A sztochasztika alapjai
MBNXK262

7. előadás: Huffman kód; szórás, kovariancia

Kevei Péter

2025/26 tavasz

Huffman-kód

$\mathcal{X} = \{x_1, x_2, \ldots, x_n\}$ véges halmaz, a *forrásábécé*.
Kód $f : \mathcal{X} \to \{$véges 0-1 sorozatok $\}$
Az f-hez tartozó lehetséges kódszavak $f(x_1), f(x_2), \ldots, f(x_n)$.

Huffman-kód

$\mathcal{X} = \{x_1, x_2, \ldots, x_n\}$ véges halmaz, a *forrásábécé*.
Kód $f : \mathcal{X} \to \{$véges 0-1 sorozatok $\}$
Az f-hez tartozó lehetséges kódszavak $f(x_1), f(x_2), \ldots, f(x_n)$.
Az f kód *prefix*, ha a lehetséges kódszavak közül egyik sem folytatása a másiknak. Jelölje $x \in \mathcal{X}$ esetén $|f(x)|$ a kódszó hosszát.

Legyen X egy véletlen betű, és eloszlása $\mathbf{P}(X = x_k) = p_k$, $k = 1, 2, \ldots, n$. Tehát p_k a x_k betű gyakorisága az adott nyelvben.

Legyen X egy véletlen betű, és eloszlása $\mathbf{P}(X = x_k) = p_k$, $k = 1, 2, \ldots, n$. Tehát p_k a x_k betű gyakorisága az adott nyelvben.

Adott f kód esetén egy hosszú szövegben az egy karakterre eső átlagos kódszóhossz:

Feltehető, hogy $p_1 \geq p_2 \geq \ldots \geq p_n$. Ha az f prefix kód optimális, akkor feltehető, hogy teljesülnek a következők:

(i) Hosszabb kódhoz ritkább betűk tartoznak, azaz

$$|f(x_1)| \leq |f(x_2)| \leq \ldots \leq |f(x_n)|.$$

Feltehető, hogy $p_1 \geq p_2 \geq \ldots \geq p_n$. Ha az f prefix kód optimális, akkor feltehető, hogy teljesülnek a következők:

(i) Hosszabb kódhoz ritkább betűk tartoznak, azaz

$$|f(x_1)| \leq |f(x_2)| \leq \ldots \leq |f(x_n)|.$$

(ii) A két legkisebb valószínűséghez tartozó kód hossza egyenlő.

Feltehető, hogy $p_1 \geq p_2 \geq \ldots \geq p_n$. Ha az f prefix kód optimális, akkor feltehető, hogy teljesülnek a következők:

(i) Hosszabb kódhoz ritkább betűk tartoznak, azaz

$$|f(x_1)| \leq |f(x_2)| \leq \ldots \leq |f(x_n)|.$$

(ii) A két legkisebb valószínűséghez tartozó kód hossza egyenlő.

(iii) $f(x_{n-1})$ és $f(x_n)$ csak az utolsó bitben térnek el.

Tétel

Tegyük fel, hogy az

$$\mathcal{X}' = \{x_1, \ldots, x_{n-2}, y_{n-1}\}$$

$(n-1)$ *elemű forrásábécé és* $p_1, \ldots, p_{n-2}, p_{n-1} + p_n$ *eloszlás esetén* g *egy optimális prefix kód. Ekkor az eredeti problémához tartozó optimális prefix kódot kapunk, ha az* x_{n-1}*, ill.* x_n *kódját úgy választjuk, hogy a* $g(y_{n-1})$ *kódszót kiegészítjük 0-val, ill. 1-gyel, a többi kódszót változatlanul hagyjuk.*

Példa

Legyen $n = 6$, $\mathcal{X} = \{x_1, \ldots, x_6\}$, és $p_1 = 0.132$, $p_2 = 0.329$, $p_3 = 0.329$, $p_4 = 0.165$, $p_5 = 0.041$, $p_6 = 0.004$.

(i) $x_5, x_6 \to x_{56}$, $p_{56} = 0.045$;

Példa

Legyen $n = 6$, $\mathcal{X} = \{x_1, \ldots, x_6\}$, és $p_1 = 0.132$, $p_2 = 0.329$, $p_3 = 0.329$, $p_4 = 0.165$, $p_5 = 0.041$, $p_6 = 0.004$.

