}{x}} =$ a határátmenet elvét alkalmazva $= e^{\lim_{x\to 0}\frac{\ln\left(\sum_{i=1}^{k} x_i^{x}\cdot g_i\right)}{x}} =$ a kitevő egy „$\frac{0}{0}$” típusú kritikus határérték, amire a L'HOSPITAL-szabályt alkalmazzuk $= e^{\lim_{x\to 0}\frac{\sum_{i=1}^{k} x_i^{x}\cdot \ln x_i \cdot g_i}{\sum_{i=1}^{k} x_i^{x}\cdot g_i}} =$ felhasználjuk, hogy $x_i^{x} \to 1$ $= e^{\frac{\sum_{i=1}^{k} \ln x_i \cdot g_i}{\sum_{i=1}^{k} g_i}} =$ a logaritmus azonosságai alapján $= \prod_{i=1}^{k} x_i^{g_i}$.

II. A K(x) függvény szigorúan monoton nő

Először: $n_1 > n_2 > 0$ esetén az $f(x) = x^{\frac{n_1}{n_2}}$ függvény konvex a pozitív valós számok halmazán, hiszen $f^{//}(x) = \frac{n_1}{n_2}\cdot\left(\frac{n_1}{n_2} - 1\right)\cdot x^{\frac{n_1}{n_2}-2} > 0$. Az $f(x) = x^{\frac{n_1}{n_2}}$ függvényre és $0 < x_1^{n_2} \le x_2^{n_2} \le x_3^{n_2} \le \ldots \le x_k^{n_2}$ értékekre alkalmazva a JENSEN-egyenlőtlenséget kapjuk, hogy

$$\left(\sum_{i=1}^{k} g_i \cdot x_i^{n_2}\right)^{\frac{n_1}{n_2}} \le \sum_{i=1}^{k} g_i \cdot \left(x_i^{n_2}\right)^{\frac{n_1}{n_2}} .$$

Mindkét oldalt $\frac{1}{n_1}$-edik hatványra emelve

$\left(\sum_{i=1}^{k} g_i \cdot x_i^{n_2}\right)^{\frac{1}{n_2}} \leq \left(\sum_{i=1}^{k} g_i \cdot x_i^{n_1}\right)^{\frac{1}{n_1}}$ adódik, ami a K(x) függvény szigorú monoton növekedését fejezi ki pozitív kitevők esetén.

Másodszor: $0 > n_2 > n_1$ esetén az $f(x) = x^{\frac{n_1}{n_2}}$ függvény konvex a pozitív valós számok halmazán, hiszen

$$f''(x) = \frac{n_1}{n_2} \cdot \left(\frac{n_1}{n_2} - 1\right) \cdot x^{\frac{n_1}{n_2} - 2} > 0 .$$

Az $f(x) = x^{\frac{n_1}{n_2}}$ függvényre és $0 < x_1^{n_2} \leq x_2^{n_2} \leq x_3^{n_2} \leq \ldots \leq x_k^{n_2}$ értékekre alkalmazva a JENSEN-egyenlőtlenséget kapjuk, hogy

$$\left(\sum_{i=1}^{k} g_i \cdot x_i^{n_2}\right)^{\frac{n_1}{n_2}} \leq \sum_{i=1}^{k} g_i \cdot \left(x_i^{n_2}\right)^{\frac{n_1}{n_2}} .$$

Mindkét oldalt $\frac{1}{n_1}$ -edik hatványra emelve $\left(\sum_{i=1}^{k} g_i \cdot x_i^{n_2}\right)^{\frac{1}{n_2}} \geq \left(\sum_{i=1}^{k} g_i \cdot x_i^{n_1}\right)^{\frac{1}{n_1}}$ adódik, ami a K(x) függvény szigorú monoton növekedését fejezi ki negatív kitevők esetén. Mivel a K(x) függvény negatív és pozitív tartományon külön-külön szigorúan monoton nő, folytonos, tehát szigorúan monoton nő az egész értelmezési tartományán.

