

Chapitre 6
Introduction à la statistique avec R


Mesure de la force de la liaison entre
deux variables quantitatives :
le coefficient de corrélation

1
Pr. Bruno Falissard
UNIVERSITÉ PARIS SUD

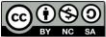
[0:01] Mesurer, quantifier, caractériser la force de l'association existant entre deux mesures est à la base de la plupart des expériences scientifiques. Ainsi le physicien va-t-il chercher à caractériser la relation qu'il observe entre la distance de chute d'un corps et sa vitesse. Le médecin, lui, sera plutôt intéressé par exemple par la relation existant entre la tension artérielle et la survenue d'un infarctus du myocarde.

Liaison et dépendance

Introduction à la statistique avec R > La corrélation




- Deux variables sont dites dépendantes quand la connaissance de l'une donne une indication sur la valeur de l'autre
- La notion de « force » d'une liaison
- Causalité et liaison



2

Pr. Bruno Falissard




[0:27] Commençons par une petite définition : deux variables sont dites dépendantes quand la connaissance de l'une donne une indication sur la valeur de l'autre.

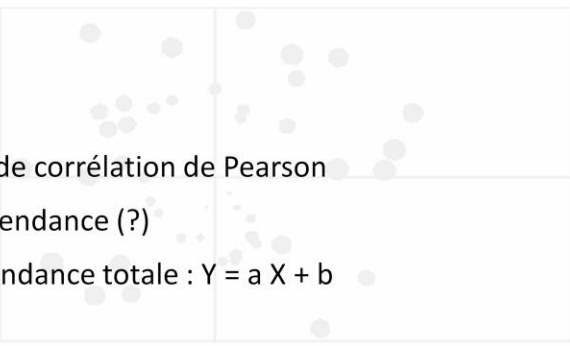
Alors bien sûr, le niveau d'indication peut être plus ou moins élevé et ça correspond à des liaisons plus ou moins fortes. Prenons un exemple : des paires de jumeaux. Si vous connaissez la taille d'un des jumeaux, ça vous donne une indication assez précise sur la taille du second jumeau, on dit que la liaison est forte. Si vous connaissez les revenus du premier jumeau, ça vous donne aussi une indication statistique sur les revenus du second jumeau, mais là la relation est moins forte. Enfin, s'ils vont tous les deux jouer dans un casino, le niveau des pertes ou des gains du premier jumeau ne va être qu'une très maigre indication des pertes ou des gains du second jumeau, en tout cas si les jeux de casino sont des jeux de hasard. Dans ce dernier cas, on dira que la force de la liaison est très faible, voire nulle.

Pour terminer, un quiproquo très fréquent est celui qu'il y a entre liaison et causalité. Ce n'est pas parce qu'il y a une liaison statistique qu'il y a une causalité dans un sens ou dans l'autre entre deux variables. Prenons un exemple : il y a une liaison entre le fait d'avoir les dents jaunes et le fait d'attraper un cancer du poumon, parce que les fumeurs ont souvent les dents jaunes et que fumer est un facteur de risque du cancer du poumon. Bien entendu, il n'y a aucune relation de cause à effet entre avoir les dents jaunes, et avoir un cancer du poumon.

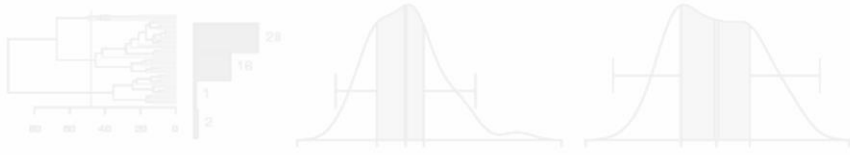
Coefficient de corrélation de Pearson


Introduction à la statistique avec R > La corrélation






- Le coefficient de corrélation de Pearson
- $r = 0 \rightarrow$ indépendance (?)
- $r = \pm 1 \rightarrow$ dépendance totale : $Y = aX + b$





3

Pr. Bruno Falissard




[1:59] Le grand classique pour quantifier la force de l'association entre deux variables aléatoires quantitatives, c'est le **coefficient de corrélation de Pearson**. Alors il faut avoir en tête tout de suite que la corrélation de Pearson, c'est un cas particulier de liaison, liaison qu'on qualifie souvent de monotone ou linéaire, c'est-à-dire si vous avez y et x deux variables, plus l'une est grande, plus l'autre est grande aussi.