(i) $x_5, x_6 \to x_{56}$, $p_{56} = 0.045$;

(ii) x_1, x_{56}, $p_{156} = 0.177$;

Példa

Legyen $n = 6$, $\mathcal{X} = \{x_1, \ldots, x_6\}$, és $p_1 = 0.132$, $p_2 = 0.329$, $p_3 = 0.329$, $p_4 = 0.165$, $p_5 = 0.041$, $p_6 = 0.004$.

(i) $x_5, x_6 \to x_{56}$, $p_{56} = 0.045$;

(ii) x_1, x_{56}, $p_{156} = 0.177$;

(iii) x_{156}, x_4, $p_{1564} = 0.342$;

Példa

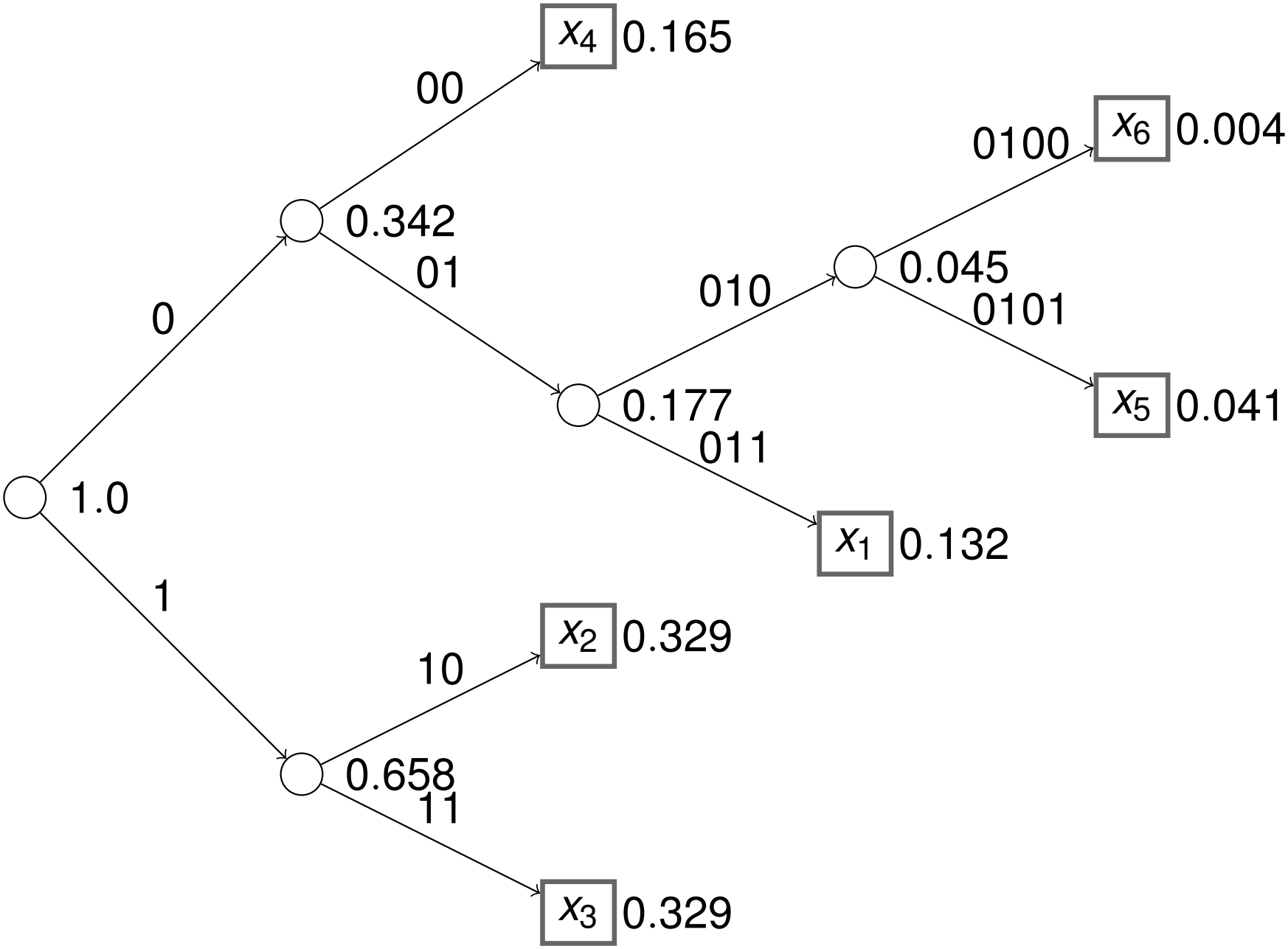
Legyen $n = 6$, $\mathcal{X} = \{x_1, \ldots, x_6\}$, és $p_1 = 0.132$, $p_2 = 0.329$, $p_3 = 0.329$, $p_4 = 0.165$, $p_5 = 0.041$, $p_6 = 0.004$.

