Vektor náhodných veličin - práce s více proměnnými

12. května 2015

Někdy k popisu nějaké situace potřebujeme více než jednu náhodnou veličinu. Např. věk, hmotnost, výšku. Mezi těmito veličinami mohou být i nějaké statistické vztahy. V tomto materiálu se stručně seznámíme s hlavními prostředky, které pro práci s náhodným vektorem (více náhodnými veličinami) budeme potřebovat.

Pracovat zde budeme se spojitými veličinami. Pokud bychom chtěli pracovat s diskrétními, nahradíme integrály sumami apod.

Sdružená, marginální a podmíněná hustota pravděpodobnosti

Sdružená hustota $f(x, y)$

Jedná se o funkci více proměnných, která:

1) Je na celém definičním intervalu nezáporná.

2) Integrál (součet) přes celý definiční obor z této funkce je 1.

Pravděpodobnost, že náhodná veličina x je mezi 3 a 5 a současně je náhodná veličina y mezi 4 a 6, spočteme podle vzorce:

$$P\left(x \in (3,5) \wedge y \in (4,6)\right) = \int_{x=3}^{5} \int_{y=4}^{6} f\left(x, y\right) dydx.$$

Marginální hustota $f(x)$

Jedná se o hustotu pravděpodobnosti jedné proměnné odvozenou ze sdružené, s tím, že mne zajímá pravděpodobnostní rozdělení pouze této jedné proměnné. Spočte se integrováním přes všechny ostatní proměnné a přes celý definiční obor:

$$f(x) = \int_{\Omega} f(x, y) \cdot dy$$

Pravděpodobnost, že náhodná veličina x je mezi 3 a 5, pokud o veličině y nevíme nic, spočteme podle vzorce:

$$P\left(x \in (3,5)\right) = \int_{x=3}^{5} f\left(x\right) dx.$$

Podmíněná hustota $f(x|y)$

Jedná se o hustotu pravděpodobnosti jedné proměnné (x) odvozenou za předpokladu, že znám hodnoty ostatních proměnných (y). Spočte se:

$$f(x|y) = \frac{f(x,y)}{f(y)}$$

Pokud $f(x|y) = f(x)$, jsou veličiny x a y nezávislé. To znamená, že znalost o jedné z nich mi nepřináší jakoukoliv informaci o druhé z nich.

Pravděpodobnost, že náhodná veličina x je mezi 3 a 5, pokud veličina y je 10, spočteme podle vzorce:

$$P\left(x \in (3,5)\,|y = 10\right) = \int_{x=3}^{5} f(x|10)dx.$$

Střední hodnota

Pokud chceme počítat střední hodnotu čehokoliv - $\check{C}$, spočteme ji jako určitý integrál přes celý definiční obor z $\check{C}$ krát hustota.

Tedy střední hodnotu pro x spočteme:

$$\mu_x = \int_{\Omega} x \cdot f\left(x,y\right) dxdy.$$

Střední hodnotu pro y spočteme:

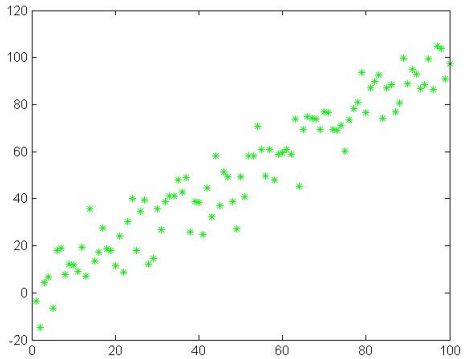
$$\mu_y = \int_{\Omega} y \cdot f\left(x,y\right) dxdy.$$

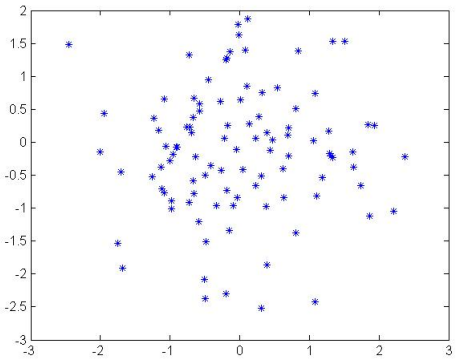
Někdy se chce střední hodnota pro vektorovou veličinu $X = (x,y)$. Tou je vektor: $\mu_X = (\mu_x, \mu_y)$.