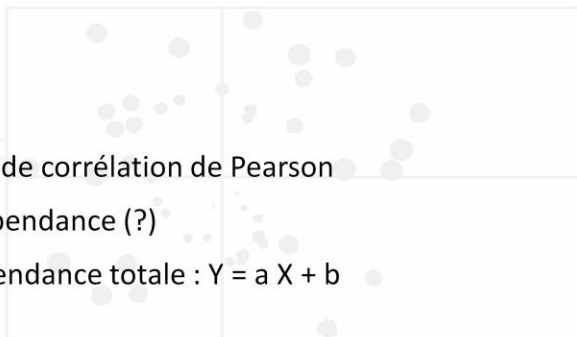
Le coefficient de corrélation est souvent désigné par la lettre r , r varie entre -1 et 1. Quand $r=0$, de façon un peu hâtive, on considère souvent que les deux variables sous-jacentes sont indépendantes ; en fait ce n'est pas complètement vrai, ce n'est vrai que dans le cas où les deux variables suivent une loi normale. Il peut arriver si l'une des deux variables ne suit pas une loi normale d'avoir $r=0$ et pourtant d'avoir une liaison entre les deux variables. Mais il faut bien reconnaître que ça n'est pas si fréquent que ça en pratique.

Quand $r=\pm 1$, la force de la liaison est tellement importante que la connaissance d'une variable donne exactement la valeur de l'autre variable. On dit que les deux variables sont mutuellement déterminées, et elles sont même mutuellement déterminées au moyen d'une relation linéaire de type $y=ax+b$. Quand r est positif, plus l'une des variables est grande, plus l'autre variable est grande ; alors que quand r est négatif, quand une variable augmente, l'autre diminue. Prenons un exemple : entre 0 et 6 ans, il y a une corrélation positive assez forte entre la taille et l'âge ; plus l'enfant vieillit, plus il grandit. A partir de 60 ans, au contraire, il y a une corrélation légère mais négative entre l'âge et la taille ; à cause de l'ostéoporose on a tendance à se tasser et donc plus on vieillit plus on rapetisse.

Coefficient de corrélation de Pearson

Introduction à la statistique avec R > La corrélation






- Le coefficient de corrélation de Pearson
- $r = 0 \rightarrow$ indépendance (?)
- $r = \pm 1 \rightarrow$ dépendance totale : $Y = aX + b$


Traduit la force de la liaison

$$r = \frac{(x_1y_1 + \dots + x_ny_n) - n \times \text{moyenne}(x) \times \text{moyenne}(y)}{(n-1) \times \text{écart_type}(x) \times \text{écart_type}(y)}$$

Standardisation
($-1 \leq r \leq 1$)




3'

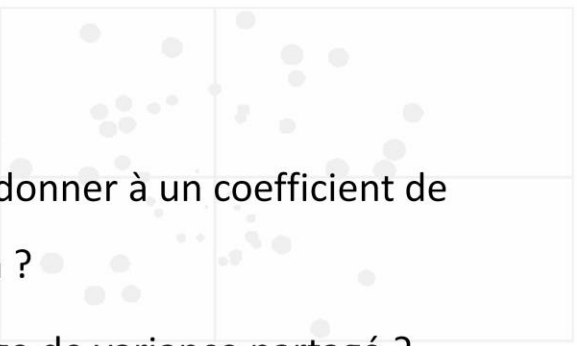
Pr. Bruno Falissard


[3:54] L'expression mathématique qui permet de calculer une corrélation n'est pas si compliquée que ça, elle est présentée ici : ça correspond à la covariance divisée par la racine carrée des variances. Nous avons vu dans un cours précédent que la variance, ce n'était pas facile à interpréter et donc se pose naturellement la question comment interpréter la taille, l'importance, la valeur d'un coefficient de corrélation.