(i) $x_5, x_6 \to x_{56}$, $p_{56} = 0.045$;

(ii) x_1, x_{56}, $p_{156} = 0.177$;

(iii) x_{156}, x_4, $p_{1564} = 0.342$;

(iv) x_2, x_3, $p_{23} = 0.658$.


x_4 0.165
00
x_6 0.004
0100
0.342
01
0.045
010
0101
0
0.177
x_5 0.041
011
1.0
x_1 0.132
1
x_2 0.329
10
0.658
11
x_3 0.329

Így az optimális kód:

x_1	x_2	x_3	x_4	x_5	x_6
011	10	11	00	0101	0100
0.132	0.329	0.329	0.165	0.041	0.004

A várható érték:

$$\begin{aligned}\mathbf{E}(f(X)) &= 0.132 \cdot 3 + 0.329 \cdot 2 + 0.329 \cdot 2 \\ &\quad + 0.165 \cdot 2 + 0.041 \cdot 4 + 0.004 \cdot 4 = 2.22\end{aligned}$$

Entrópia

Az optimális várható kódhosszra teljesül, hogy

$$\sum_{k=1}^{n} p_k \log_2 \frac{1}{p_k} \leq \mathbf{E}(|f(\xi)|) < \sum_{k=1}^{n} p_k \log_2 \frac{1}{p_k} + 1.$$

Pl.: JPEG, MP3.

Momentumok

Definíció

Az ξ véletlen változó k-adik momentuma $\mathbf{E}(\xi^k)$, és k-adik centrális momentuma $\mathbf{E}[(\xi - \mathbf{E}\xi)^k]$, $k = 1, 2, \ldots$.

$$\mathbf{E}(\xi^k) = \begin{cases} \sum_i x_i^k \mathbf{P}(\xi = x_i), & \text{ha } \xi \text{ diszkrét,} \\ \int_{-\infty}^{\infty} x^k f(x) \mathrm{d}x, & \text{ha } \xi \text{ folytonos.} \end{cases}$$

Szórás

Definíció

Az ξ véletlen változó szórása $\mathbf{D}(\xi) = \sqrt{\mathbf{E}(\xi - \mathbf{E}(\xi))^2}$.

Szórás

Definíció

Az ξ véletlen változó szórása $\mathbf{D}(\xi) = \sqrt{\mathbf{E}(\xi - \mathbf{E}(\xi))^2}$.

Állítás (Szórás tulajdonságai)

Tetszőleges ξ véletlen változó és a, b valós számok esetén

(i) $\mathbf{D}^2(\xi) = \mathbf{E}(\xi^2) - (\mathbf{E}(\xi))^2$;

Szórás

Definíció

Az ξ véletlen változó szórása $\mathbf{D}(\xi) = \sqrt{\mathbf{E}(\xi - \mathbf{E}(\xi))^2}$.

Állítás (Szórás tulajdonságai)

Tetszőleges ξ véletlen változó és a, b valós számok esetén

(i) $\mathbf{D}^2(\xi) = \mathbf{E}(\xi^2) - (\mathbf{E}(\xi))^2$;

(ii) $\mathbf{D}^2(a\xi + b) = a^2\mathbf{D}^2(\xi)$;

Szórás

Definíció

Az ξ véletlen változó szórása $\mathbf{D}(\xi) = \sqrt{\mathbf{E}(\xi - \mathbf{E}(\xi))^2}$.