Podobně pracujeme s rozptylem nebo jinými charakteristikami.

Kovariance a korelace

Chci umět říci, jak moc jsou dvě veličiny na sobě lineárně závislé. První obrázek ukazuje velkou závislost, druhý nulovou.





Na prvním obrázku vidíme, že pokud je x větší než průměr, i y bývá větší než průměr. Důležité jsou tedy odchylky od průměru. Základem tedy nebudou vektory hodnot, ale vektory hodnot, od nichž je odečtena střední hodnota. Např. $X = x - \mu_x$.

Z těchto vektorů udělám skalární součin $X \cdot Y = \sum (x - \mu_x) \cdot (y - \mu_y)$.

Protože skalární součin bude tím větší, čím víc položek budou mít vektory, podělím skalární součin počtem položek a dostanu kovarianci.

Pokud pracuji s výběrovým souborem, nedělím n, ale $n - 1$. Je to úplně stejné jako u rozptylů. Pozor na to!

Mimochodem - kovariance vektoru se sebou samým dává právě rozptyl.

Základní soubor:

$$cov(x, y) = \frac{\sum (x - \mu_x)(y - \mu_y)}{n}$$

Výběrový soubor:

$$cov(x, y) = \frac{\sum (x - \bar{x})(y - \bar{y})}{n - 1}$$

Užitečnější než kovariance je korelační koeficient.

Možná si vzpomenete na vzoreček ze střední školy pro skalární součin: $\cos \alpha = \frac{\vec{u} \cdot \vec{v}}{|u| \cdot |v|}$. Když je kosinus blízký jedničce, ukazují oba vektory téměř stejným směrem. Když je blízký -1, ukazují téměř opačně. Když je kolem nuly, ukazují zcela jinam.

A to je vlastně korelační koeficient. Když si za vektory vezmu ty s odečtenou střední hodnotou ($X = x - \mu_x$, $Y = y - \mu_y$), můžeme korelační koeficient definovat jako:

$$r_{xy} = \frac{X \cdot Y}{|X| \cdot |Y|} = \frac{\sum (x - \mu_x)(y - \mu_y)}{\sqrt{\sum (x - \mu_x)^2 \cdot \sum (y - \mu_y)^2}}$$

Jinak můžeme použít i definici:

$$r_{xy} = \frac{cov(x, y)}{\sigma_x \cdot \sigma_y}$$

Tento vzorec platí pro základní i výběrový soubor, protože faktor $n - 1$ resp. n se vykrátí.

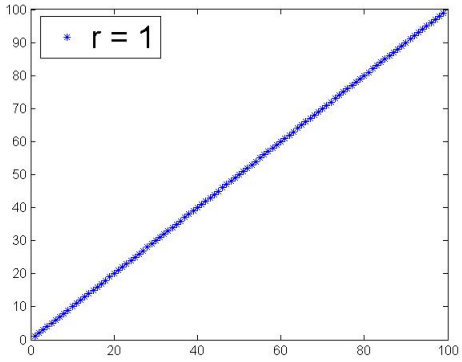
Hodnota korelačního koeficientu je vždy mezi -1 a 1. Pokud je blízko krajním hodnotám, veličiny na sobě silně lineárně závisí. Pokud je blízko nuly, lineární závislost je slabá nebo žádná.