La question du sens


Introduction à la statistique avec R > La corrélation






- Quel sens donner à un coefficient de corrélation ?
- Pourcentage de variance partagé ?
- Illustration :



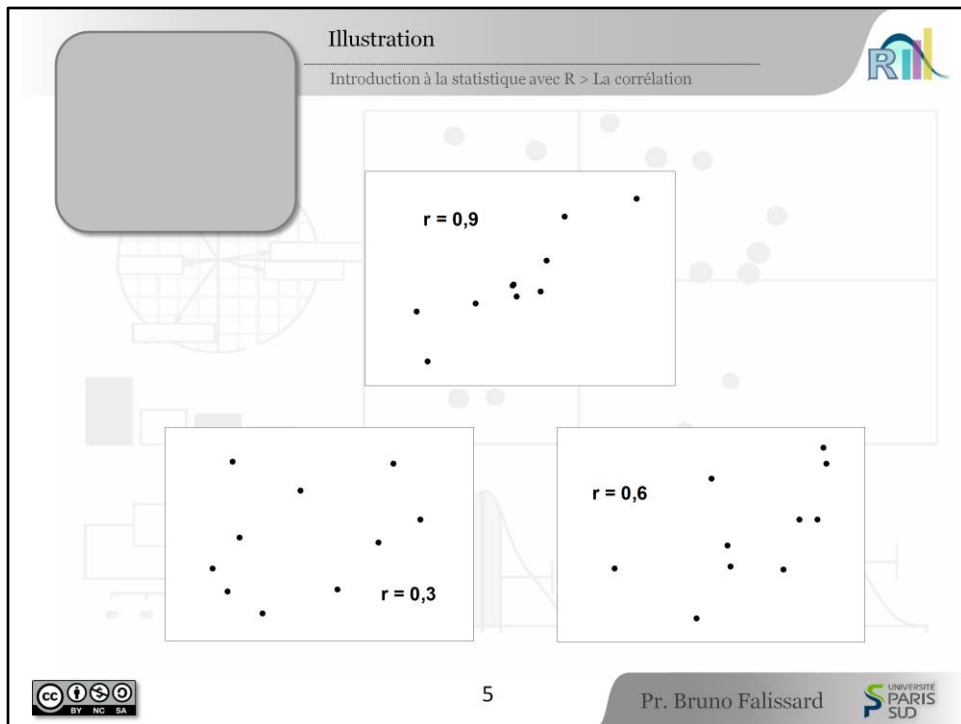

4

Pr. Bruno Falissard


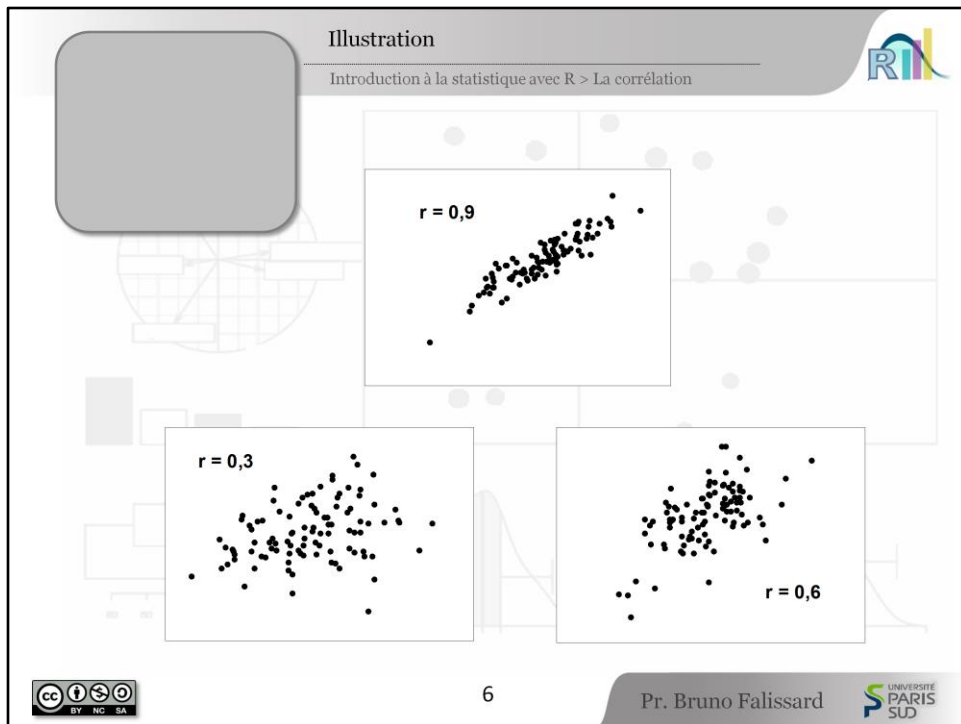
[4:20] Si une corrélation vaut 0,6, ça veut dire quoi ? Et bien malheureusement, ça ne veut pas dire grand-chose. Ce qu'on peut dire et qui marche à tous les coups, c'est qu'une corrélation de 0,6, elle est supérieure à une corrélation de 0,4, et que du coup, la force de la liaison sous-jacente, elle est bien plus grande, ça, ça marche. Par contre le 0,6 en soi, on ne peut pas en faire grand-chose.