Állítás (Szórás tulajdonságai)

Tetszőleges ξ véletlen változó és a, b valós számok esetén

(i) $\mathbf{D}^2(\xi) = \mathbf{E}(\xi^2) - (\mathbf{E}(\xi))^2$;

(ii) $\mathbf{D}^2(a\xi + b) = a^2\mathbf{D}^2(\xi)$;

(iii) $\mathbf{D}(\xi) = 0$ *akkor és csak akkor, ha* $\xi = \mathbf{E}(\xi)$, *azaz* ξ *konstans véletlen változó.*

Véletlen vektorváltozók

Definíció

$\xi = (\xi_1, \ldots, \xi_n) : \Omega \to \mathbb{R}^n$ véletlen vektorváltozó, ha minden komponense véletlen változó. Eloszlásfüggvénye

$$F(x_1, \ldots, x_n) = \mathbf{P}(\xi_1 < x_1, \ldots, \xi_n < x_n).$$

$(\xi_1, \ldots, \xi_n)$ véletlen vektorváltozó diszkrét, ha értékkészlete megszámlálható.
ξ_i, $i = 1, 2, \ldots, n$, eloszlása a peremeloszlás, vagy marginális eloszlás

Trinomiális eloszlás

Szabályos dobókockával n-szer dobunk. ξ a hatosok, η egyesek száma

Függetlenség

Definíció

$\xi_1, \ldots, \xi_n$ függetlenek, ha minden $x_1, \ldots, x_n \in \mathbb{R}$ esetén

$$\mathbf{P}(\xi_1 < x_1, \ldots, \xi_n < x_n) = \mathbf{P}(\xi_1 < x_1) \ldots \mathbf{P}(\xi_n < x_n).$$

Függetlenség

Definíció

$\xi_1, \ldots, \xi_n$ függetlenek, ha minden $x_1, \ldots, x_n \in \mathbb{R}$ esetén

$$\mathbf{P}(\xi_1 < x_1, \ldots, \xi_n < x_n) = \mathbf{P}(\xi_1 < x_1) \ldots \mathbf{P}(\xi_n < x_n).$$

Állítás

$\xi_1, \ldots, \xi_n$ diszkrét véletlen változók úgy, hogy ξ_i lehetséges értékei $x_1^{(i)}, x_2^{(i)}, \ldots$, $i = 1, 2, \ldots, n$. Ekkor $\xi_1, \ldots, \xi_n$ pontosan akkor függetlenek, ha

$$\mathbf{P}(\xi_1 = x_{i_1}^{(1)}, \ldots, \xi_n = x_{i_n}^{(n)}) = \mathbf{P}(\xi_1 = x_{i_1}^{(1)}) \ldots \mathbf{P}(\xi_n = x_{i_n}^{(n)})$$

tetszőleges $i_i, \ldots, i_n$ indexekre.

Állítás

ξ, η független diszkrét véletlen változók.

$$\mathbf{E}\left(g_1(\xi)g_2(\eta)\right) = \mathbf{E}\left(g_1(\xi)\right)\mathbf{E}\left(g_2(\eta)\right).$$

Speciálisan, ha ξ és η függetlenek, akkor $\mathbf{E}(\xi\eta) = \mathbf{E}(\xi)\mathbf{E}(\eta)$.

Kovariancia, korreláció

Definíció

Az ξ és η véletlen változók *kovarianciája*

$$\mathbf{Cov}(\xi, \eta) = \mathbf{E}[(\xi - \mathbf{E}(\xi))(\eta - \mathbf{E}(\eta))],$$

korrelációja

$$\rho(\xi, \eta) = \frac{\mathbf{Cov}(\xi, \eta)}{\mathbf{D}(\xi)\mathbf{D}(\eta)}.$$

Tulajdonságok

Állítás

Tetszőleges $\xi, \xi_1, \ldots, \xi_n, \eta, \eta_1, \ldots, \eta_m$ *véletlen változók és* a, b *valós számok esetén igazak az alábbiak.*

(i) $\mathbf{Cov}(\xi, \xi) = \mathbf{D}^2(\xi)$;

