

Dr. Balogh Albert:

A statisztikai adatfeldolgozás néhány érdekessége

2009/05/19

Kérdések:

1. Hogyan becsüljük a tapasztalati eloszlásfüggvényt?
2. Mi az a rendezett minta?
3. Mi az a medián rang és milyen becslések vannak?
4. Hogyan becsüljük a hibaarány 50%-os felső konfidencia határát?
5. Miért tér el az Excel és Minitab kvartilis-számítása?

1.A tapasztalati eloszlásfüggvényt rendszerint a Weibull és a normális eloszlás esetében grafikus módszerrel becsülik.

Ekkor a becsléseket például Gauss(Weibull)papíron ábrázolva normális eloszlás esetében egy egyenest kapunk.

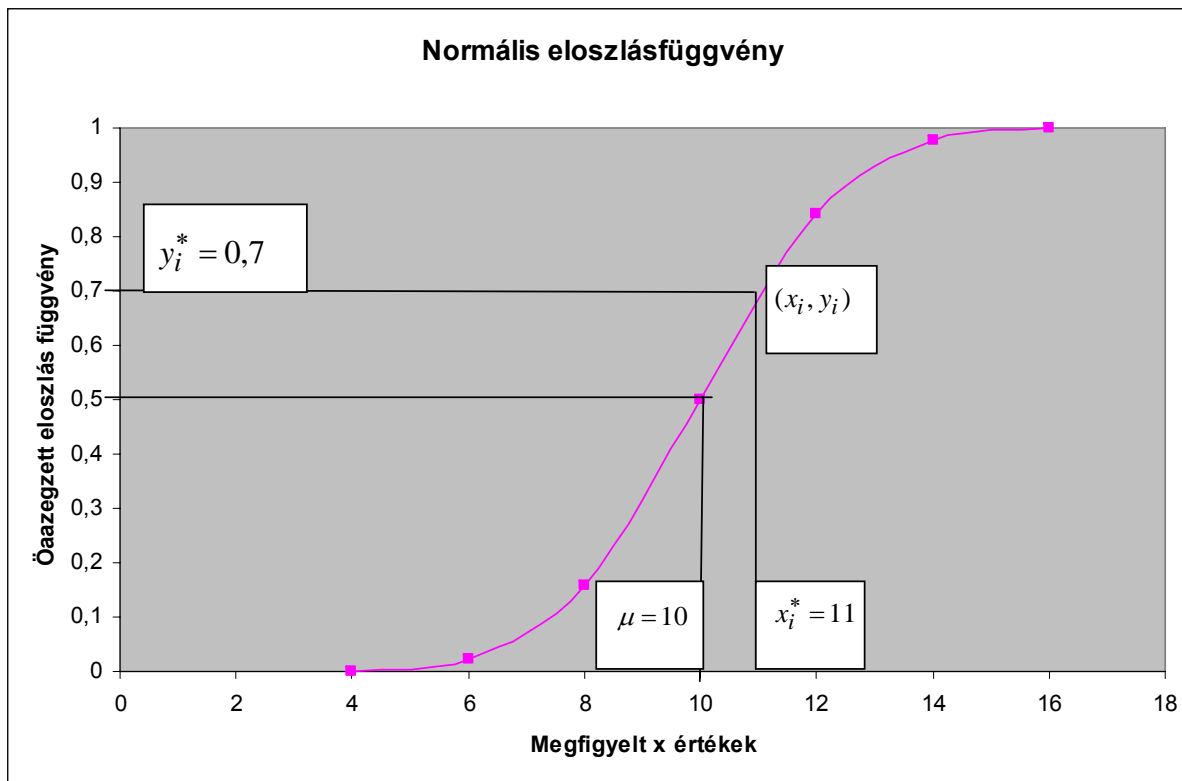
2.A tapasztalati eloszlásfüggvényt a rendezett mintaelemek eloszlásának jellemzőiből határozzuk meg.

Legyenek a rendezett mintaelemek nagyság szerint növekvők:

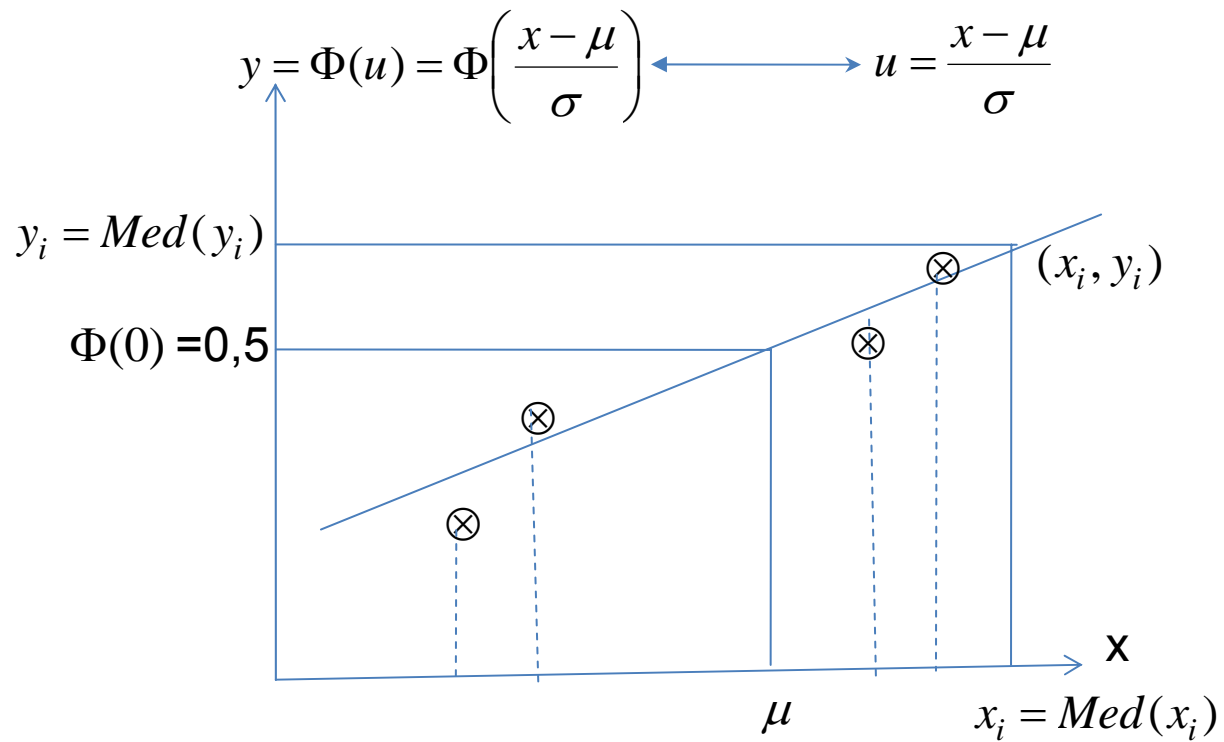
$$x_1 < x_2 < \dots < x_i < \dots < x_n$$

Ekkor az ezekhez tartozó $y_i = F(x_i)$ értékek is rendezett növekvő mintát adnak.

$$y_1 < y_2 < \dots < y_i < \dots < y_n.$$



1. ábra A normális eloszlásfüggvény



⊗ Kérdés: mivel becsüljük az 1. 2...i-edik mintaelemhez tartozó $y_1, y_2, \dots, y_i - t$

Egyenes Gauss-papíron ábrázolva

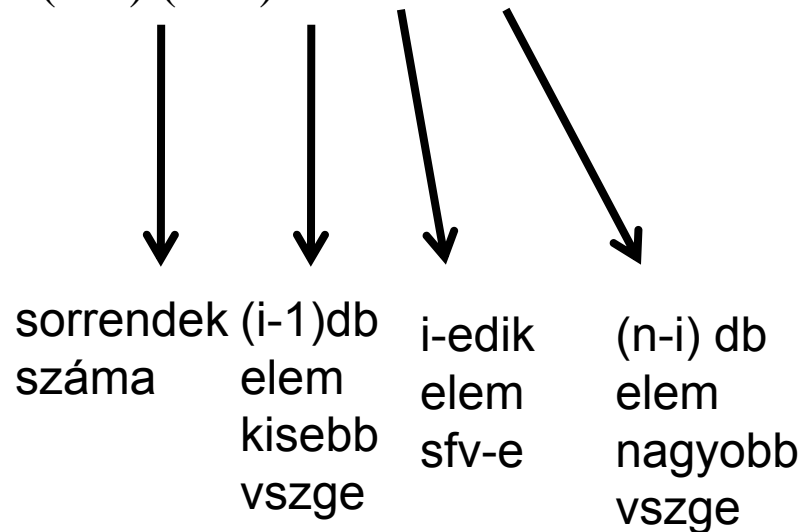
3. Mi az a medián rang?

Az $x_1, <x_2, \dots, <x_i, \dots, <x_n$ rendezett mintaelemek sorszámát a **rang**.

Az ezekhez tartozó y_i eloszlásfüggvény-értékek is rendezett mintát alkotnak, azaz $y_1 < y_2 < \dots < y_i < \dots < y_n$ is rendezett minta, ezek sorszámát is rang.

Az y_i rendezett mintaelem (rangja i) a $[0,1]$ intervallumban egyenletes eloszlású valószínűségi változó, amelynek sűrűségfüggvénye:

$$g_i(y) = \frac{n!}{(i-1)!(n-i)!} y^{i-1} \cdot 1 \cdot (1-y)^{n-i}; (0 \leq y \leq 1).$$



Ennek az eloszlásnak az eloszlásfüggvénye az $y_{\text{medián}}$ helyen veszi fel a 0,5 értéket.

Angol: median rank.

Magyar: **a rang mediánja**.

3.A tapasztalati eloszlásfüggvény szokásos becsléseit részben y_i eloszlásából származtatják.

