Statistiques pour les sciences (MAT-4681)

Arthur Charpentier

14 - Corrélation

été 2022

Covariance et corrélation (mathématiques) I

Covariance

On appelle covariance d'un couple de variable aléatoire

$$\text{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$

(à condition que $\mathbb{E}[X^2] < \infty$ et $\mathbb{E}[Y^2] < \infty$).

Covariance

Soient X, X_1, X_2 et Y des variables aléatoires,

- $\text{Cov}(X, Y) = \text{Cov}(Y, X)$,
- $\text{Cov}(X, X) = \text{Var}(X)$,
- $\text{Cov}(aX + b, cY + d) = ac\text{Cov}(X, Y)$,
- $\text{Cov}(X_1 + X_2, Y) = \text{Cov}(X_1, Y) + \text{Cov}(X_2, Y)$,

Covariance et corrélation (mathématiques) II

Covariance

Soient X et Y des variables aléatoires,

$$X \perp\!\!\!\perp Y \Longrightarrow \text{Cov}(X, Y) = 0$$

Corrélation

On appelle corrélation d'un couple de variable aléatoire

$$\text{Cor}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X) \cdot \text{Var}(Y)}}$$

(à condition que $\mathbb{E}[X^2] < \infty$ et $\mathbb{E}[Y^2] < \infty$).

Covariance et corrélation (mathématiques) III

Corrélation

Soient X, X_1, X_2 et Y des variables aléatoires,

- $\text{Cor}(X, Y) = \text{Cor}(Y, X)$,
- $\text{Cor}(X, X) = 1$,
- $\text{Cor}(aX + b, cY + d) = \text{signe}(ac) \cdot \text{Cor}(X, Y)$,
- $\text{Cor}(X, Y) = \in [-1, +1]$,

Corrélation ± 1

Soient X et Y des variables aléatoires,

$$\text{Cor}(X, Y) = \pm 1 \Longleftrightarrow Y = aX + b$$

(admis)

Covariance et corrélation (données) I

Corrélation empirique

On appelle corrélation de $\{(x_1, y_1), \cdots, (x_n, y_n)\}$

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

Cette corrélation est appelée corrélation de Pearson.

Au lieu de calculer la corrélation entre $\boldsymbol{x}$ et $\boldsymbol{y}$, on peut calculer la corrélation entre $\boldsymbol{r}$ et $\boldsymbol{s}$, désignant respectivement les rangs dans les deux échantillons (de x_i dans $\boldsymbol{x}$ et de y_i dans $\boldsymbol{y}$). On parle alors de corrélation de Spearman.

Covariance et corrélation (données) II

Exemple dans Truett & Cicchetti (1976) on a

- relation entre le poids de l'animal et le poids du cerveau
- relation entre le poids du cerveau de l'animal et le nombre d'heures de sommeil

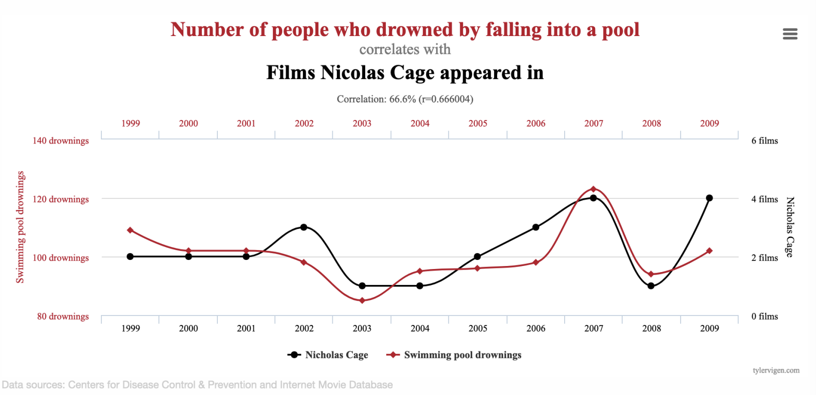
```
> attach(Sleep)
> logBrain = log(BrainWeight)
> logBody  = log(BodyWeight)
> cor(logBrain, logBody, method="pearson")
[1] 0.9520962
```

Note $\text{Cor}(\log X, \log Y) \neq \text{Cor}(X, Y)$

```
> attach(Sleep)
> logBrain = log(BrainWeight)
> logBody  = log(BodyWeight)
> cor(logBrain, logBody, method="pearson")
[1] 0.9520962
```