Tulajdonságok

Állítás

Tetszőleges $\xi, \xi_1, \ldots, \xi_n, \eta, \eta_1, \ldots, \eta_m$ *véletlen változók és* a, b *valós számok esetén igazak az alábbiak.*

(i) $\mathbf{Cov}(\xi, \xi) = \mathbf{D}^2(\xi)$;

(ii) $\mathbf{Cov}(\xi, \eta) = \mathbf{Cov}(\eta, \xi)$;

Tulajdonságok

Állítás

Tetszőleges $\xi, \xi_1, \ldots, \xi_n, \eta, \eta_1, \ldots, \eta_m$ *véletlen változók és* a, b *valós számok esetén igazak az alábbiak.*

(i) $\mathbf{Cov}(\xi, \xi) = \mathbf{D}^2(\xi)$;

(ii) $\mathbf{Cov}(\xi, \eta) = \mathbf{Cov}(\eta, \xi)$;

(iii) $\mathbf{Cov}(a(\xi + c), b(\eta + d)) = ab\mathbf{Cov}(\xi, \eta)$;

Tulajdonságok

Állítás

Tetszőleges $\xi, \xi_1, \ldots, \xi_n, \eta, \eta_1, \ldots, \eta_m$ *véletlen változók és* a, b *valós számok esetén igazak az alábbiak.*

(i) $\mathbf{Cov}(\xi, \xi) = \mathbf{D}^2(\xi)$;

(ii) $\mathbf{Cov}(\xi, \eta) = \mathbf{Cov}(\eta, \xi)$;

(iii) $\mathbf{Cov}(a(\xi + c), b(\eta + d)) = ab\mathbf{Cov}(\xi, \eta)$;

(iv) $\mathbf{Cov}\left(\sum_{i=1}^{n} \xi_i, \sum_{j=1}^{m} \eta_j\right) = \sum_{i=1}^{n} \sum_{j=1}^{m} \mathbf{Cov}(\xi_i, \eta_j)$;

Tulajdonságok

Állítás

Tetszőleges $\xi, \xi_1, \ldots, \xi_n, \eta, \eta_1, \ldots, \eta_m$ *véletlen változók és* a, b *valós számok esetén igazak az alábbiak.*

(i) $\mathbf{Cov}(\xi, \xi) = \mathbf{D}^2(\xi)$;

(ii) $\mathbf{Cov}(\xi, \eta) = \mathbf{Cov}(\eta, \xi)$;

(iii) $\mathbf{Cov}(a(\xi + c), b(\eta + d)) = ab\mathbf{Cov}(\xi, \eta)$;

(iv) $\mathbf{Cov}\left(\sum_{i=1}^{n} \xi_i, \sum_{j=1}^{m} \eta_j\right) = \sum_{i=1}^{n} \sum_{j=1}^{m} \mathbf{Cov}(\xi_i, \eta_j)$;

(v) *ha* ξ *és* η *függetlenek, akkor* $\mathbf{Cov}(\xi, \eta) = 0$.

Állítás

(i) *Bunyakovszkij–Cauchy–Schwarz-egyenlőtlenség:*

$$|\mathbf{Cov}(\xi, \eta)| \leq \mathbf{D}(\xi)\mathbf{D}(\eta).$$

Innen adódik, hogy $\rho(\xi, \eta) \in [-1, 1]$*;*

Állítás

(i) *Bunyakovszkij–Cauchy–Schwarz-egyenlőtlenség:*

$$|\mathbf{Cov}(\xi, \eta)| \leq \mathbf{D}(\xi)\mathbf{D}(\eta).$$

Innen adódik, hogy $\rho(\xi, \eta) \in [-1, 1]$*;*

(ii) *ha* $\rho(\xi, \eta) = 1$*, akkor*

$$\xi = \mathbf{E}(\xi) + \frac{\mathbf{D}(\xi)}{\mathbf{D}(\eta)}(\eta - \mathbf{E}(\eta));$$

