Poprzednim razem

$$\boxed{\begin{aligned}&cov(X,Y)=E\left[(X-E(X))(Y-E(Y))\right]=\\&=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}(x-E(X))(y-E(Y))f_{X,Y}(x,y)\,dx\,dy\end{aligned}}$$

$$\boxed{\rho_{X,Y}\equiv\rho(X,Y)\equiv corr(X,Y)=\frac{cov(X,Y)}{\sqrt{var(X)var(Y)}}}$$

$$\boxed{-1\le\rho_{X,Y}\le 1}$$

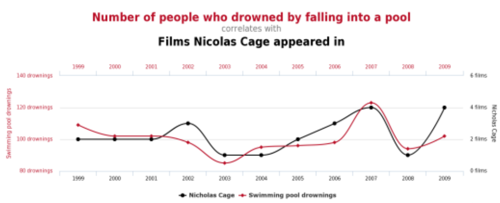
$$\boxed{\begin{aligned}&\rho_{X,Y}=1\quad dla\quad Y=aX+b\quad a>0\\&\rho_{X,Y}=-1\quad dla\quad Y=aX+b\quad a<0\end{aligned}}$$

1

http://tylervigen.com/spurious-correlations

2

http://tylervigen.com/spurious-correlations



Korelacje przypadkowe (fałszywe) czy prawdziwe?

Analiza ekspercka
Zmienne współzależne: X→Y, X→Z prowadzi do korelacji (Y,Z)
Zmienne pośredniczące: X→Y→Z prowadzi do korelacji (X,Z)
Analiza zebrania próby: zmienne kontrolne, badania z randomizacją, dobór próby
Analiza wyników: grupowanie, analiza regresji, analiza wariancji, sieci zależności

3

Wektor losowy – przypadek n>2

- Syt.: dany jest n-wymiarowy wektor losowy $\vec{X}$, z którego tworzymy wielowymiarową zmienną losową $\vec{Y}$:

$$\vec{Y}=\vec{Y}(\vec{X})$$

rozwijamy wokół Y=E(X), zostawiając człony liniowe

$$Y_i=\underbrace{Y_i(E(\vec{X}))}_{E(Y_i)}+\sum_j\left.\left(\frac{\partial Y_i}{\partial X_j}\right)\right|_{\vec{X}=E(\vec{X})}\left(X_j-E(X_j)\right)+\ldots$$

$$Y_i-E(Y_i)=\sum_j\left.\left(\frac{\partial Y_i}{\partial X_j}\right)\right|_{\vec{X}=E(\vec{X})}\left(X_j-E(X_j)\right)$$

4

Prawo przenoszenia błędów

$$cov(Y_k,Y_q)=E\left[(Y_k-E(Y_k))(Y_q-E(Y_q))\right]=$$

$$=E\left[\sum_l\left.\left(\frac{\partial Y_k}{\partial X_l}\right)\right|_{\vec{X}=E(\vec{X})}(X_l-E(X_l))\sum_m\left.\left(\frac{\partial Y_q}{\partial X_m}\right)\right|_{\vec{X}=E(\vec{X})}(X_m-E(X_m))\right]=$$

$$=\sum_{l,m}\left.\left(\frac{\partial Y_k}{\partial X_l}\right)\right|_{\vec{X}=E(\vec{X})}\left.\left(\frac{\partial Y_q}{\partial X_m}\right)\right|_{\vec{X}=E(\vec{X})}E\left[(X_l-E(X_l))(X_m-E(X_m))\right]=$$

$$=\boxed{\sum_{l,m}\left.\left(\frac{\partial Y_k}{\partial X_l}\right)\right|_{\vec{X}=E(\vec{X})}\left.\left(\frac{\partial Y_q}{\partial X_m}\right)\right|_{\vec{X}=E(\vec{X})}cov(X_l,X_m)}$$

$$[C(X)]_{l,m}\equiv cov(X_l,X_m);\quad [C(Y)]_{k,q}\equiv cov(Y_k,Y_q);\quad T_{i,j}\equiv\left.\left(\frac{\partial Y_i}{\partial X_j}\right)\right|_{\vec{X}=E(\vec{X})}$$

$$[C(Y)]_{k,q}=\sum_{l,m}T_{k,l}T_{q,m}[C(X)]_{l,m}=\sum_{l,m}T_{k,l}[C(X)]_{l,m}T^T_{m,q}$$