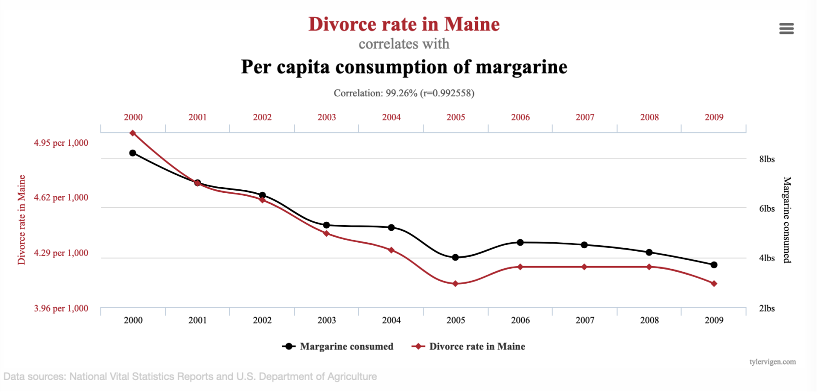
Corrélation et causalité I

via https://www.tylervigen.com/spurious-correlations



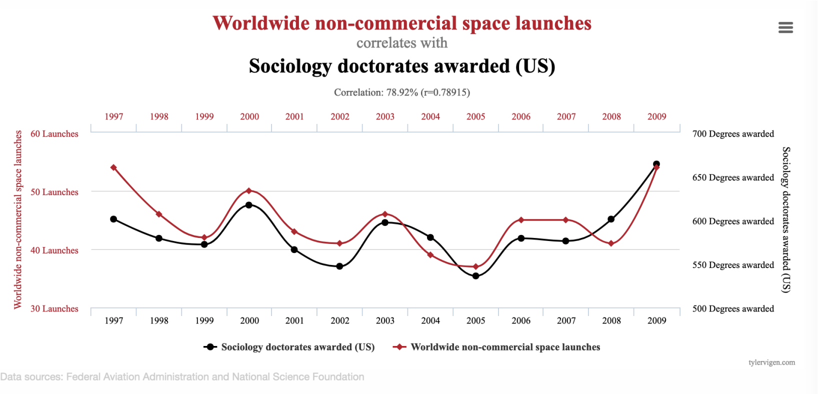
Corrélation et causalité II

via https://www.tylervigen.com/spurious-correlations



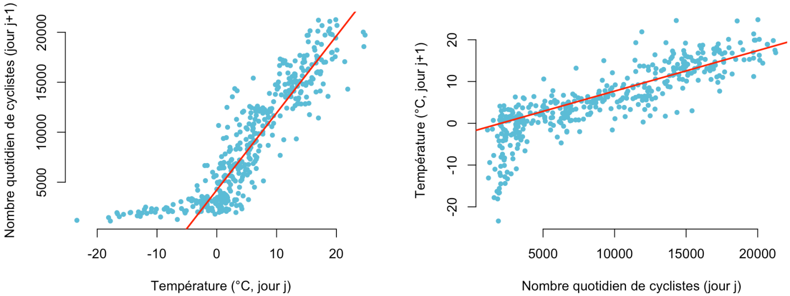
Corrélation et causalité III

via https://www.tylervigen.com/spurious-correlations



Corrélation et causalité IV

```
> df = read.csv("http://freakonometrics.free.fr/
    cyclistsTempHKI.csv")
> n = nrow(df)
> df1 = data.frame(meanTemp = df$meanTemp[1:(n-1)],
    cyclists = df$cyclists[2:n])
> df2 = data.frame(meanTemp = df$meanTemp[2:n],
    cyclists = df$cyclists[1:(n-1)])
```



Corrélation et causalité V

On peut regarder la corrélation entre la température le jour j et le nombre de cyclistes le jour $j+1$

```
> cor.test(df1$meanTemp,df1$cyclists)

  Pearson's product-moment correlation

data:  df1$meanTemp and df1$cyclists
t = 35.758, df = 422, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal
    to 0
95 percent confidence interval:
 0.8413353 0.8889236
sample estimates:
      cor
0.8670943
```

Corrélation et causalité VI

On peut regarder la corrélation entre le nombre de cyclistes le jour j et la température le jour $j+1$

```
> cor.test(df2$meanTemp,df2$cyclists)

  Pearson's product-moment correlation

data:  df2$meanTemp and df2$cyclists
t = 30.07, df = 421, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal
    to 0
95 percent confidence interval:
 0.7931397 0.8540990
sample estimates:
      cor
0.8260198
```

Test de corrélation I

Test H_0 : cor$(x,y) = 0$ contre H_1 : cor$(x,y) \neq 0$, $\mathcal{N}(\mu_\cdot, \sigma_\cdot^2)$