Állítás

(i) *Bunyakovszkij–Cauchy–Schwarz-egyenlőtlenség:*

$$|\mathbf{Cov}(\xi,\eta)| \le \mathbf{D}(\xi)\mathbf{D}(\eta).$$

Innen adódik, hogy $\rho(\xi,\eta) \in [-1,1]$*;*

(ii) *ha* $\rho(\xi,\eta) = 1$*, akkor*

$$\xi = \mathbf{E}(\xi) + \frac{\mathbf{D}(\xi)}{\mathbf{D}(\eta)}(\eta - \mathbf{E}(\eta));$$

(iii) *ha* $\rho(\xi,\eta) = -1$*, akkor*

$$\xi = \mathbf{E}(\xi) - \frac{\mathbf{D}(\xi)}{\mathbf{D}(\eta)}(\eta - \mathbf{E}(\eta)).$$

Összeg szórásnégyzete

Legyenek ξ, η véletlen változók, ρ a korrelációjuk. Ekkor

$$\mathbf{D}^2(\xi + \eta) =$$

Összeg szórásnégyzete

Ha ξ és η függetlenek, akkor $\rho = 0$, így

$$\mathbf{D}^2(\xi + \eta) =$$

Összeg szórásnégyzete

Ha ξ és η függetlenek, akkor $\rho = 0$, így

$$\mathbf{D}^2(\xi + \eta) = \mathbf{D}^2(\xi) + \mathbf{D}^2(\eta).$$

Összeg szórásnégyzete

Ha ξ és η függetlenek, akkor $\rho = 0$, így

$$\mathbf{D}^2(\xi + \eta) = \mathbf{D}^2(\xi) + \mathbf{D}^2(\eta).$$

Indukcióval, ha $\xi_1, \xi_2, \ldots, \xi_n$ páronként független (korrelálatlan) véletlen változók, akkor

$$\mathbf{D}^2\left(\sum_{i=1}^{n} \xi_i\right) = \sum_{i=1}^{n} \mathbf{D}^2(\xi_i).$$

Kovariancia

Példa

Egy szabályos dobókockával n-szer dobunk. Jelölje ξ a hatosok, η egyesek számát! Adjuk meg a várható értéket, szórást, kovarianciát, korrelációt!

Lineáris regresszió felé

Példa

3, külsőre egyforma érmével a fejdobás valószínűsége $1/4, 2/4, 3/4$. Véletlenszerűen választunk egy érmét, és azzal kétszer dobunk. Legyen η a választott érmével dobva a fej valószínűsége, ξ a dobott fejek száma. ξ-ből szeretnénk η értékére következtetni.

Lineáris regresszió felé

(ξ, η) együttes eloszlása:

$\eta \backslash \xi$	0	1	2	$\sum$
$\frac{1}{4}$	$\frac{9}{48}$	$\frac{6}{48}$	$\frac{1}{48}$	$\frac{1}{3}$
$\frac{1}{2}$	$\frac{1}{12}$	$\frac{1}{6}$	$\frac{1}{12}$	$\frac{1}{3}$
$\frac{3}{4}$	$\frac{1}{48}$	$\frac{6}{48}$	$\frac{9}{48}$	$\frac{1}{3}$
$\sum$	$\frac{14}{48}$	$\frac{20}{48}$	$\frac{14}{48}$	1

Lineáris regresszió

(ξ, η) véletlen vektorváltozó. Az η változót tekintem *függő* változónak, ennek az értékére szeretnék következtetni a ξ *független változó* értékéből. Vagyis ismert ξ esetén szeretném megmondani η-t. Keressük azokat az a, b valós számokat, melyre a $\eta - (a\xi + b)$ változó kicsi. A kicsiséget négyzetes hibában mérve, keressük az

$$h(a, b) = \mathbf{E}\left[(\eta - (a\xi + b))^2\right]$$

függvény minimumhelyét, azaz a legjobb a, b választást.

Lineáris regresszió

Ezek szerint a legjobb lineáris közelítést a

$$g(x) = \frac{\mathbf{Cov}(\eta, \xi)}{\mathbf{D}^2(\xi)}(x - \mathbf{E}(\xi)) + \mathbf{E}(\eta)$$

függvény adja. Ő a *regressziós egyenes*.
Ha ξ és η korrelálatlanok, azaz $\mathbf{Cov}(\xi, \eta) = 0$, akkor a legjobb közelítés $\mathbf{E}(\eta)$, vagyis ξ semmi információt nem ad η értékéről.