$\boxed{C(Y)=TC(X)T^T}$ ← macierzowo

5

Prawo przenoszenia błędów - przykład

$$cov(Y_k,Y_q)=\sum_{l,m}\left.\left(\frac{\partial Y_k}{\partial X_l}\right)\right|_{\vec{X}=E(\vec{X})}\left.\left(\frac{\partial Y_q}{\partial X_m}\right)\right|_{\vec{X}=E(\vec{X})}cov(X_l,X_m)$$

- Zał: Y_k (nowa zmienna) zależy od $X_1,\ldots X_N$ (stare zmienne). Dokonujemy pomiaru wszystkich X_i otrzymując $E(X_i)$ oraz $\sigma(X_i)$

$$\sigma(Y_k)=\sqrt{\mathrm{var}(Y_k)}=\sqrt{\mathrm{cov}(Y_k,Y_k)}=\sqrt{\sum_{l,m}\left.\left(\frac{\partial Y_k}{\partial X_l}\right)\right|_{\vec{X}=E(\vec{X})}\left.\left(\frac{\partial Y_k}{\partial X_m}\right)\right|_{\vec{X}=E(\vec{X})}cov(X_l,X_m)}=\ldots$$

$$X_l\ i\ X_m\ niezależne\rightarrow cov(X_l,X_m)=0$$

$$\ldots=\sqrt{\sum_l\left(\left.\left(\frac{\partial Y_k}{\partial X_l}\right)\right|_{\vec{X}=E(\vec{X})}\right)^2cov(X_l,X_l)}=\sqrt{\sum_l\left(\left.\left(\frac{\partial Y_k}{\partial X_l}\right)\right|_{\vec{X}=E(\vec{X})}\right)^2(\sigma(X_l))^2}$$

6

Transformacje wektorów losowych

- Syt. $\vec{Z} = g(\vec{X})$
- Przypadek dyskretny: grupujemy te wartości $\vec{X}$, które dają te same wartości $\vec{Z}$ i dodajemy prawdopodobieństwa. Otrzymujemy rozkład prawdopodobieństwa wektora losowego $\vec{Z}$.
- Przypadek ciągły:

I sposób: (zał. Z jest jednowymiarowe czyli $Z = g(X_1, X_2, \ldots, X_n)$)

$$F_Z(z) = \underbrace{\int \ldots \int}_{\substack{\text{po wszystkich zdarzeniach} \\ \text{dla których } Z(A) \le z}} f_{\vec{X}}(x_1, x_2, \ldots, x_n)\, dx_1 \ldots dx_n$$

Znając dystrybuantę obliczamy (łączną) funkcję gęstości prawdopodobieństwa.

Dla Z n-wymiarowej całkujemy po wszystkich zdarzeniach dla których $Z_1(A) \le z_1 \wedge Z_2(A) \le z_2 \wedge \ldots \wedge Z_n(A) \le z_n$.

Transformacje wektorów losowych

- Przypadek ciągły:

II sposób:

zał.:

1. n=2 czyli (X,Y)→(V,W) czyli V=g_1(X,Y), W=g_2(X,Y)
2. istnieją jednoznaczne funkcje h_1: X=h_1(V,W) i h_2:Y=h_2(V,W)
3. dla każdego x,y funkcje g_1 i g_2 mają ciągłe pochodne cząstkowe
4. Jacobian

$$J_{(X,Y)} \equiv \begin{vmatrix} \frac{\partial g_1}{\partial x} & \frac{\partial g_1}{\partial y} \\ \frac{\partial g_2}{\partial x} & \frac{\partial g_2}{\partial y} \end{vmatrix} \ne 0$$