Alors vous verrez dans certains livres, on dit que le pourcentage de variance partagé entre deux variables est égal au carré du coefficient de corrélation. Alors bien sûr, mathématiquement c'est vrai, mais ça veut dire quoi ? Si vous avez x et y deux variables, elles corréleront à 0,6. $0,6^2=0,36$. Ça veut dire que le pourcentage de variance partagé entre x et y , c'est 36%, mais ça veut dire quoi ? Alors là les gens interprètent ça en se disant "s'il y a un pourcentage de variance partagé, ça veut dire que les variables se ressemblent, elles se ressemblent à 36%, c'est-à-dire en gros, elles sont pour 36% pareilles". Et bien non, ça c'est faux. Donc au total, je vous incite à être vraiment extrêmement prudents avec un pourcentage de variance partagé.

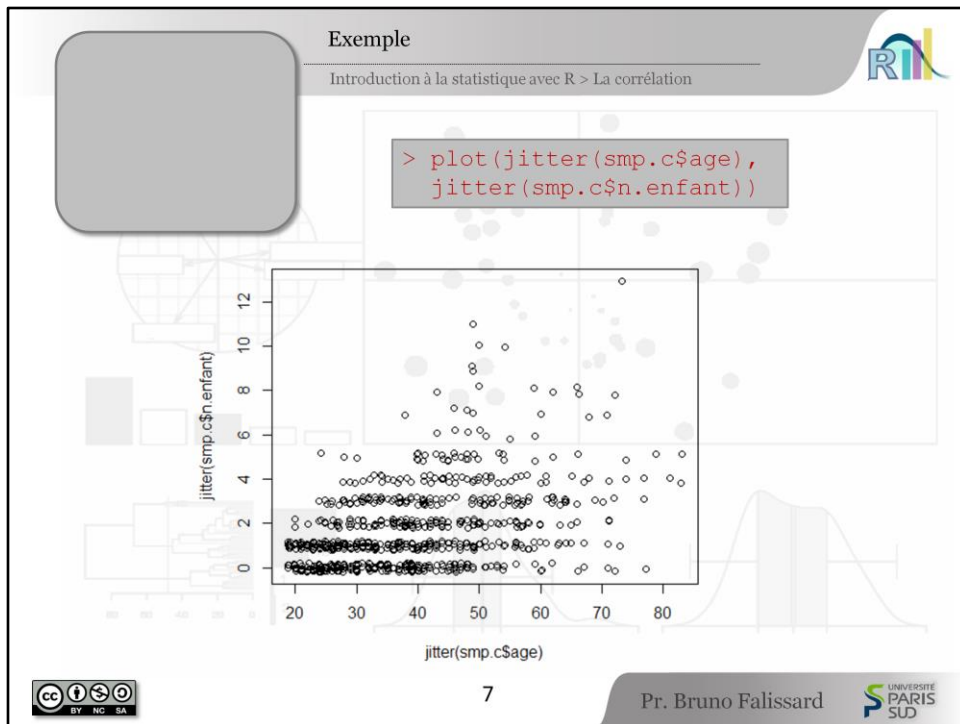
Vous verrez aussi dans certains livres, en particulier dans les sciences de la vie, dans les sciences humaines et sociales, des auteurs qui disent qu'une corrélation de 0,8 c'est très fort, de 0,6 c'est plutôt fort, de 0,4 c'est moyen, de 0,2 c'est faible, de 0,1 c'est extrêmement faible. Tout ça est un peu vrai, mais ça ne repose pas sur des bases très rigoureuses. Une façon de toucher du doigt ce que c'est qu'une corrélation de 0,6 ou de 0,4, c'est de faire des simulations et de regarder des graphiques en x/y .



[6:03] Voici trois simulations où deux variables x et y ont été mesurées sur 10 individus, et ce pour des corrélations de 0,3 0,6 ou 0,9.