A szokásos becslések egy része gyakorlati megfontolások alapján a következők:

$$\varphi_1(i) = \frac{i}{n} \quad \varphi_2(i) = \frac{i-1}{n} \quad \varphi_3(i) = \frac{i - \frac{1}{2}}{n} \rightarrow (\text{Montgomery})$$

$$\varphi_4(i) = \frac{i}{n+1}, \quad \text{Ez } y_i \text{ eloszlásának várható értéke.}$$

$$\varphi_5(i) = \frac{i-1}{n-1}, \quad \text{Ez } y_i \text{ eloszlásának módusa.}$$

$$\varphi_6(i) \approx \frac{i-0,3}{n+0,4}. \quad \text{Ez } y_i \text{ eloszlásának közelítő mediánja, vagyis a medián rang.}$$

$$\frac{i}{n+1} < \frac{i-0,3}{n+0,4} < \frac{i-1}{n-1}, \quad \text{ha} \quad \frac{n+1}{2} < i$$

$$g_i(y) = \frac{n!}{(i-1)!(n-i)!} y^{i-1} (1-y)^{n-i}; (0 \leq y \leq 1). \quad \text{Ez } y_i \text{ sűrűségfüggvénye, ebből } G_i(y):$$

$$\tilde{r}_i(y) = \sum_{k=i}^n \binom{n}{k} y^k (1-y)^{n-k} = 1 - \sum_{k=0}^{i-1} \binom{n}{k} y^k (1-y)^{n-k} = 1 - \sum_{k=n-i+1}^n \binom{n}{k} y^k (1-y)^{n-k} = 1 - G_{n+1-i}(1-y)$$

A medián rang közelítő képletének származtatása:

$$G_i(y) = \sum_{k=i}^n \binom{n}{k} y^k (1-y)^{n-k} = 1 - \sum_{k=0}^{i-1} \binom{n}{k} y^k (1-y)^{n-k} = 1 - \sum_{k=n-i+1}^n \binom{n}{k} y^k (1-y)^{n-k} = 1 - G_{n+1-i}(1-y)$$

$$G_i(y) = G_{n+1-i}(1-y).$$

$$\varphi(n-i+1) = 1 - \varphi(i).$$

Keressük a becslést $\varphi(i) = \frac{i-a}{n+b}$ alakban.