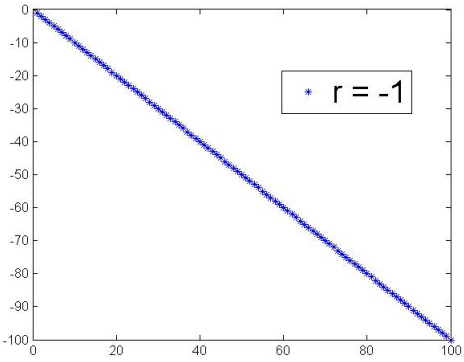
Kovarianční a korelační matice

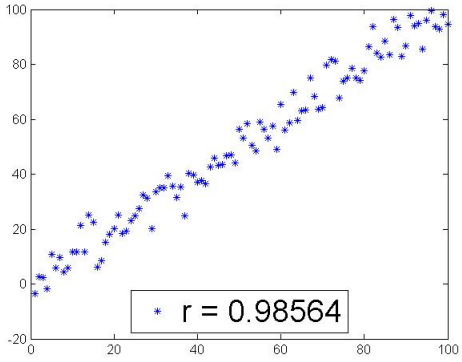
Když mám vektorů více, mohu sestrojit kovarianční nebo korelační matici, kam napíšu kovarianci resp. korelaci každého vektoru s každým.

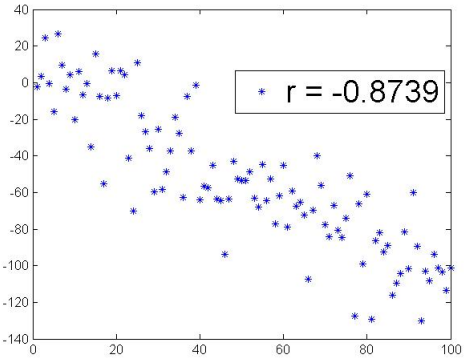
Na diagonále kovarianční matice tak dostanu rozptyly. Na diagonále korelační matice jedničky.

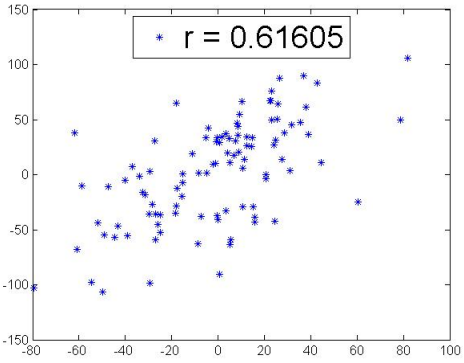
Následuje několik ilustračních grafů s hodnotami korelačního koeficientu.

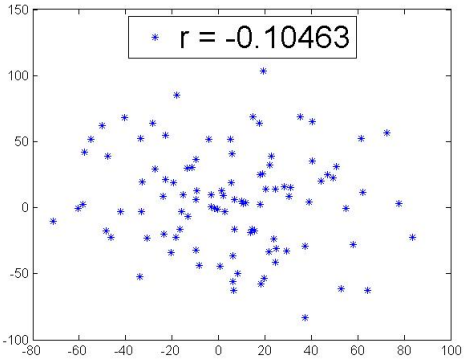

r = 1

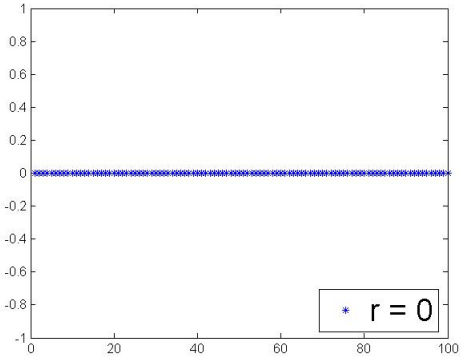

r = -1

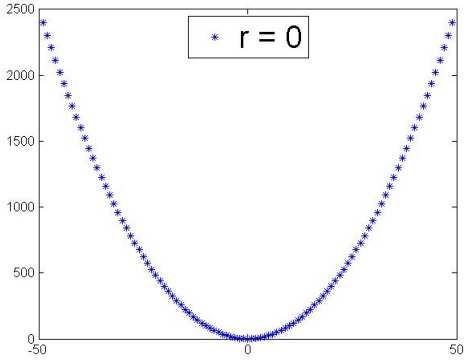

r = 0.98564


r = -0.8739


r = 0.61605


r = -0.10463


r = 0


r = 0

V předposledním obrázku získáváme nulový korelační koeficient, protože y není závislé na x. Pro jakékoliv x je stále nula.

V posledním obrázku vidíme, že korelační koeficient měří jen lineární závislost. Složitější závislost není schopen zachytit.