Závislosť náhodných udalostí a premenných

Formálna definícia závislosti udalostí:

Náhodné udalosti $A_1,...,A_n$ sú združene nezávislé, ak pre akýkoľvek výber z nich $A_{i1},...,A_{im}$ platí

$$P[A_{i1} \& ... \& A_{im}] = P[A_{i1}] ... P[A_{im}].$$

Intuitívne:

$A_1,...,A_n$ sú združene nezávislé, ak informácia, že nastala udalosť vytvorená z $A_{k1},...,A_{km}$, nezmení naše očakávanie, že nastala udalosť vytvorená z iných udalostí $A_{j1},...,A_{jl}$ („z iných udalostí“ = {k1,...,km} a {j1,...,jl} sú disjunktné).

Formálna definícia závislosti náhodných premenných:

Náhodné premenné $X_1,...,X_n$ sú združene nezávislé, ak pre akékoľvek $B_1,...,B_n$ sú [X_1 patrí do B_1], ..., [X_n patrí do B_n] združene nezávislé udalosti.

Intuitívne:

$X_1,...,X_n$ sú združene nezávislé, ak informácia o tom, aké hodnoty nadobudli $X_{k1},...,X_{km}$, nezmení naše očakávanie ohľadom hodnôt, ktoré nadobudli iné premenné $X_{j1},...,X_{jl}$ („iné premenné“ = {k1,...,km} a {j1,...,jl} sú disjunktné).

Nezávislosť sa typicky vyskytuje v:

Merania (pozorovania/pokusy/výbery), o ktorých výsledku nám neposkytujú žiadnu informáciu výsledky iných meraní (pozorovaní/pokusov/výberov).

Závislosť sa typicky vyskytuje v:

Merania (pozorovania/pokusy/výbery), o ktorých výsledku nám poskytujú informáciu výsledky iných (napríklad časovo predchádzajúcich) meraní (pozorovaní/pokusov/výberov)

Štatistické testovanie nezávislosti:

To, či sú dáta realizáciou nezávislých premenných vieme (nedokonale!) testovať štatistickými testami, napríklad:

- Testy nulovosti rôznych mier „konkordancie“ dvojice náhodných premenných (napr. Pearsonovho korelačného koeficientu a iných),

- Testy rovnosti združeného rozdelenia a súčinu marginálnych rozdelení (napr. chíkvadrát test nezávislosti a mnohé iné).

Kovariancia náhodných premenných

Formálna definícia kovariancie a vlastnosti:

Pre náhodné premenné X,Y definujeme

$$\text{cov}(X,Y)=E[(X-EX)(Y-EY)].$$

cov(X,Y)=cov(Y,X), cov(X+a,Y+b)=cov(X,Y), cov(aX, bY)= a.b.cov(X,Y) pre všetky a,b. Ak sú X,Y nezávislé, cov(X,Y)=0.

Intuitívne:

Kovariancia je symetrická miera „lineárnej“ závislosti premenných X a Y, na ktorú nemá vplyv posun hodnôt premenných X a Y, avšak má vplyv zmena jednotiek X a Y.

Interpretácia:

- cov(X,Y)>0: s vyšším X sa dajú čakať vyššie Y,
- cov(X,Y)<0: s vyšším X sa dajú čakať nižšie Y,
- cov(X,Y)=0: zvyšovanie X nevplýva na stredný rast/pokles Y,
- Zo samotnej hodnoty cov(X,Y) je ťažké niečo vyčítať...

Výberový odhad kovariancie z n nezávislých dvojíc realizácií:

$$\widehat{cov}(X,Y)=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$$

V tomto vzťahu sú realizácie $(x_1,y_1),\ldots,(x_n,y_n)$ a $\bar{x},\bar{y}$ znamenajú aritmetický priemer.