III. A K(x) függvény –∞-ben a legkisebb értékhez, ∞-ben pedig a legnagyobb értékhez tart.

$$\lim_{x \to -\infty} K(x) = \lim_{x \to -\infty} x_1 \cdot \left(g_1 + \sum_{i=2}^{k} g_i \cdot \left(\frac{x_i}{x_1}\right)^x\right)^{\frac{1}{x}} = x_1 \text{, hasonlóan}$$

$$\lim_{x \to \infty} K(x) = \lim_{x \to \infty} x_k \cdot \left(g_k + \sum_{i=1}^{k-1} g_i \cdot \left(\frac{x_i}{x_k}\right)^x\right)^{\frac{1}{x}} = x_k$$

A most bemutatott tulajdonságokból néhány egyszerű következményt szűrhetünk le. A K(x) függvény a legkisebb és a legnagyobb átlagolandó érték között minden értéket fölvesz – pontosan egyszer. A monoton növekedésből következik, hogy például $x = -1,0,1,2$ értékekre $K(-1) \leq K(0) \leq K(1) \leq K(2)$. Határozzuk meg ezeket a hatványközepeket!

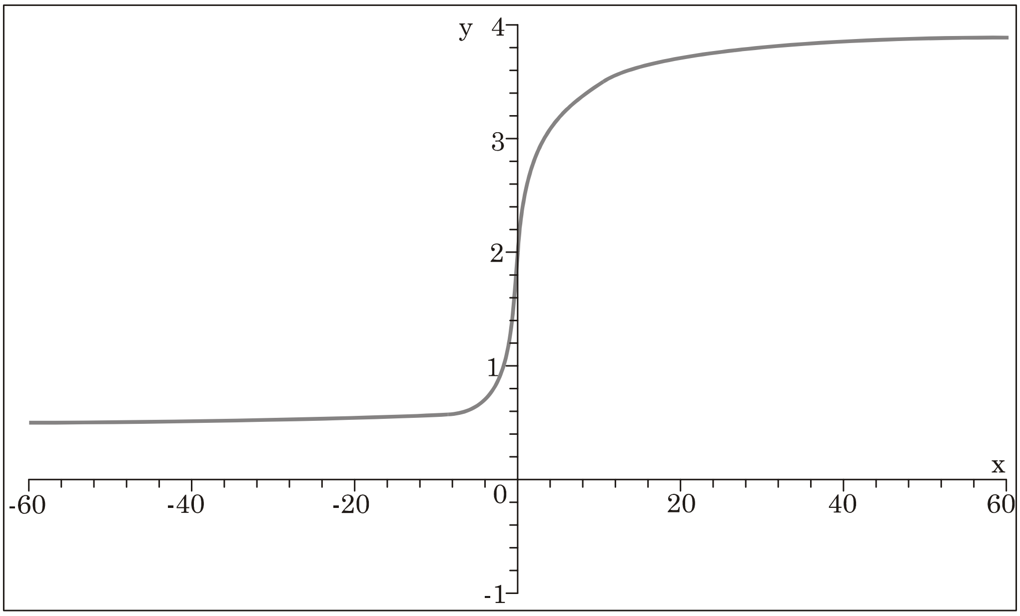
$$K(-1)=\left(\sum_{i=1}^{k} x_i^{-1}\cdot g_i\right)^{\frac{1}{-1}}=\frac{1}{\sum\limits_{i=1}^{k}\frac{g_i}{x_i}}$$ harmonikus közép,

$$K(0)=\prod_{i=1}^{k} x_i^{g_i}$$ geometriai közép,

$$K(1)=\left(\sum_{i=1}^{k} x_i^{1}\cdot g_i\right)^{\frac{1}{1}}=\sum_{i=1}^{k} x_i\cdot g_i$$ számtani közép,

$$K(2)=\left(\sum_{i=1}^{k} x_i^{2}\cdot g_i\right)^{\frac{1}{2}}=\sqrt{\sum_{i=1}^{k} x_i^{2}\cdot g_i}$$ négyzetes közép.

