Pièges des corrélations: les coefficients de corrélation de Pearson et de Spearman

Ulrike Held
Horten-Zentrum, UniversitätsSpital, Zürich

En recherche clinique, il arrive fréquemment que l'on mesure plusieurs paramètres chez le patient, par ex. le poids corporel et la pression artérielle systolique. Chacun des paramètres, appelés variables, est examiné séparément. Mais parfois, il est intéressant d'étudier s'il existe des relations entre les deux variables. Par exemple, on pourrait examiner le comportement d'une variable lorsque l'autre diminue ou qu'elle augmente, la nature de la relation, les possibilités de quantification de la relation, c.-à-d. son intensité et son sens. Il existe diverses façons de mesurer la relation ou la liaison, selon qu'il s'agisse de variables *métriques (quantitatives)*, telles que la pression systolique, ou *ordinales* (tels que l'état dans lequel se sent un patient sur une échelle de 1 à 7). Contrairement à la régression linéaire, qui exige de définir une variable explicative et une variable dépendante, cela n'est pas nécessaire dans le cas de la corrélation.

Reprenons l'exemple fictif cité dans l'article «Les questionnements scientifiques de la médecine ont besoin de modèles statistiques»[1]: on a relevé la pression systolique et le poids de 20 patients. Les observations figurent au tableau 1.

Il est à noter qu'il manque la mesure de la variable «poids» chez le patient portant le numéro 6, et celle de la variable «pression systolique» chez le patient 8.

Avant de procéder à un calcul particulier de la mesure de la force de la relation statistique entre deux variables métriques, par ex. à celui de leur coefficient de corrélation, il faut commencer par représenter graphiquement les deux séries de données dans un diagramme de dispersion appelé nuage de points, ou *scatter plot*. Aussi bien la représentation graphique que le calcul du coefficient de corrélation exigent que les couples d'observations soient complets. Dans les deux séries de données de notre exemple, seuls 18 couples seront ainsi pris en compte, et les données des personnes pour qui il manque une des observations (en l'occurrence celles des patients 6 et 8) seront donc exclues de l'analyse statistique. Il faudrait s'assurer encore que l'absence d'une donnée soit indépendante de sa valeur: par ex., il ne faut pas que le manque de données concerne trop fréquemment les personnes dont la pression artérielle est très élevée ou très faible, car les résultats en seraient biaisés. Dans la pratique, il n'existe pas de stratégie universellement valable pour traiter les données manquantes, mais il faut plutôt juger de cas en cas. Dans la figure 1, chaque couple d'observations complet est représenté sous la forme d'un point dont les coordonnées sont les données d'un individu qui correspondent aux deux variables à étudier.

Le coefficient de corrélation de Pearson permet d'évaluer l'intensité et le sens de la relation linéaire entre deux séries de données provenant de l'échantillonnage de deux variables métriques. Le coefficient de corrélation indique le degré de relation linéaire entre les deux séries de données, et il prend des valeurs situées entre –1 et 1. S'il n'y a pas de relation linéaire entre les deux séries de données, le coefficient de corrélation est très proche de zéro, et on dira que les deux variables ne sont pas corrélées. Dans ce cas, il pourrait tout de même y avoir une relation entre les deux variables, mais alors elle ne sera pas linéaire.

Le signe du coefficient de corrélation indique le sens de la corrélation: s'il est positif, la valeur d'une des variables tend à augmenter en même temps que celle de l'autre variable, s'il est négatif, la valeur d'une variable tend à diminuer quand celle de l'autre augmente. Pour interpréter les corrélations dans le cadre de la relation examinée, nous pouvons nous en tenir à certaines valeurs indicatives. Les valeurs situées entre 0,3 et 0,5 (resp. entre –0,3 et –0,5) indiquent une corrélation de faible intensité positive (resp. négative), celles situées entre 0,5 et 0,8 (resp. –0,5 et –0,8) indiquent une corrélation d'intensité moyenne, et au-dessus de 0,8 (resp. en-dessous de –0,8), la corrélation entre les deux variables est considérée comme de forte intensité.

Tableau 1. Mesures du poids corporel et de la pression systolique prises auprès de 20 patients de sexe masculin.

Patient	Poids (kg)	Pression systolique (mm Hg)
1	92,8	153,6
2	82,9	116,2
3	101,6	157,8
4	97,7	153,5
5	111,3	153,2
6	manque	123,3
7	73,3	128,4
8	87,2	manque
9	114,7	126,1
10	113,1	167,9
11	97,4	130,5
12	90,2	141,2
13	95,0	109,9
14	89,4	89,8
15	90,4	137,6
16	92,2	114,1
17	105,1	130,6
18	89,4	138,1
19	88,7	109,4
20	86,3	103,5

Ulrike Held

L'auteur déclare ne pas être en conflit d'intérêt en relation avec cette contribution.

1 U. Held. «Les questionnements scientifiques de la médecine ont besoin de modèles statistiques». Forum Médical Suisse. 2010;10(32):528–30.

