1.2.6. Definíció. *A* $\xi_1, \xi_2, \ldots, \xi_n$ *mintából képezett* ***rendezett minta:***

$$\xi^*_{1:n} \leqslant \xi^*_{2:n} \leqslant \ldots \leqslant \xi^*_{n:n}.$$

A rendezett minta elemei statisztikák, ugyanis például $\xi^*_{1:n} = T_n(\xi_1, \xi_2, \ldots, \xi_n)$, ahol

$$T_n(x_1, x_2, \ldots, x_n) = \min\{x_1, x_2, \ldots, x_n\}$$

folytonos függvény. Megjegyezzük, hogy a $\xi^*_{1:n}, \xi^*_{2:n}, \ldots, \xi^*_{n:n}$ valószínűségi változók nem függetlenek és nem azonos eloszlásúak.

1.2.7. Definíció. *A* $\xi_1, \xi_2, \ldots, \xi_n$ *mintából képezett* ***tapasztalati medián:***

$$\begin{cases} \xi^*_{m+1:n}, & ha \quad n = 2m+1 \quad páratlan, \\ (\xi^*_{m:n} + \xi^*_{m+1:n})/2, & ha \quad n = 2m \quad páros. \end{cases}$$

A $\xi_1, \xi_2, \ldots, \xi_n$ *minta* ***terjedelme:***

$$\xi^*_{n:n} - \xi^*_{1:n}.$$

1.2.8. Definíció. *A* $\xi_1, \xi_2, \ldots, \xi_n$ *mintából képezett* ***tapasztalati (empirikus) eloszlás-függvény:*** $\mathcal{F}_n : \mathbb{R} \times \Omega \to \mathbb{R}$,

$$\mathcal{F}_n(x) = \mathcal{F}_n(x, \omega) := \sum_{i:\, \xi_i(\omega) < x} \frac{1}{n} = \begin{cases} 0, & ha \quad x \leqslant \xi^*_{1:n}, \\ \dfrac{k}{n}, & ha \quad \xi^*_{k:n} < x \leqslant \xi^*_{k+1:n}, \\ 1, & ha \quad x > \xi^*_{n:n}. \end{cases}$$

Vagyis minden rögzített $x \in \mathbb{R}$ esetén $\mathcal{F}_n(x, \cdot) : \Omega \to \mathbb{R}$ valószínűségi változó, ezért statisztika, mégpedig a $\{\omega \in \Omega : \xi_i(\omega) < x\} \in \mathcal{A}$ esemény relatív gyakorisága. Továbbá minden rögzített $\omega \in \Omega$ esetén $\mathcal{F}_n(\cdot, \omega) : \mathbb{R} \to \mathbb{R}$ eloszlásfüggvény, mégpedig annak a diszkrét eloszlásnak az eloszlásfüggvénye, melynek a lehetséges értékei a mintaelemek $\xi_1(\omega), \xi_2(\omega), \ldots, \xi_n(\omega)$ realizációi, és ezen értékeket egyaránt $1/n$ valószínűséggel veheti fel.

2) Glivenko - Cantelli alaptétel: Minden valószínűségi változónak van várható értéke

$E(\xi) = \frac{x_1 + x_2 + \ldots + x_n}{n} = \bar{x}$ $\quad E(x) = m \quad D^2(x) = \sigma^2$

variancia

mintaátlag (empirikus várható érték)

empirikus szórásnégyzet: $D^2(\xi) = \frac{(x_1 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n} = s_n^2$

mintaátlag valsz. vált: $\bar{X} = \frac{X_1 + X_2 + \ldots X_n}{n}$

$E(\bar{x}) = m = E(x)$

középérték

$E(S_n^2) = \frac{n-1}{n} \cdot \sigma^2$ majdnem torzítatlanul becsül.

empirikus szórásnégyzet

korrigált empirikus szórásnégyzet $S_n^{*2} = \frac{n}{n-1} \cdot S_n^2$

Kolmogorov–Szmirnov-tételkör

A Glivenko–Cantelli-tétel arról szól, hogy az empirikus eloszlásfüggvény 1 valószínűséggel (majdnem minden realizációra) az egész számegyenesen egyenletesen tart az elméleti eloszlásfüggvényhez. Tehát kellő számú mintát véve tetszőleges pontossággal közelíteni tudjuk a valódi eloszlásfüggvényt. De adott pontossághoz vajon hány elemű mintát kell vennünk? A konvergencia sebességére vonatkozóan újabb tételeket fogunk kimondani. Ezek azt jelzik, hogy n kísérlet kb. $1/\sqrt{n}$ nagyságrendű közelítéshez elegendő.