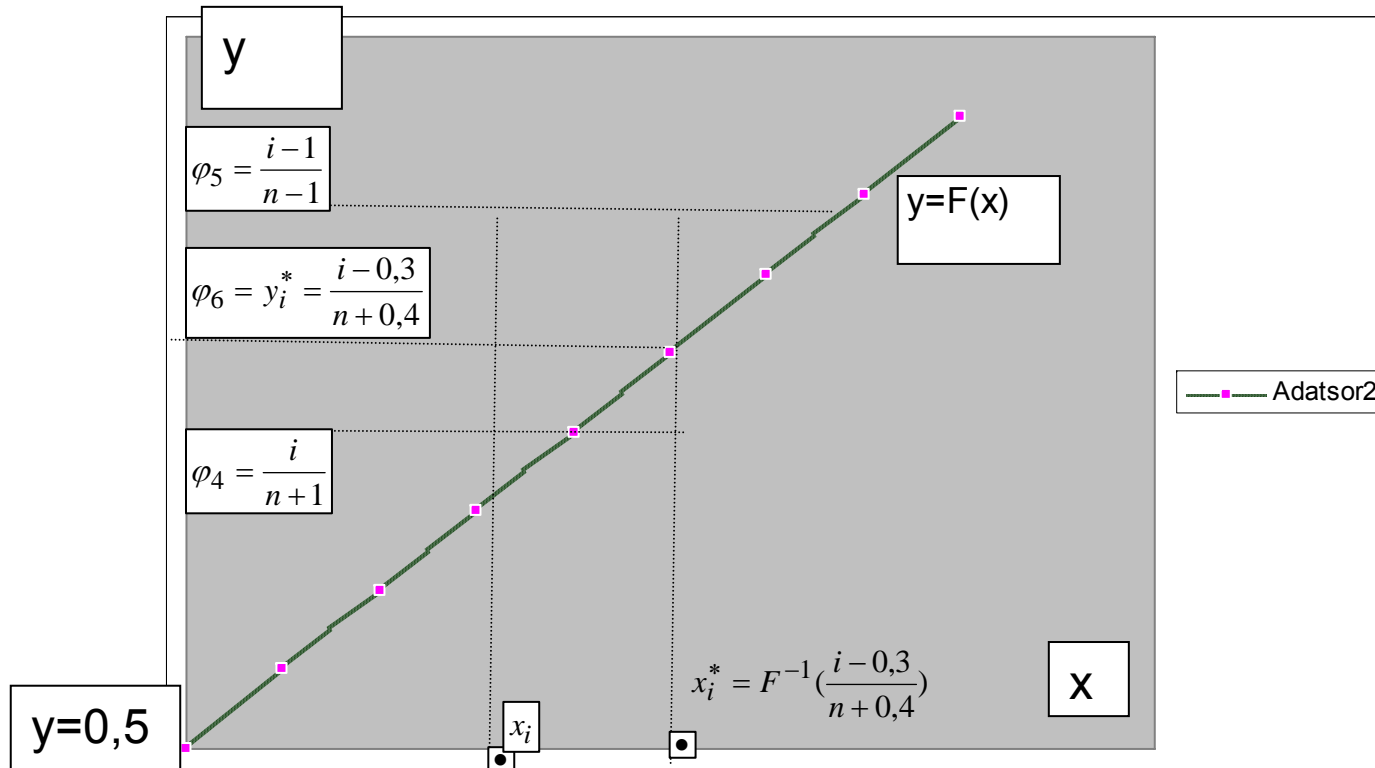
$$1 - G_{n-i+1}(1-y_i^*) = 0,5$$

$$1 - \frac{i-a}{n+b} = \frac{n+1-i+a}{n+b} \rightarrow b=1-2a$$

$$\varphi(i) = \frac{i-a}{n+1-2a} \approx \frac{i-0,3}{n+0,4}$$

$$\sum_{k=0}^{i-1} \binom{n}{k} \left(\frac{i-a}{n+1-2a} \right)^k \left(1 - \frac{i-a}{n+1-2a} \right)^{n-k} = 0,5$$

Ha $n \rightarrow \infty$, akkor a fenti képlet az $(i-a)$ paraméterű Poisson eloszlással közelíthető, és kapjuk, hogy a jó közelítéssel 0,3.



A három becslési módszer ábrázolása

A becslések tulajdonságai és összehasonlításuk:

1. Az $i/(n+1)$ becslés az eseteknek több, mint felében az egyenes alatt van.

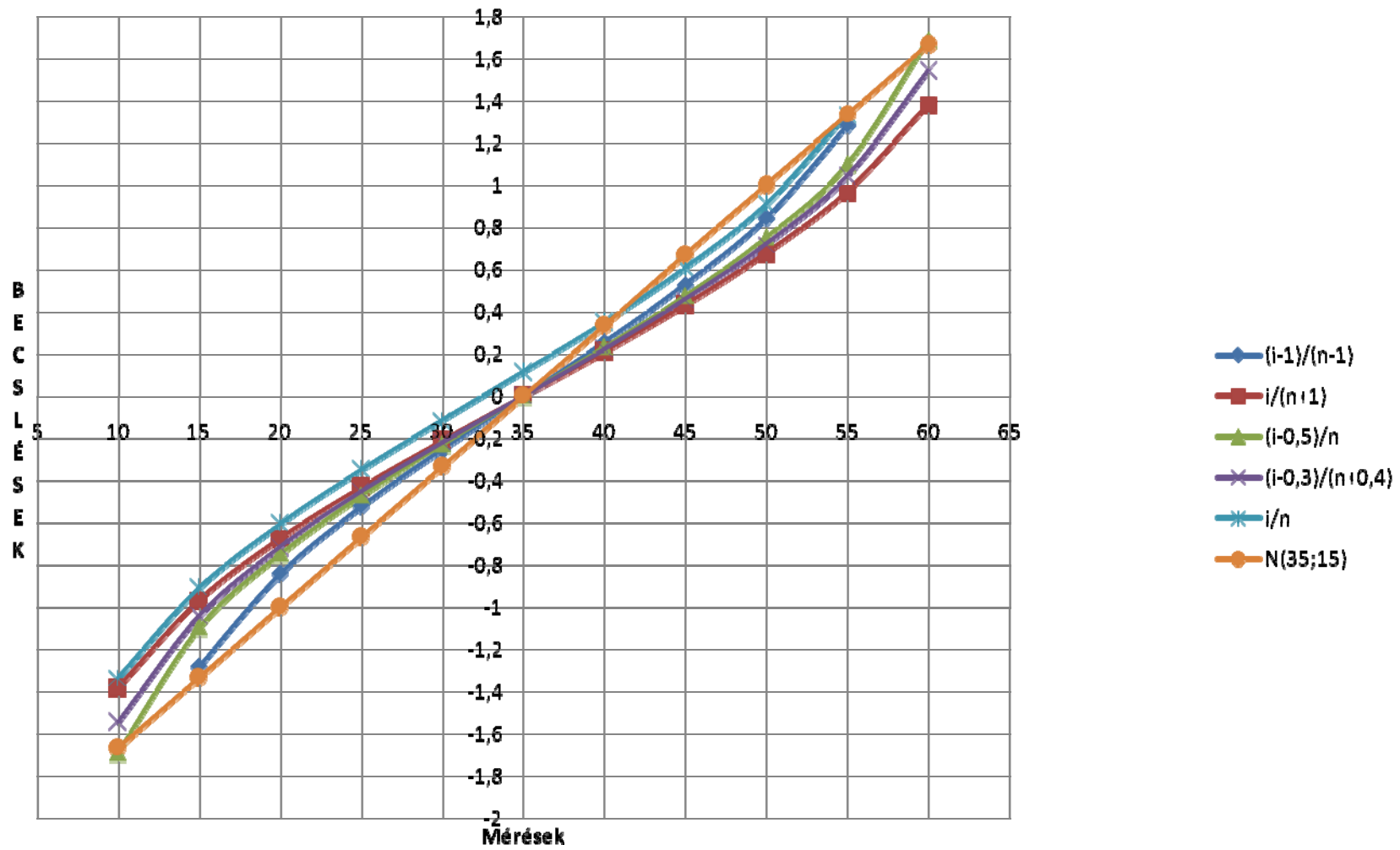
2. Az $(i-1)/(n-1)$ becslés az eseteknek több, mint felében az egyenes felett van.