A középiskolában tanult nevezetes közepek közti reláció egyszerű behelyettesítéssel vezethető vissza a K(x) szigorú monoton növekedésére.



2. ábra
Az $x_1 = 0,5$; $x_2 = 2$; $x_3 = 3$; $x_4 = 4$, és
$g_1 = g_2 = g_3 = g_4 = ¼$ esetén a K(x) grafikonja

A hatványközepek segítségével definiálhatóak olyan függvények, amiknek a vizsgálata több statisztikai mutató elemi tulajdonságainak a felszínre hozatalát, és az egymáshoz való viszonyát tisztázza. Egy geometriai tulajdonságai miatt érdekes függvény, a *δ(x)* vizsgálata a *medián* elemzéséhez nyújt segítséget.

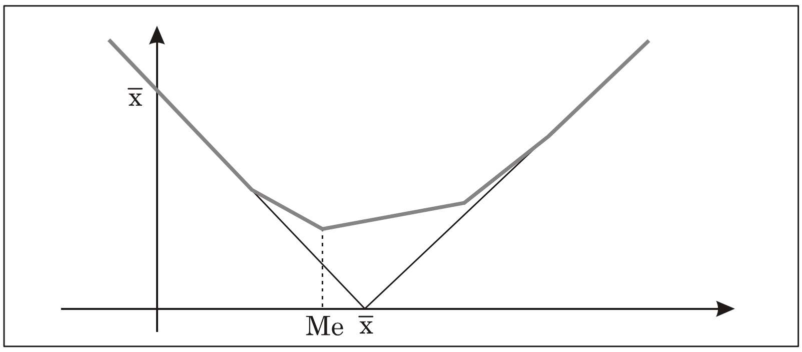
Definíció: Legyenek adva a $0 < x_1 \le x_2 \le x_3 \le \ldots \le x_k$ nem mind egyenlő valós számok, és a $g_1, g_2, g_3, \ldots, g_k$ pozitív súlyok, amikre $\sum_{i=1}^{k} g_i = 1$ teljesül. Tetszőleges valós x-re $\delta(x) = \sum_{i=1}^{k} g_i \cdot |x - x_i|$.

A $\delta(x)$ függvény megmutatja az x valós szám $0 < x_1 \le x_2 \le x_3 \le \ldots \le x_k$ értékektől vett átlagos eltérését. Meghatározzuk a függvény kifejtett alakját, és megvizsgáljuk a tulajdonságait.

Ha $\delta(x) = \sum_{i=1}^{k} g_i \cdot |x - x_i|$ alakjában felbontjuk az abszolútérték-jeleket, a következőt kapjuk:

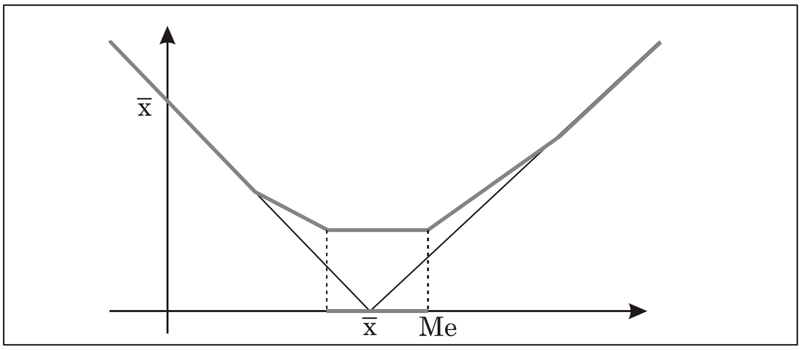
$$\delta(x) = \begin{cases} -x + \bar{x} & , \text{ha } x \le x_1 \\ \vdots & \\ (2g_i{}' - 1)x - (2z_i{}' - 1)\bar{x} & , \text{ha } x_i < x \le x_{i+1} \\ \vdots & \\ x - \bar{x} & , \text{ha } x_k < x \end{cases}$$