Legyen a háttéreloszlás F eloszlásfüggvénye folytonos, F_n^* pedig jelölje az n-elemű mintához tartozó empirikus eloszlásfüggvényt. Akkor

2.1. Tétel (Szmirnov).

$$\lim_{n\to\infty} \mathbb{P}\left(\sqrt{n}\sup_{x\in\mathbb{R}}(F_n^*(x)-F(x))<z\right)=S(z), \qquad \forall z\in\mathbb{R},$$

ahol

$$S(z)=\begin{cases}0, & ha \quad z\le 0,\\ 1-e^{-2z^2}, & ha \quad z>0,\end{cases} \tag{2.1}$$

az ún. Szmirnov-eloszlásfüggvény.

2.2. Tétel (Kolmogorov).

$$\lim_{n\to\infty} \mathbb{P}\left(\sqrt{n}\sup_{x\in\mathbb{R}}|F_n^*(x)-F(x)|<z\right)=K(z), \qquad \forall z\in\mathbb{R},$$

ahol

$$K(z)=\begin{cases}0, & ha \quad z\le 0,\\ \sum_{i=-\infty}^{\infty}(-1)^i e^{-2i^2z^2}=1-2\sum_{i=1}^{\infty}(-1)^{i-1}e^{-2i^2z^2}, & ha \quad z>0,\end{cases} \tag{2.2}$$

az ún. Kolmogorov-eloszlásfüggvény.

Legyen most az X illetve Y háttérváltozó (nem feltétlenül ismert) eloszlásfüggvénye a folytonos F illetve G függvény, F_n^* illetve G_m^* pedig jelölje az n-elemű $X_1,\dots,X_n$ illetve az m-elemű $Y_1,\dots,Y_m$, egymástól is független mintákhoz tartozó empirikus eloszlásfüggvényeket. Tegyük fel továbbá, hogy $F(x)=G(x)$, $\forall x\in\mathbb{R}$. Akkor

2.3. Tétel (Szmirnov).

$$\lim_{n,m\to\infty} \mathbb{P}\left(\sqrt{\frac{nm}{n+m}}\sup_{x\in\mathbb{R}}(F_n^*(x)-G_m^*(x))<z\right)=S(z), \qquad \forall z\in\mathbb{R}.$$

1.2.6. Definíció. *A* $\xi_1, \xi_2, \ldots, \xi_n$ *mintából képezett* ***rendezett minta:***

$$\xi^*_{1:n} \leqslant \xi^*_{2:n} \leqslant \ldots \leqslant \xi^*_{n:n}.$$

A rendezett minta elemei statisztikák, ugyanis például $\xi^*_{1:n} = T_n(\xi_1, \xi_2, \ldots, \xi_n)$, ahol

$$T_n(x_1, x_2, \ldots, x_n) = \min\{x_1, x_2, \ldots, x_n\}$$

folytonos függvény. Megjegyezzük, hogy a $\xi^*_{1:n}, \xi^*_{2:n}, \ldots, \xi^*_{n:n}$ valószínűségi változók nem függetlenek és nem azonos eloszlásúak.

1.2.7. Definíció. *A* $\xi_1, \xi_2, \ldots, \xi_n$ *mintából képezett* ***tapasztalati medián:***

$$\begin{cases} \xi^*_{m+1:n}, & ha \;\; n = 2m+1 \;\; páratlan, \\ (\xi^*_{m:n} + \xi^*_{m+1:n})/2, & ha \;\; n = 2m \;\; páros. \end{cases}$$

A $\xi_1, \xi_2, \ldots, \xi_n$ *minta* ***terjedelme:***

$$\xi^*_{n:n} - \xi^*_{1:n}.$$

1.2.8. Definíció. *A* $\xi_1, \xi_2, \ldots, \xi_n$ *mintából képezett* ***tapasztalati (empirikus) eloszlás-függvény:*** $\mathcal{F}_n : \mathbb{R} \times \Omega \to \mathbb{R}$,

$$\mathcal{F}_n(x) = \mathcal{F}_n(x, \omega) := \sum_{i:\, \xi_i(\omega) < x} \frac{1}{n} = \begin{cases} 0, & ha \;\; x \leqslant \xi^*_{1:n}, \\ \dfrac{k}{n}, & ha \;\; \xi^*_{k:n} < x \leqslant \xi^*_{k+1:n}, \\ 1, & ha \;\; x > \xi^*_{n:n}. \end{cases}$$

Vagyis minden rögzített $x \in \mathbb{R}$ esetén $\mathcal{F}_n(x, \cdot) : \Omega \to \mathbb{R}$ valószínűségi változó, ezért statisztika, mégpedig a $\{\omega \in \Omega : \xi_i(\omega) < x\} \in \mathcal{A}$ esemény relatív gyakorisága. Továbbá minden rögzített $\omega \in \Omega$ esetén $\mathcal{F}_n(\cdot, \omega) : \mathbb{R} \to \mathbb{R}$ eloszlásfüggvény, mégpedig annak a diszkrét eloszlásnak az eloszlásfüggvénye, melynek a lehetséges értékei a mintaelemek $\xi_1(\omega), \xi_2(\omega), \ldots, \xi_n(\omega)$ realizációi, és ezen értékeket egyaránt $1/n$ valószínűséggel veheti fel.