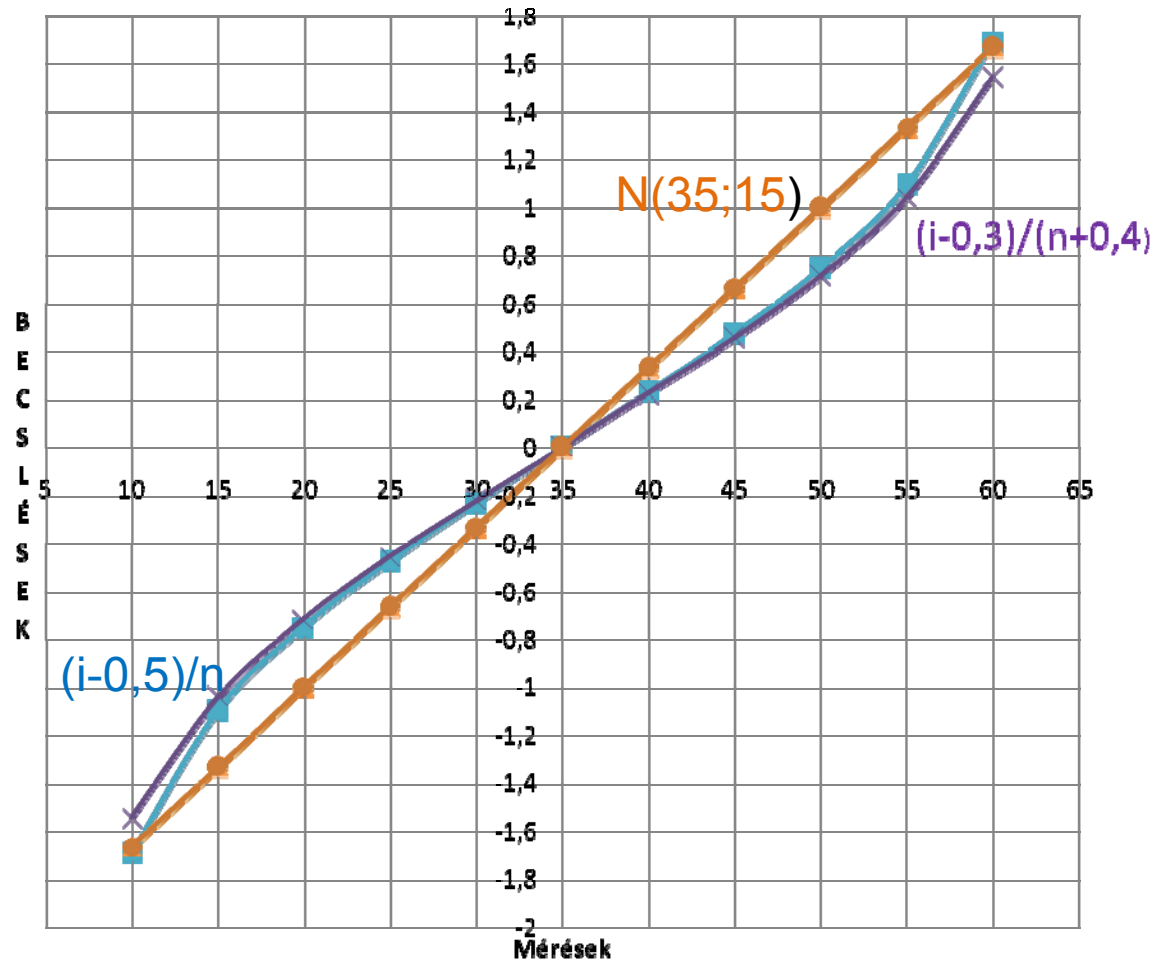
3. Mivel $\frac{i-0,5}{n} > \frac{i-0,3}{n+0,4}, i > \frac{n+1}{2}$, erre is teljesül a fenti megállapítás.