A függvény, mivel folytonos függvények összege, folytonos. A fenti felírásból kitűnik, hogy szakaszonként lineáris, esetleg konstans – amennyiben valamilyen i-re a $2g_i' - 1 = 0$. Mivel $2g_i' - 1 < 2g_{i+1}' - 1$, a függvény szakaszonkénti meredeksége szigorúan monoton nő, a függvény konvex. A konvexitás miatt a szélső értéke minimum lehet. A minimum helyét a $2g_i' - 1$ meredekségek előjelváltásából állapíthatjuk meg: ahol $2g_i' - 1 = 0$ teljesül, vagy előjelet vált. Ez $2g_i' - 1 \le 0 < 2g_{i+1}' - 1$ megoldásához vezet. Az egyenlőtlenségrendszert átrendezve kapjuk, hogy $g_i{}' \le \frac{1}{2} < g_{i+1}{}'$. Az x_i elem, amelyik a $g_i{}' \le \frac{1}{2} < g_{i+1}{}'$ tulajdonsággal rendelkezik, a medián.



3. ábra

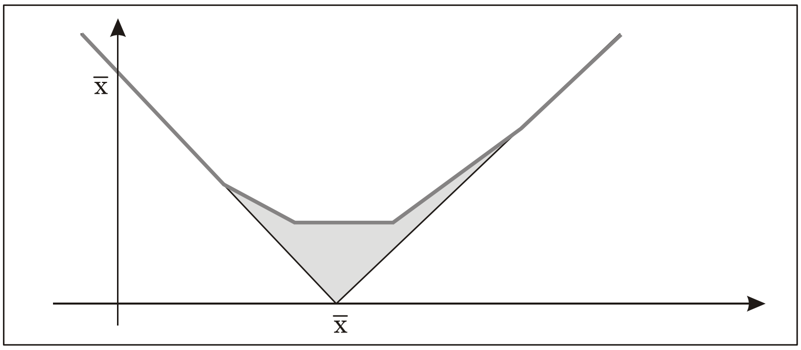
Előfordulhat viszont, hogy a $\delta(x)$ a minimumát nem egy helyen veszi föl, hanem $2g_i' - 1 = 0$ esetén egy egész szakaszon. Ilyenkor a szakasz összes pontja teljesíti a mediánnal szemben támasztható követelményeket – beleértve a végpontokat is.



4. ábra

A $\delta(x)$ függvény tehát az ismérvértékek szóródásáról ad információt: ha az adatok *x* értéktől vett abszolút eltérését számtani középpel átlagoljuk, az a medián esetén a minimális.

A $\delta(x)$ függvény egy másik érdekes tulajdonsága, hogy a szórásról – pontosabban a szórásnégyzetről – is szemléletes képet ad. A beszínezett terület éppen a szórásnégyzettel egyenlő (*5. ábra*).



5. ábra

Ha a grafikon pontjai az $y = -x + \bar{x}$ és $y = x - \bar{x}$ egyenesekhez esnek közel, akkor kisebb a szórás, ha pedig távolabb esnek, a szórás növekszik.

A $\delta(x)$ definíciójából egyszerűen adódik, hogy $G = \sum_{i=1}^{k} g_i \cdot \delta(x_i)$, ahol a G a GINI-féle együtthatót jelöli. Figyelembe véve, hogy a $\delta(x)$ konvex, a JENSEN-egyenlőtlenséget alkalmazva kapjuk, hogy

$$\delta\left(\overline{x}\right)=\delta\left(\sum_{i=1}^{k} g_i \cdot x_i\right) \le G = \sum_{i=1}^{k} g_i \cdot \delta(x_i)\,.$$