(Pearsonova) korelácia náhodných premenných

Formálna definícia:

Pre náhodné premenné X,Y definujeme

$$cor(X,Y) = \frac{cov(X,Y)}{\sqrt{DX}\sqrt{DY}}$$

cor(X,Y)=cor(Y,X), cor(X+a,Y+b)=cor(X,Y) pre všetky a,b, cor(aX,bY)=cor(X,Y) pre všetky a>0, b>0. Ak sú X,Y nezávislé, cor(X,Y)=0. Tiež

$$-1 \leq cor(X,Y) \leq 1$$

Intuitívne:

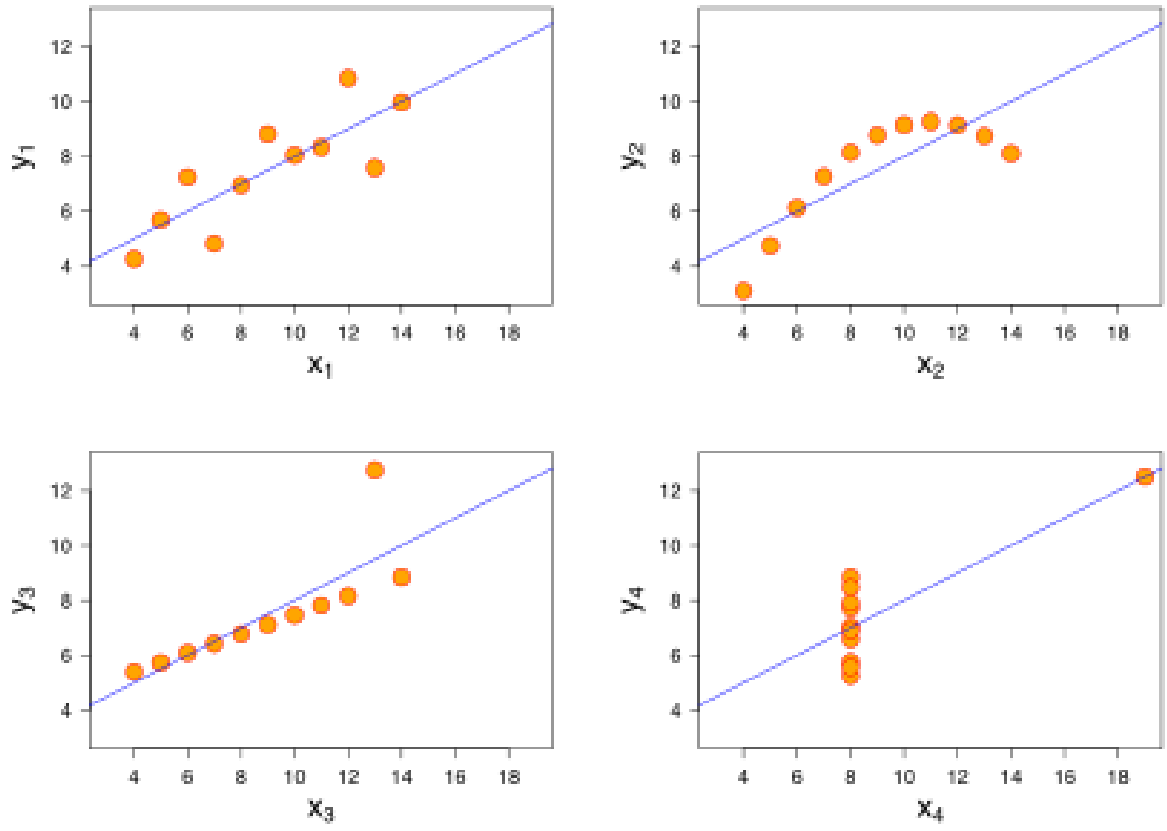
Korelácia je symetrická, normalizovaná miera „lineárnej“ závislosti premenných X a Y, na ktorú nemá vplyv posun hodnôt premenných X a Y, ani zmena jednotiek X a Y.

Interpretácia:

- cor(X,Y)>0: s vyšším X sa dajú čakať vyššie Y,
- cor(X,Y)<0: s vyšším X sa dajú čakať nižšie Y,
- cor(X,Y)=0: zvyšovanie X nevplýva na stredný rast/pokles Y,
- Ak je cor(X,Y) blízke -1, máme skoro presnú klesajúcu lineárnu závislosť premenných X a Y,
- Ak je cor(X,Y) blízke 1, máme skoro presnú rastúcu lineárnu závislosť premenných X a Y.