En calculant la corrélation de Pearson entre les deux ensembles de données de notre exemple, nous obtenons une intensité de corrélation 0,48.
Dans les cas où le coefficient de corrélation de Pearson semble fortement influencé par des valeurs extrêmes ou aberrantes, et dans ceux où les données proviennent de variables qui ne sont pas métriques, mais ordinales, il est bon d'avoir recours au coefficient de corrélation des rangs selon Spearman. Ce coefficient ne fait pas appel aux valeurs quantitatives, mais simplement à leur rang. Cela signifie que l'on classe les valeurs d'observation par ordre de grandeur et que l'on remplace ensuite la valeur quantitative par le numéro du rang. Le coefficient de corrélation des rangs se situe dans un intervalle de –1 à 1, et les règles d'interprétation sont les mêmes que pour le coefficient de Pearson. Les données de notre exemple sont indiquées dans le tableau 2, rangées selon l'ordre croissant des poids corporels des patients.
Nous pouvons observer que dans la variable poids, la valeur 89,4 apparaît précisément deux fois. Dans ce cas, on attribue deux fois la valeur moyenne de 6,5 au numéro de rang en lieu et place des numéros 6 et 7. On appelle ces rangs des rangs *ex aequo* ou rangs liés.
Dans notre exemple ci-dessus, le coefficient de corrélation des rangs entre poids corporel et pression systolique est estimé à 0,54. Comme seuls 18 couples d'observations sont à disposition, il faut donner dans ce cas la préférence au coefficient de corrélation des rangs plutôt qu'à celui de Pearson et conclure à une corrélation modérée entre le poids corporel et la pression systolique.

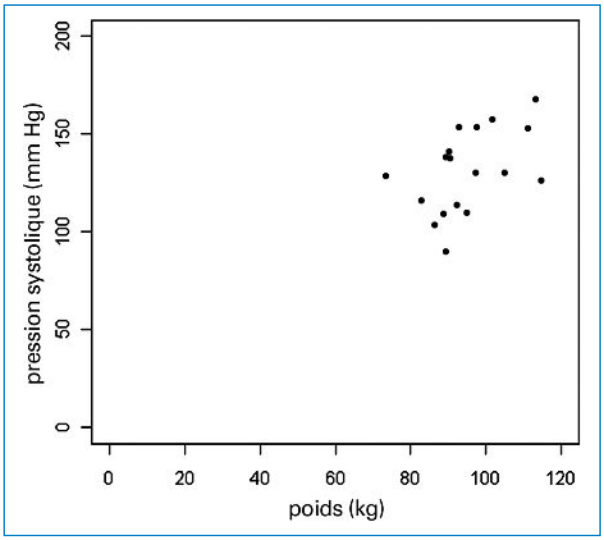

Figure 1
Diagramme de dispersion des 18 couples d'observation.

Tableau 2. Valeurs mesurées et numérotation selon leur rang par ordre croissant de poids.

Patient	Poids (kg)	Rang poids	Pression systolique (mm Hg)	Rang pression systolique
7	73,3	1	128,4	9
2	82,9	2	116,2	6
20	86,3	3	103,5	2
8	87,2	4		
19	88,7	5	109,4	3
14	89,4	6,5	89,8	1
18	89,4	6,5	138,1	13
12	90,2	8	141,2	14
15	90,4	9	137,6	12
16	92,2	10	114,1	5
1	92,8	11	153,6	17
13	95,0	12	109,9	4
11	97,4	13	130,5	10
4	97,7	14	153,5	16
3	101,6	15	157,8	18
17	105,1	16	130,6	11
5	111,3	17	153,2	15
10	113,1	18	167,9	19
9	114,7	19	126,1	8
6			123,3	7

Glossaire

Métrique
On nomme les paramètres «métriques» s'ils se mesurent sur une échelle continue. Exemples: tension artérielle, taille en cm.

Ordinal
On nomme les paramètres «ordinaux» s'ils sont mesurés sur une échelle de rang ordonnancée selon un ordre approprié. Exemples: la satisfaction par rapport à un produit (très satisfait, satisfait, peu satisfait), la performance scolaire (très bien, bien, satisfaisant, etc.).

Rang
Les observations d'une variable ou d'une série de données sont triées par ordre croissant et numérotées: la plus petite observation obtient le numéro de rang 1, la suivante le numéro 2, etc.

Diagramme de dispersion
Dans un diagramme de dispersion, chaque couple d'observations complet est représenté sous la forme d'un point dont les coordonnées sont les données d'un individu qui correspondent aux deux variables à étudier, ce qui donne lieu à un «nuage de points». Anglais: *scatter plot*.

Correspondance:
Dr Ulrike Held
Horten-Zentrum
UniversitätsSpital Zürich
Postfach Nord
CH-8091 Zürich
ulrike.held@usz.ch

Références recommandées
– Bland JM, Altman DG. Statistics notes: correlation, regression, and repeated data. BMJ. 1994;308:896.
– Held L, Rufibach C, Seifert B. Einführung in die Biostatistik. 4. Auflage. Zürich: Abteilung Biostatistik, Institut für Sozial- und Präventivmedizin der Universität Zürich; Juli 2009. http://www.biostat.uzh.ch.
– Hüsler J, Zimmermann H. Statistische Prinzipien für medizinische Projekte. 4. Auflage. Bern: Huber-Verlag; 2006.
– Kreienbrock L, Schach S. Epidemiologische Methoden. 4. Auflage. München: Elsevier-Verlag; 2005.
– R Development Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2008. ISBN 3-900051-07-0. URL http://www.R-project.org.