Wtedy

$$\boxed{f_{V,W}(v,w) = f_{X,Y}(x(v,w), y(v,w)) \left| J_{(X,Y)} \right|^{-1}}$$

Transformacje wektorów losowych - przykład

- Niezależne zmienne losowe X,Y mają rozkłady jednorodne na (0,1). Znaleźć łączną funkcję gęstości prawdopodobieństwa zmiennych $V \equiv \sqrt{-2\ln(X)}\cos(2\pi Y)$ oraz $W \equiv \sqrt{-2\ln(X)}\sin(2\pi Y)$
- Rozwiązanie:

Z treści zadania: $f_{X,Y}(x,y) = f_X(x) \cdot f_Y(y) = 1 \cdot 1 = 1$

Potrzebujemy: $f_{V,W}(v,w) = f_{X,Y}(x(v,w), y(v,w)) \left| J_{(X,Y)} \right|^{-1}$

Liczymy

$$\tfrac{\partial V}{\partial x} = \tfrac{\partial}{\partial x}\left(\sqrt{-2\ln(X)}\cos(2\pi Y)\right) = \cos(2\pi Y)\tfrac{\partial}{\partial x}\left(\sqrt{-2\ln(X)}\right) = \tfrac{\cos(2\pi Y)}{2\sqrt{-2\ln(X)}} \cdot \tfrac{(-2)}{X}$$

$$\tfrac{\partial V}{\partial y} = \tfrac{\partial}{\partial y}\left(\sqrt{-2\ln(X)}\cos(2\pi Y)\right) = \sqrt{-2\ln(X)}\tfrac{\partial}{\partial y}\left(\cos(2\pi Y)\right) = \sqrt{-2\ln(X)} \cdot (-1)\sin(2\pi Y) \cdot 2\pi$$

$$\tfrac{\partial W}{\partial x} = \tfrac{\partial}{\partial x}\left(\sqrt{-2\ln(X)}\sin(2\pi Y)\right) = \sin(2\pi Y)\tfrac{\partial}{\partial x}\left(\sqrt{-2\ln(X)}\right) = \tfrac{\sin(2\pi Y)}{2\sqrt{-2\ln(X)}} \cdot \tfrac{(-2)}{X}$$

$$\tfrac{\partial W}{\partial y} = \tfrac{\partial}{\partial y}\left(\sqrt{-2\ln(X)}\sin(2\pi Y)\right) = \sqrt{-2\ln(X)}\tfrac{\partial}{\partial y}\left(\sin(2\pi Y)\right) = \sqrt{-2\ln(X)} \cdot \cos(2\pi Y) \cdot 2\pi$$

Transformacje wektorów losowych - przykład

$V \equiv \sqrt{-2\ln(X)}\cos(2\pi Y)$

$W \equiv \sqrt{-2\ln(X)}\sin(2\pi Y)$

$$J_{(X,Y)} = \det\begin{pmatrix} \frac{\partial V}{\partial x} & \frac{\partial V}{\partial y} \\ \frac{\partial W}{\partial x} & \frac{\partial W}{\partial y} \end{pmatrix} = \frac{\partial V}{\partial x} \cdot \frac{\partial W}{\partial y} - \frac{\partial V}{\partial y} \cdot \frac{\partial W}{\partial x} = \left(\cos(2\pi Y)\right)^2 \cdot \tfrac{(-2\pi)}{X} - \left(\sin(2\pi Y)\right)^2 \cdot \tfrac{(2\pi)}{X} =$$

$$= \tfrac{(-2\pi)}{X} \cdot \left(\left(\cos(2\pi Y)\right)^2 + \left(\sin(2\pi Y)\right)^2\right) = \tfrac{(-2\pi)}{X}$$

$$\left| J_{(X,Y)} \right|^{-1} = \left| \tfrac{(-2\pi)}{X} \right|^{-1} = \left(\tfrac{2\pi}{X}\right)^{-1} = \tfrac{X}{2\pi}$$

$$V^2 + W^2 = \left(\sqrt{-2\ln(X)}\cos(2\pi Y)\right)^2 + \left(\sqrt{-2\ln(X)}\sin(2\pi Y)\right)^2 = \left(\sqrt{-2\ln(X)}\right)^2 = -2\ln(X)$$

$$\rightarrow \quad X = e^{\frac{-1}{2}\left(V^2 + W^2\right)} \quad \rightarrow \quad \left| J_{(X,Y)} \right|^{-1} = \tfrac{1}{2\pi} e^{\frac{-1}{2}\left(V^2 + W^2\right)}$$