[6:16] Ici, nous avons les mêmes simulations, mais cette fois-ci pour 100 individus. On voit que pour une corrélation de 0,9, l'association entre y et x est évidente. Alors que pour 0,3, à l'œil nu, il est bien difficile de dire que plus x est grand, plus y est grand.



[6:37] Alors venons-en maintenant à la pratique : comment calculer avec R un coefficient de corrélation de Pearson ? Je vous propose de reprendre l'exemple de l'âge et du nombre d'enfants chez les détenus de l'étude *santé mentale en prison*. Pour nous raviver la mémoire, et parce que c'est toujours bien de faire des représentations graphiques avant de faire des calculs, faisons un diagramme en x/y de l'âge et du nombre d'enfants. Ensuite, pour calculer la corrélation, il suffit d'utiliser la fonction `cor()` ...

Exemple

Introduction à la statistique avec R > La corrélation



```
> str(smp.c)
'data.frame': 799 obs. of 10 variables:
 $ age      : int  31 49 50 47 23 34 24 52 42 45 ...
 $ prof     : Factor w/ 8 levels "agriculteur",...: 3 NA 7 6 8 6 ...
 $ dep.cons : int  0 0 0 0 1 0 1 0 1 0 ...
 $ scz.cons : int  0 0 0 0 0 0 0 0 0 0 ...
 $ grav.cons: int  1 2 2 1 2 1 5 1 5 5 ...
 $ n.enfant : int  2 7 2 0 1 3 5 2 1 2 ...
 $ rs       : int  2 2 2 2 2 1 3 2 3 2 ...
 $ ed       : int  1 2 3 2 2 2 3 2 3 2 ...
 $ dr       : int  1 1 2 2 2 1 2 2 1 2 ...
 $ ed.b     : num  0 0 1 0 0 0 1 0 1 0 ...


> cor(smp.c$age, smp.c$n.enfant, use="complete.obs")
[1] 0.4326039
```




[7:10] ... avec nos deux variables age et nombre d'enfants et ne pas oublier l'instruction `use="complete.obs"` qui permet de gérer les données manquantes. Si vous n'écrivez pas ça et si vous avez des données manquantes, R marquera qu'il ne peut pas calculer la corrélation. Nous obtenons ici un résultat de 0,43.

Remarques


Introduction à la statistique avec R > La corrélation



- Le coefficient de corrélation ne répond pas à toutes les questions
- Les relations quadratiques (en « U »)
- La concordance



9

Pr. Bruno Falissard
 

[7:31] Il est important d'avoir à l'esprit que le coefficient de corrélation ne répond pas à toutes les questions relatives à la quantification de l'association entre deux variables quantitatives.

Non seulement deux variables peuvent être liées alors que leur corrélation est nulle. C'est par exemple le cas si sur un diagramme en x/y vous avez la relation entre y et x qui prend la forme d'une courbe en "U", alors la corrélation sera nulle, même si vous avez une courbe tellement précise que vous pouvez déterminer y en fonction de x .

Mais en plus il y a des situations où la forme de la liaison entre y et x qui intéresse les expérimentateurs ne correspond pas à la corrélation de Pearson. C'est notamment le cas quand on s'intéresse à la concordance entre deux variables aléatoires. Concordance entre deux juges qui évaluent, par exemple, le niveau de la dépression de sujets évalués sur des vidéos. Ou alors, deux biologistes qui tentent de calibrer un appareil, un nouvel appareil qui ne coûte pas très cher au moyen d'un autre appareil qui lui est une référence parce qu'il coûte très cher. Dans ce type de situation, si le nouvel appareil dit systématiquement un résultat deux fois moins important que l'appareil de référence, alors la corrélation sera parfaite, elle vaudra 1 et pourtant le nouvel appareil sera bien mauvais puisque systématiquement il donne un résultat faux.

Conclusion

Introduction à la statistique avec R > La corrélation

```
plot(jitter(smp.c$age), jitter(smp.c$n.enfant))
cor(smp.c$age, smp.c$n.enfant, use="complete.obs")
```

10

Pr. Bruno Falissard

[8:55] Maintenant, vous pouvez refaire les calculs qui ont été abordés dans ce chapitre, il n'y en a pas beaucoup cette fois-ci.