4. Az $(i-0,3)/(n+0,4)$ becslés közel azonos számú esetben van az egyenes alatt és felett.

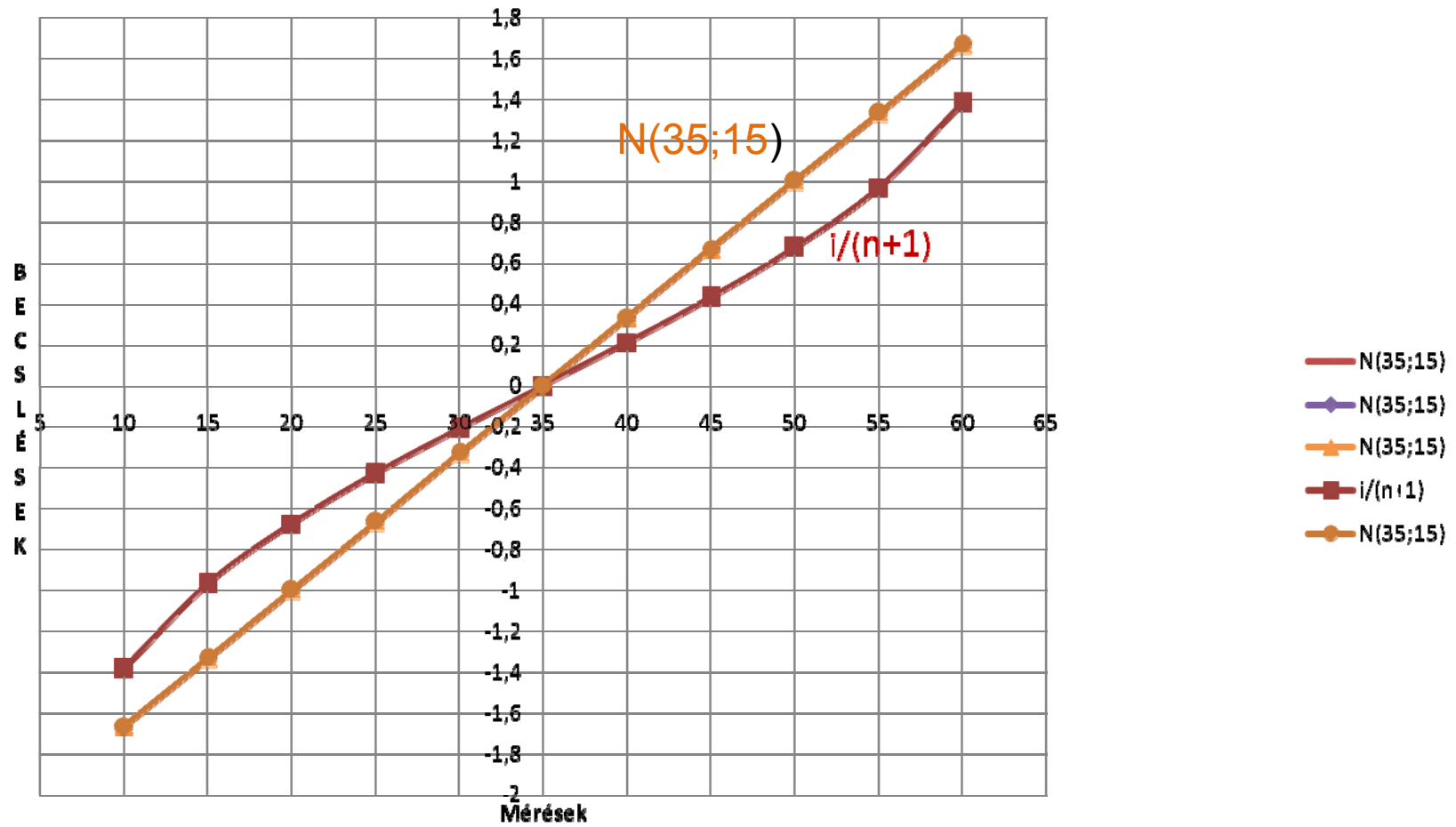
5. A 2. és 3. esetben alábecsülik a normális eloszlás szórását, az 1. esetben pedig túl nagy szórást becsülnek. Ez azért van, mert az egyenes meredeksége fordítottan arányos a szórással. Az egyenes 0,5 ordinátájú pontjához tartozó x érték becsüli a várható értéket, az egyenes meredeksége pedig a szórás reciprok értéke.