2) Glivenko - Cantelli alaptétel: Minden valószínűségi változónak van várható értéke

$E(\bar{x}) = \frac{x_1 + x_2 + \ldots + x_n}{n} = \bar{x}$ $\quad E(x) = m \quad D^2(x) = \sigma^2$

variancia

mintaátlag (empirikus várható érték)

empirikus szórásnégyzet: $D^2(\bar{x}) = \frac{(x_1-\bar{x})^2 + \ldots + (x_n-\bar{x})^2}{n} = s_n^2$

mintaátlag valószn. vált: $\bar{X} = \frac{X_1 + X_2 + \ldots X_n}{n}$

$E(\bar{x}) = m = E(x)$

középérték

$E(S_n^2) = \frac{n-1}{n} \cdot \sigma^2$ majdnem torzítatlanul becsül.

empirikus szórásnégyzet

korrigált empirikus szórásnégyzet $S_n^{*2} = \frac{n}{n-1} \cdot S_n^2$

Kolmogorov–Szmirnov-tételkör

A Glivenko–Cantelli-tétel arról szól, hogy az empirikus eloszlásfüggvény 1 valószínűséggel (majdnem minden realizációra) az egész számegyenesen egyenletesen tart az elméleti eloszlásfüggvényhez. Tehát kellő számú mintát véve tetszőleges pontossággal közelíteni tudjuk a valódi eloszlásfüggvényt. De adott pontossághoz vajon hány elemű mintát kell vennünk? A konvergencia sebességére vonatkozóan újabb tételeket fogunk kimondani. Ezek azt jelzik, hogy n kísérlet kb. $1/\sqrt{n}$ nagyságrendű közelítéshez elegendő.

Legyen a háttéreloszlás F eloszlásfüggvénye folytonos, F_n^* pedig jelölje az n-elemű mintához tartozó empirikus eloszlásfüggvényt. Akkor

2.1. Tétel (Szmirnov).

$$\lim_{n\to\infty} \mathbb{P}\left(\sqrt{n}\sup_{x\in\mathbb{R}}(F_n^*(x) - F(x)) < z\right) = S(z), \qquad \forall z \in \mathbb{R},$$

ahol

$$S(z) = \begin{cases} 0, & ha \quad z \leq 0, \\ 1 - e^{-2z^2}, & ha \quad z > 0, \end{cases} \tag{2.1}$$

az ún. Szmirnov-eloszlásfüggvény.

2.2. Tétel (Kolmogorov).

$$\lim_{n\to\infty} \mathbb{P}\left(\sqrt{n}\sup_{x\in\mathbb{R}}|F_n^*(x) - F(x)| < z\right) = K(z), \qquad \forall z \in \mathbb{R},$$

ahol

$$K(z) = \begin{cases} 0, & ha \quad z \leq 0, \\ \sum_{i=-\infty}^{\infty}(-1)^i e^{-2i^2z^2} = 1 - 2\sum_{i=1}^{\infty}(-1)^{i-1}e^{-2i^2z^2}, & ha \quad z > 0, \end{cases} \tag{2.2}$$

az ún. Kolmogorov-eloszlásfüggvény.

Legyen most az X illetve Y háttérváltozó (nem feltétlenül ismert) eloszlásfüggvénye a folytonos F illetve G függvény, F_n^* illetve G_m^* pedig jelölje az n-elemű $X_1,\dots,X_n$ illetve az m-elemű $Y_1,\dots,Y_m$, egymástól is független mintákhoz tartozó empirikus eloszlásfüggvényeket. Tegyük fel továbbá, hogy $F(x) = G(x)$, $\forall x \in \mathbb{R}$. Akkor

2.3. Tétel (Szmirnov).

$$\lim_{n,m\to\infty} \mathbb{P}\left(\sqrt{\frac{nm}{n+m}}\sup_{x\in\mathbb{R}}(F_n^*(x) - G_m^*(x)) < z\right) = S(z), \qquad \forall z \in \mathbb{R}.$$