$$\rightarrow \quad f_{V,W}(v,w) = f_{X,Y}(x(v,w), y(v,w)) \left| J_{(X,Y)} \right|^{-1} = 1 \cdot \tfrac{1}{2\pi} e^{\frac{-1}{2}\left(V^2 + W^2\right)} = \tfrac{1}{2\pi} e^{\frac{-1}{2}\left(V^2 + W^2\right)}$$

$$\rightarrow \quad f_{V,W}(v,w) = \tfrac{1}{\sqrt{2\pi}} e^{\frac{-1}{2}V^2} \cdot \tfrac{1}{\sqrt{2\pi}} e^{\frac{-1}{2}W^2} = N_V(0,1) \cdot N_W(0,1) = f_V(v) \cdot f_W(w)$$

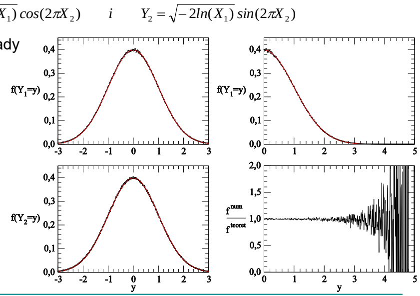
Dwa niezależne rozkłady N(0,1)

Transformacja Box-Mullera

- X_1, X_2 mają rozkłady jednorodne na przedziale (0,1); wtedy

$Y_1 = \sqrt{-2ln(X_1)}\, cos(2\pi X_2)$ i $Y_2 = \sqrt{-2ln(X_1)}\, sin(2\pi X_2)$

mają rozkłady N(0,1).



Wielowymiarowy rozkład normalny

- Łączna gęstość prawdopodobieństwa

$$\boxed{f_{\vec{X}}(\vec{x}) = \left(\frac{\det(K^{-1})}{(2\pi)^n}\right)^{1/2} e^{\frac{-1}{2}(\vec{x}-\vec{a})^T K^{-1} (\vec{x}-\vec{a})}}$$

$\vec{X}$ – wektor n zmiennych losowych

K – macierz kowariancji n×n

$\vec{a}$ – stały wektor n liczb (E($\vec{X}$)=$\vec{a}$)

- dla n=2

$$K^{-1} = \frac{1}{var(X_1)var(X_2) - (cov(X_1,X_2))^2}\begin{pmatrix} var(X_2) & -cov(X_1,X_2) \\ -cov(X_1,X_2) & var(X_1) \end{pmatrix}$$

- Standaryzacja $U_i = \dfrac{X_i - a_i}{\sigma(X_i)}$ dla n=2 prowadzi do

$$f_{\vec{U}}(u_1,u_2) = \left(\frac{\det(K^{-1})}{(2\pi)^2}\right)^{1/2} e^{\frac{-1}{2}\vec{u}^T K^{-1} \vec{u}}$$

ρ to współczynnik korelacji u_1 i u_2

$$K^{-1} = \frac{1}{1-\rho^2}\begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix}$$

Wielowymiarowy rozkład normalny - elipsa kowariancji

- Łączna gęstość prawdopodobieństwa

$$f_{\vec{U}}(u_1, u_2) = const \to \vec{u}^T K^{-1} \vec{u} = const \to$$

$$\frac{1}{1-\rho^2}(u_1^2 + u_2^2 - 2u_1u_2\rho) = const \to$$

$$\frac{1}{1-\rho^2}\left(\frac{(x_1-a_1)^2}{(\sigma(X_1))^2} + \frac{(x_2-a_2)^2}{(\sigma(X_2))^2} - 2\rho\frac{(x_1-a_1)}{\sigma(X_1)}\frac{(x_2-a_2)}{\sigma(X_2)}\right) = const$$

- Jest to równanie elipsy zwanej dla const=1 elipsą kowariancji. Prawdopodobieństwo wystąpienia zmiennych losowych wewnątrz elipsy kowariancji jest niezależne od $\sigma(X_1)$,$\sigma(X_2)$ i ρ.