Soit $(\boldsymbol{x},\boldsymbol{y}) = \{(x_1,y_1),\cdots,(x_n,y_n)\}$ où x est de loi $\mathcal{N}(\mu_x,\sigma_x^2)$ et y de loi $\mathcal{N}(\mu_y,\sigma_y^2)$. Pour tester H_0 : cor$(x,y) = 0$ contre H_1 : cor$(x,y) \neq 0$, on utilise la statistique de test

$$T_0 = \sqrt{n-2}\frac{r}{\sqrt{1-r^2}} \text{ où } r = \frac{\sum_{i=1}^n (x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^n (x_i-\bar{x})^2}\sqrt{\sum_{i=1}^n (y_i-\bar{y})^2}}$$

Si H_0 est vraie, $T_0 \sim \mathcal{S}td(n-2)$, et

- on rejette H_0 si $|t_0| > T_{n-2}^{-1}(1-\alpha/2)$

où T_ν est la fonction de répartition de la loi de Student $\mathcal{S}td(\nu)$

Test de corrélation II

Test $H_0 : \text{cor}(x,y) = 0$ contre $H_1 : \text{cor}(x,y) \neq 0$, $\mathcal{N}(\mu_\cdot, \sigma_\cdot^2)$

Soit $(\boldsymbol{x}, \boldsymbol{y}) = \{(x_1, y_1), \cdots, (x_n, y_n)\}$ où x est de loi $\mathcal{N}(\mu_x, \sigma_x^2)$ et y de loi $\mathcal{N}(\mu_y, \sigma_y^2)$. Pour tester $H_0 : \text{cor}(x,y) = 0$ contre $H_1 : \text{cor}(x,y) \neq 0$, on utilise la statistique de test

$$Z_0 = \frac{\sqrt{n-3}}{2} \log\left(\frac{1+r}{1-r}\right)$$

Si H_0 est vraie, $Z_0 \sim \mathcal{S}td(n-2)$, et

- on rejette H_0 si $|z_0| > T^{-1}(1-\alpha/2)$

où T_ν est la fonction de répartition de la loi de Student $\mathcal{S}td(\nu)$

Test de corrélation III

```
> cor.test(logBrain, logBody, method = "pearson")

      Pearson's product-moment correlation

data: logBrainWeight and logBodyWeight
t = 19.193, df = 38, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal
    to
0
95 percent confidence interval:
0.9106836 0.9745630
sample estimates:
cor
0.9520962
```

La transformation $h : r \mapsto \dfrac{1}{2}\log\left(\dfrac{1+r}{1-r}\right)$ est appelée transformation de Fisher. Elle permet de construire un intervalle de confiance pour la corrélation.

Test de corrélation IV

Intervalle de confiance pour la corrélation $\mathcal{N}(\mu_\cdot, \sigma_\cdot^2)$

Soit $(\boldsymbol{x}, \boldsymbol{y}) = \{(x_1, y_1), \cdots, (x_n, y_n)\}$ où x est de loi $\mathcal{N}(\mu_x, \sigma_x^2)$ et y de loi $\mathcal{N}(\mu_y, \sigma_y^2)$. Un intervalle de confiance de niveau α pour $\text{Cor}(X, Y)$ est

$$\left[h^{-1}(z_-); h^{-1}(z_+)\right] \text{ où } h^{-1}(z) = \frac{e^{2z} - 1}{e^{2z} + 1}$$

où

$$z^- = h(r) - \frac{u_{1-\alpha/2}}{n-3} \text{ et } z^+ = h(r) + \frac{u_{1-\alpha/2}}{n-3}$$

Test de corrélation V

On peut faire des tests sur la corrélation, par exemple entre la taille et le poids (d'élèves)

```
> x = Davis$height
> y = Davis$weight
> cor(x,y)
> h = function(r) .5*log((1+r)/(1-r))
> hinv = function(z) (exp(2*z)-1)/(exp(2*z)+1)
> hinv(h(cor(x,y))+qnorm(c(.025,.975))/sqrt(length(x))
  )
[1] 0.7086076 0.8215484
```

que l'on retrouve avec

```
> cor.test(x,y)

  Pearson's product-moment correlation

95 percent confidence interval:
 0.7080838 0.8218898
```

Test de corrélation VI

Parfois, on peut vouloir tester $H_0 : \text{cor}(X, Y) = r_0$ avec $r_0 \neq 0$.