"Nedokonalosti" cor(X,Y) z praktického hľadiska:

- Rôzne "formy" závislosti môžu viesť na ten istý korelačný koeficient, viď napr. *Anscombe's quartet* (obrázok je z wiki)



- Pozri tiež *Datasaurus dozen* (napr. https://cran.r-project.org/web/packages/datasauRus/vignettes/Datasaurus.html)

- Závislé náhodné premenné môžu mať nulový korelačný koeficient, ako je vidieť z Datasaurus dozen

Výberový odhad korelácie z n nezávislých dvojíc realizácií:

$$\widehat{cor}(X,Y) = \frac{\widehat{cov}(X,Y)}{S_X S_Y},$$

kde S_X je odmocnina výberového rozptylu realizácií $x_1, \dots, x_n$ náhodnej premennej X a analogicky S_Y pre náhodnú premennú Y.

Súvis výberovej korelácie s regresnou priamkou:

Ak $\hat{a}x + \hat{b}$ je regresná priamka vypočítaná z dvojíc $(x_1, y_1), \dots, (x_n, y_n)$ a $\widehat{cor}(X,Y)$ je výberový korelačný koeficient vypočítaný z tých istých dvojíc, tak

$$\hat{a} = \widehat{cor}(X,Y)\frac{S_Y}{S_X}.$$

Test nulovosti korelačného koeficientu:

$$t = \widehat{cor}(X,Y)\sqrt{\frac{n-2}{1-\widehat{cor}(X,Y)^2}}$$

má tazvané Studentovo rozdelenie s n-2 stupňami voľnosti, ak sú $(x_1, y_1), \dots, (x_n, y_n)$ realizácie dvojrozmerného normálneho rozdelenia a X,Y majú nulový korelačný koeficient.

Iné typy mier závislosti:

Rôzne druhy korelačných koeficientov (Spearmanovo rho, Kendallovo tau), alebo úplne odlišné miery závislosti

Kauzalita

Definícia (Collins dictionary): *Causality is the relationship of cause and effect. The cause of an event is the thing that makes it happen. The effect of one thing on another is the change that the first thing causes in the second thing.* :)

Otázky týkajúce sa kauzality, ktoré tu nemôžeme riešiť:

- Čo presne je kauzalita?

Jedna zo základných otázok filozofie spadajúca do oblasti ontológie a epistemológie. Táto otázka je zásadná (a nevyjasnená) pre mnohé vedecké disciplíny.

- Ako matematicky formalizovať uvažovanie o kauzalite?

Množstvo prístupov: pravdepodobnostné teórie kauzality, kauzálny kalkulus, teória kontrafaktuálov a iné. Ide o zložité teórie, ktoré si vyžadujú dlhodobé štúdium.

- Ako zistiť, či nejaká premenná kauzálne ovplyvňuje inú premennú?

Viacero prístupov, jeden z nich sa opiera o kontrolované experimenty, resp. o "intervenciu" (o "vynútenie" hodnôt alebo rozdelenia nejakej premennej).

Korelácia nie je kauzalita:

Spresnenie: Nenulová korelácia premenných X a Y neimplikuje kauzálnu súvislosť* medzi X a Y (čiže ani to, že X je príčinou Y, ani to, že Y je príčinou X).

*Nech by sme kauzalitu (pojmy príčiny a následku) formalizovali akýmkoľvek „rozumným“ spôsobom.

To, že premenné X a Y sú korelované je často dôsledkom nie kauzálnej závislosti X od Y alebo Y od X, ale tým, že existuje latentná premenná Z, ktorá kauzálne ovplyvňuje X aj Y.

Príklad: Nech X je hmotnosť žiaka základnej školy a Y je veľkosť jeho slovnej zásoby. X a Y sú korelované, avšak je zjavné*, že nemôžeme tvrdiť, že X je príčinou Y alebo naopak. Existuje však iná premenná, V, napríklad vek (alebo vyspelosť) žiaka, ktorá je kladne korelovaná s X aj s Y, a o ktorej môžeme tvrdiť, že kauzálne ovplyvňuje aj X aj Y.

Úloha: Štúdia v časopise Nature (https://www.nature.com/articles/20094) zistila významnú kladnú koreláciu medzi krátkozrakosťou detí a tým, koľko boli vystavené nočnému svetlu ako novorodenci. Dá sa z toho usúdiť, že nočné svetlo spôsobuje krátkozrakosť?