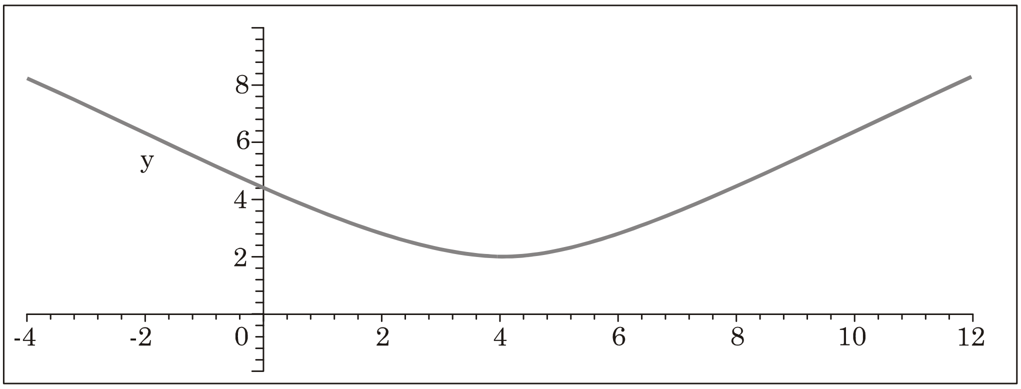
Definíció: Legyenek adva a $0 < x_1 \le x_2 \le x_3 \le \ldots \le x_k$ nem mind egyenlő valós számok, és a $g_1, g_2, g_3, \ldots, g_k$ pozitív súlyok, amikre $\sum_{i=1}^{k} g_i = 1$ teljesül. Tetszőleges valós x-re $\sigma(x) = \sqrt{\sum_{i=1}^{k} g_i (x - x_i)^2}$.

A $\sigma(x)$ függvény megmutatja az x valós szám $0 < x_1 \le x_2 \le x_3 \le \ldots \le x_k$ értékektől vett átlagos eltérését, ha négyzetes középpel átlagolunk. Meghatározzuk a függvény kifejtett alakját, és megvizsgáljuk a tulajdonságait.

Ha $\sigma(x) = \sqrt{\sum_{i=1}^{k} g_i (x - x_i)^2}$ alakjában felbontjuk zárójeleket, a következőt kapjuk:

$$\sigma(x) = \sqrt{\left(x - \overline{x}\right)^2 + \sigma^2}$$

A függvény folytonos. A fenti felírásból kitűnik, hogy nem érzékeny az x_i értékekre, egyedül az átlagtól és a szórástól függ. A függvény képe az elemi geometriából jól ismert hiperbola, mint ilyen, a függvény konvex. A konvexitás miatt a szélső értéke minimum lehet. A minimum helyét a $\sigma(x) = \sqrt{\left(x - \overline{x}\right)^2 + \sigma^2}$ alakjából következően, az átlagnál veszi fel (a $\delta(x)$-szel szemben mindig egyértelműen). A minimum értéke megegyezik a szórással.



6. ábra

A hiperbola aszimptótái az $y = -x + \overline{x}$ és $y = x - \overline{x}$ egyenesek.

A $\sigma(x)$ definíciójából kiolvasható, hogy $\sqrt{\sum_{i=1}^{k} g_i \cdot \sigma^2(x_i)}$ az x_i adatok egymástól vett eltérésének négyzetes átlaga. Ez a G GINI-féle együtthatóhoz hasonlít, ami hasonlóan definiálható, az alapadatok egymástól vett abszolút eltérésének *számtani* átlaga. Kimutatható, hogy $\sqrt{\sum_{i=1}^{k} g_i \cdot \sigma^2(x_i)} = \sqrt{2} \cdot \sigma$. Ez arra világít rá, hogy a szórás nem csak mint átlagtól vett átlagos eltérés értelmezhető, hanem mint az alapadatok egymástól vett eltérésének az átlaga is.

Irodalomjegyzék

HAJDÚ – PINTÉR – RAPPAI – RÉDEY: Statisztika I-II. JPTE, Pécs, 1994.

HUNYADI – VITA: Statisztika I. Aula Kiadó, Budapest, 1991.

KERÉKGYÁRTÓ – MUNDRUCZÓ: Statisztikai módszerek a gazdasági elemzésben. Aula Kiadó, Budapest, 1995.

KORPÁS – MOLNÁR – SZŰTS: Általános statisztika I. rész. Tankönyvkiadó, Budapest, 1992.

KÖVES – PÁRNICZKY: Általános statisztika I-II. Közgazdasági és Jogi Könyvkiadó, Budapest, 1981.