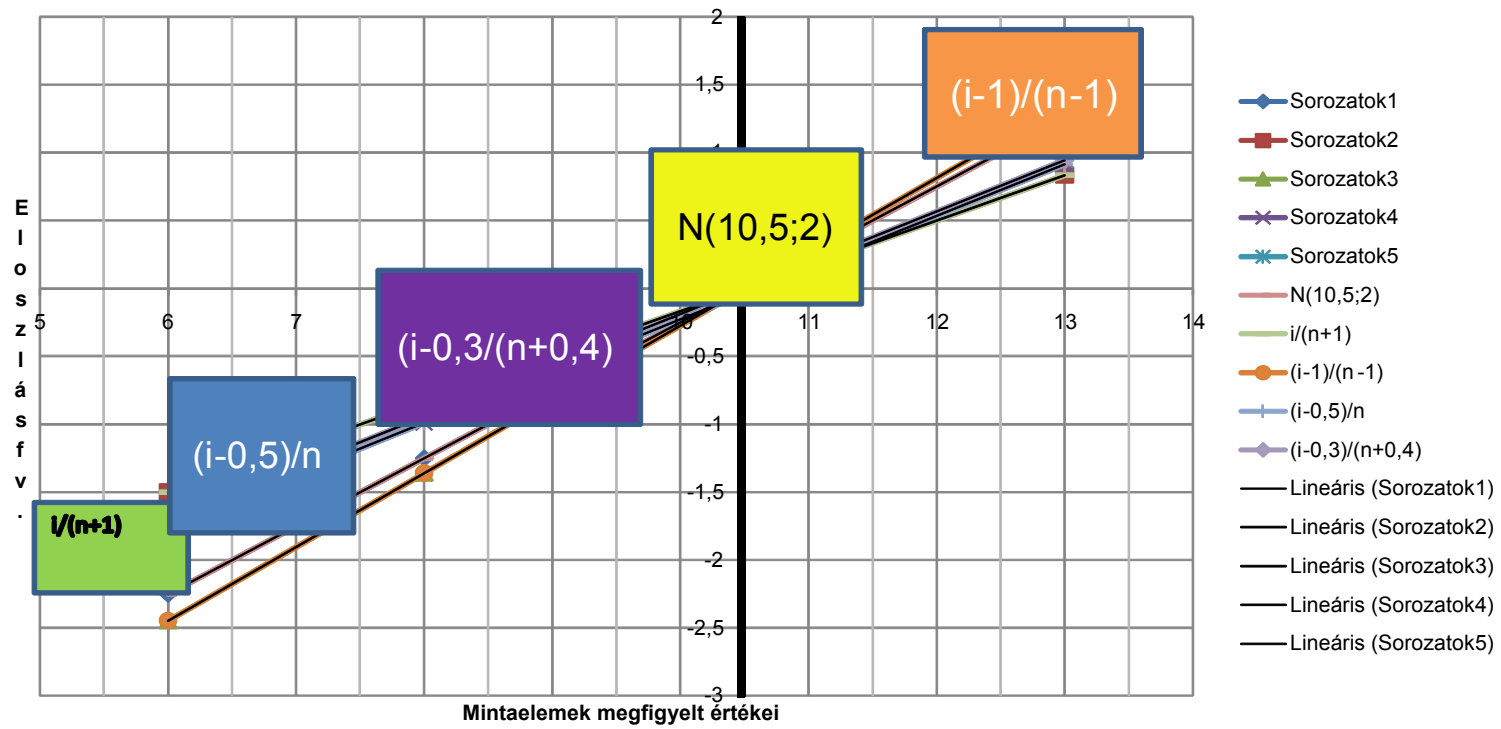
Becslések összehasonlítása

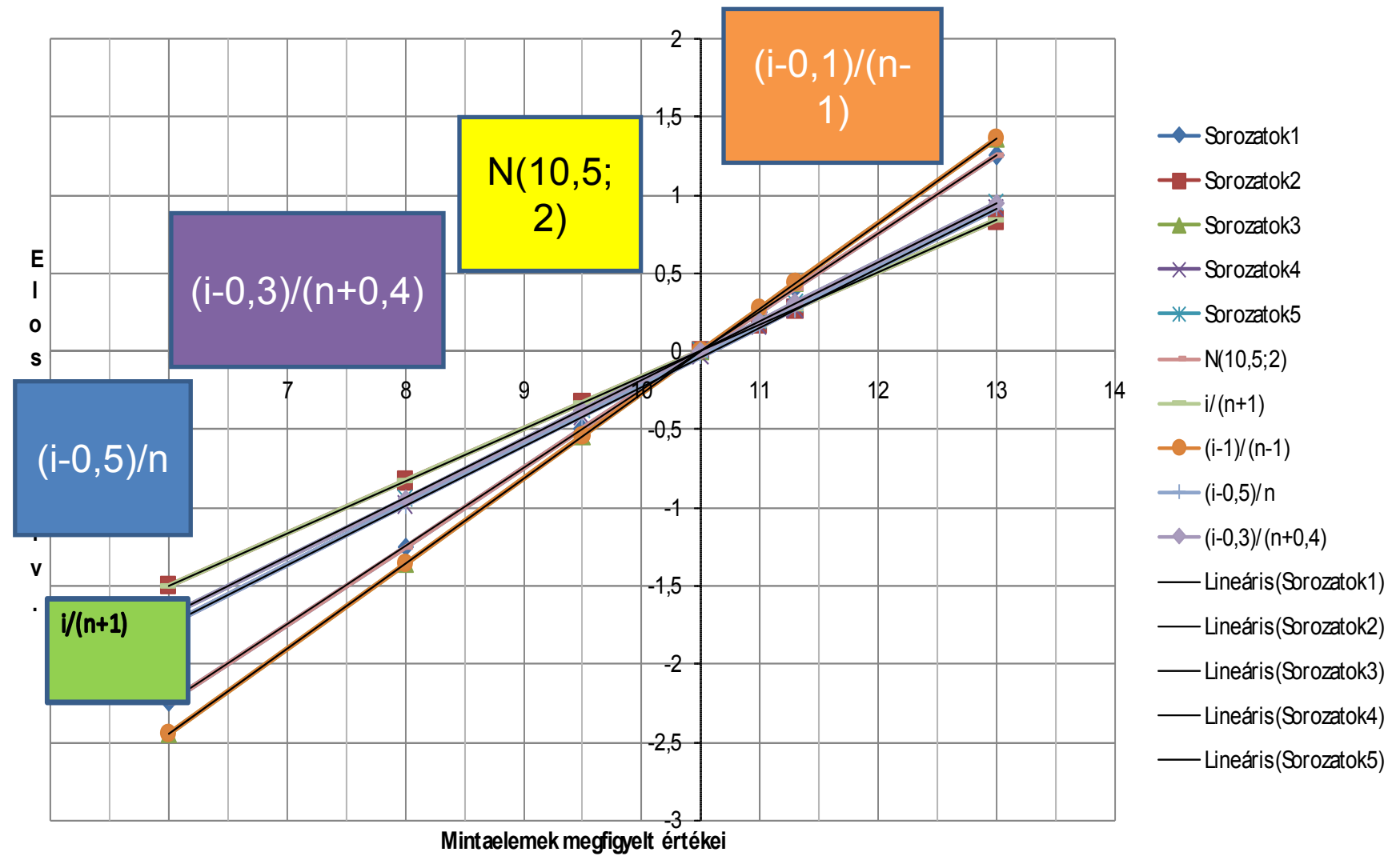




Becklések összehasonlítása







4.A hibaarány 50%-os felső konfidencia határa:

$$C_U = \left[1 - \sum_{k=0}^i \binom{n}{k} p^k \times (1-p)^{n-k} \right] \times 100\%.$$

Binomiális eloszlásból kiszámítva.

$$\hat{p} = \frac{1}{1 + \left(\frac{n-i}{i+1} \right) \times F_{0,50}(2n-2i, 2i+2)},$$

A fenti képletből adódik a pontos képlet.

$$\hat{p} \approx \frac{(i+1) - 0,3}{n + 0,4} = \frac{i + 0,7}{n + 0,4},$$

Ez a közelítő képlet a medián rangból.

Értékelési mód	Gyakorlati megfontolás			Excel (i-1)/(n-1)			Minitab program(i/(n+1))			Montgomery (i -0,5)/n		
	25 %	50 %	75 %	25 %	50 %	75 %	25 %	50 %	75 %	25 %	50 %	75 %
Minta												
1,4,5,6,9,10,12	4	6	10	4,5	6,0	9,5	4	6	10	4,5	6	9.75
1,4,6,9,12,15	4	7,5	12	4,5	7,5	11,25	3,25	7,5	12,75	4	7,5	12

5. Kvartilisek és a medián százalékos értékeinek összehasonlító táblázata

Kvartilisek számítási képletei:

Az Excel a $\hat{p} = \frac{i-1}{n-1}$ képletből indul ki és így $p=1/4$ esetén

$i = \frac{1}{4}(n-1) + 1$; Ennek a számnak egész részét kell venni, ezt a sorszámú tagot kell kiinduló értéknek tekinteni és ehhez hozzá kell adni ennek a számnak a törtrészének és következő mintaelemtől való távolságának szorzatát. $p=3/4$ esetén hasonló az eljárás.

A Minitab a $\hat{p} = \frac{i}{n+1}$ képletből indul ki és így $p=1/4$ -re $i=1/4(n+1)$; ezt követően az eljárás azonos.

$\hat{p} = \frac{i-0,5}{n}$ esetén. ha $p=1/4$, akkor $i=(1/4).n + 0,5$, ezután az eljárás azonos.

Ennek megfelelően az Excel képletei a 25 és 75%-os kvartilisekre, ahol $[x]$ x egész része, $\{x\}$ x törtrésze:

$$X_{0,25} = X_{[w'_1]} + \{w'_1\}(X_{[w'_1+1]} - X_{[w'_1]}); X_{0,75} = X_{[w'_3]} + \{w'_3\}(X_{[w'_3+1]} - X_{[w'_3]})$$

A Minitab képletei:

$$X_{0,25} = X_{[w_1]} + \{w_1\}(X_{[w_1+1]} - X_{[w_1]}); X_{0,75} = X_{[w_3]} + \{w_3\}(X_{[w_3+1]} - X_{[w_3]})$$

$$\begin{aligned}
\{w'\} &= 0,25; ha, n = 4k + 2; excel; \\
\{w'\} &= 0,5; ha, n = 4k + 3; excel; \\
\{w'\} &= 0,75; ha, n = 4k; excel; \\
\{w'\} &= 0; ha, n = 4k + 1; excel; \\
\{w\} &= 0,25; ha, n = 4k; minitab; \\
\{w\} &= 0,5; ha, n = 4k + 1; minitab; \\
\{w\} &= 0,75; ha, n = 4k + 2; minitab; \\
\{w\} &= 0; ha, n = 4k + 3; minitab; \\
|\{w\} - \{w'\}| &= 0,5; minitab; excel
\end{aligned}$$

p- kvantilis mintabeli becslése

p becslése	w értéke	p-kvantilis
$p=i/(n+1)$	$w=(n+1)p$	$x_p = x_{[w]} + \{w\}(x_{[w]+1} - x_{[w]})$
$p=(i-1)/(n-1)$	$w=(n-1)p+1$	$x_p = x_{[w]} + \{w\}(x_{[w]+1} - x_{[w]})$
$p=(i-0,5)/n$	$w=np+0,5$	$x_p = x_{[w]} + \{w\}(x_{[w]+1} - x_{[w]})$
$p=(i-0,3)/(n+0,4)$	$w=(n+0,4)p+0,3$	$x_p = x_{[w]} + \{w\}(x_{[w]+1} - x_{[w]})$

Jelölések: $[w]$ w egész része; $\{w\}$ w tört része; $F(x_p)=p$.