Test $H_0 : \text{cor}(x, y) = r_0$ contre $H_1 : \text{cor}(x, y) \neq r_0$

Soit $(\boldsymbol{x}, \boldsymbol{y}) = \{(x_1, y_1), \cdots, (x_n, y_n)\}$ où x est de loi $\mathcal{N}(\mu_x, \sigma_x^2)$ et y de loi $\mathcal{N}(\mu_y, \sigma_y^2)$. Pour tester $H_0 : \text{cor}(x, y) = r_0$ contre $H_1 : \text{cor}(x, y) \neq r_0$, on utilise la statistique de test

$$T_r = \frac{\sqrt{n-3}}{2}\left(\log\left(\frac{1+r}{1-r}\right) - \log\left(\frac{1+r_0}{1-r_0}\right)\right)$$

Si H_0 est vraie, $T_r \sim \mathcal{S}td(n-2)$, et

- on rejette H_0 si $|t_r| > T^{-1}(1-\alpha/2)$

où T_ν est la fonction de répartition de la loi de Student $\mathcal{S}td(\nu)$.

Test de corrélation VII

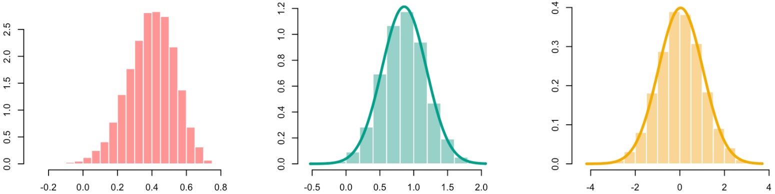
On peut générer des échantillons $\mathcal{N}(0,1)$ avec une corrélation r en considérant

$$X \sim \mathcal{N}(0,1) \text{ et } Y = rX + \sqrt{1-r^2}\cdot Z \text{ avec } Z \sim \mathcal{N}(0,1)$$

```
> r = 0.4
> z =matrix(NA,10000,3)
> for(s in 1:10000){
 x = rnorm(n=40)
 y = r*x+sqrt(1-r^2)*rnorm(n=40)
 cr = cor(x,y)
 t = .5*sqrt(40-3)*(log((1+cr)/(1-cr))-log((1+r)/(1-r)
   ))
 z[s,] = c(cr,log((1+cr)/(1-cr)),t)
}
```

Test de corrélation VIII

On peut visualiser la distribution de r, de $h(r)$ et de z



les deux dernières sont Gaussiennes.

Test de corrélation IX

Test $H_0 : \text{cor}(x, y) = r_0$ contre $H_1 : \text{cor}(x, y) \neq r_0$

À partir de deux échantillons (de tailles respectives m et n), pour tester $H_0 : \text{cor}(X_1, Y_1) = \text{cor}(X_2, Y_2)$ (avec les hypothèses alternatives usuelles), la statistique de test est

$$Z = \frac{1}{2\sqrt{\dfrac{1}{m-3} + \dfrac{1}{n-3}}} \left(\log\left(\frac{1+r_1}{1-r_1}\right) - \log\left(\frac{1+r_2}{1-r_2}\right) \right)$$

Si H_0 est vraie, $Z \approx \mathcal{N}(0, 1)$, et

- on rejette H_0 si $|z| > \Phi^{-1}(1 - \alpha/2)$

Test de corrélation X

On peut faire des tests sur la corrélation, par exemple entre la taille et le poids (d'élèves)

```
> x = Davis$height
> y = Davis$weight
> cor.test(x,y, alternative="two.sided", method="
    pearson")

  Pearson's product-moment correlation

data:  x and y
t = 17.04, df = 198, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal
    to 0
95 percent confidence interval:
 0.7080838 0.8218898
sample estimates:
      cor
0.7710743
```

Test de corrélation XI

On peut se demander si la corrélation entre la taille et le poids est identique, entre les garçons et les filles,

```
> x1 = Davis$height[Davis$sex == "M"]
> y1 = Davis$weight[Davis$sex == "M"]
> x2 = Davis$height[Davis$sex == "F"]
> y2 = Davis$weight[Davis$sex == "F"]
> m = length(x1)
> n = length(x2)
> r1 = cor(x1,y1)
> r2 = cor(x2,y2)
```

La statistique de test est

```
> (z = 0.5*(log((1+r1)/(1-r1))-log((1+r2)/(1-r2)))/(
    sqrt(1/(m-3)+1/(n-3))))
[1] 0.244443
```

et la *p*-value est

```
> 2*min(pnorm(-abs(z)),1-pnorm(-abs(z)))
[1] 0.8068878
```