Tutorial

KARDIOTECHNIK 2022; 031(4):146–149
https://doi.org/10.47624/kt.031.BAND9962

Statistik Teil 4: Korrelationen

M. Kohl, F. Münch

Die KARDIOTECHNIK stellt in der Rubrik Tutorials relevante Methoden für wissenschaftliche Arbeiten zur klinischen Perfusion vor.

Fazitbox

Pro und Contra Korrelationsanalysen:

Pro
- Mit Rangkorrelationen lassen sich auf einfache Weise monotone Zusammenhänge zwischen zwei Variablen untersuchen.
- Rangkorrelationen weisen eine gewisse Robustheit gegenüber Ausreißern auf und sollten verstärkt in medizinischen Anwendungen anstelle der Pearson-Korrelation zum Einsatz kommen. Konfidenzintervalle können z. B. mit Hilfe von Bootstrap berechnet werden.

Contra
- Korrelation darf nicht mit einem Kausalzusammenhang gleichgesetzt werden.
- Die Pearson-Korrelation reagiert sehr sensibel auf Ausreißer und sollte außerdem nur bei Vorliegen eines linearen Zusammenhangs angewendet werden.
- Korrelationsanalysen eignen sich nicht zur Untersuchung von nicht-monotonen Zusammenhängen.

Einführung

In der modernen Medizin ist man fortwährend auf der Suche nach möglichen Biomarkern, mit denen man etwa Erkrankungen diagnostizieren, den Therapieverlauf verfolgen oder eine Prognose über den Krankheitsverlauf stellen kann. Statistisch gesprochen geht es darum, klinisch relevante Zusammenhänge (Assoziationen) zwischen erhobenen Messgrößen (Variablen) zu erkennen und zu quantifizieren. Eine einfache statistische Möglichkeit hierfür besteht in der sogenannten Korrelationsanalyse, mit deren Hilfe gewisse Zusammenhänge zwischen zwei Variablen X und Y bestimmt werden können. Es ergeben sich bei der Korrelation immer Werte zwischen -1 und +1, welche die Stärke und die Richtung des Zusammenhangs widerspiegeln. Aber Achtung, eine (signifikante) statistische Korrelation zwischen zwei Variablen ist nicht gleichbedeutend mit einem Kausalzusammenhang zwischen diesen beiden Variablen. Man spricht in diesem Fall auch von einer Scheinkorrelation („spurious correlation“) [1], wobei man dies besser eine Scheinkausalität nennen sollte. Ein klassisches Beispiel hierfür ist der Zusammenhang zwischen der Geburtenrate und der Anzahl der Störche. Matthews (2000) etwa präsentiert Daten aus den Jahren 1980–1990, die eine signifikant positive Korrelation von 0,62 zwischen der Geburtenrate beim Menschen und der Anzahl brütender Storchpaare in 17 europäischen Ländern belegen [2]. Eine Vielzahl weiterer Beispiele für Scheinkorrelationen finden sich z. B. auf den Internetseiten von N. Zellmer (https://scheinkorrelation.jimdofree.com/) und T. Vigen (https://www.tylervigen.com/spurious-correlations). Da sich aus verschiedensten Gründen eine statistisch signifikante Korrelation ergeben kann (Abb. 1), sollte diese immer auf seine fachliche Plausibilität geprüft und mit großer Vorsicht als ein möglicher Kausalzusammenhang interpretiert werden.

Prof. Dr. Matthias Kohl
Department of Medical and Life Sciences
Institute of Precision Medicine
Hochschule Furtwangen
Jakob-Kienzle-Str. 17,
78054 Villingen-Schwenningen (Germany)
E-Mail: kohl@hs-furtwangen.de
www.hs-furtwangen.de
www.life-data-science.org

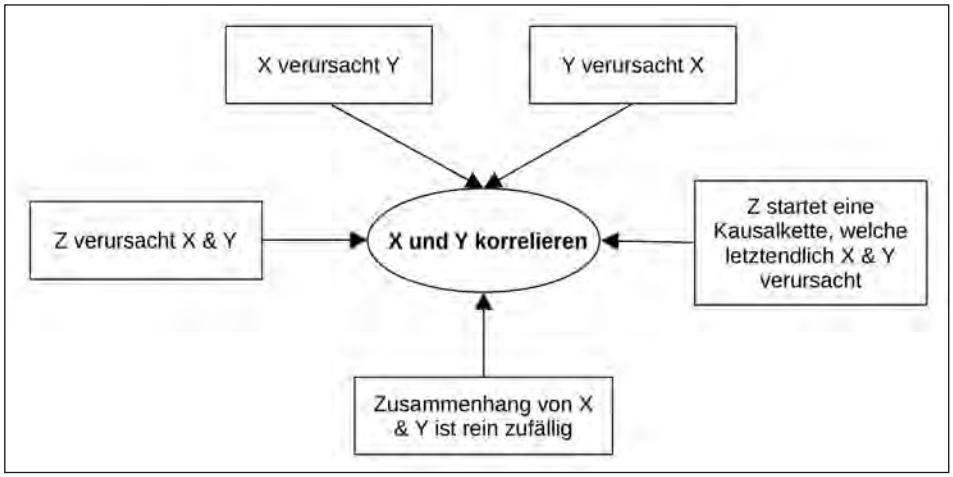

Abb. 1: Möglichkeiten für eine Korrelation zwischen den Variablen X und Y; in Anlehnung an [3]

Bravais-Pearson-Korrelation

Der wohl am häufigsten verwendete Korrelationskoeffizient ist die sogenannte Bravais-Pearson-Korrelation oder auch kurz Pearson-Korrelation (Abb. 2).

Es kann damit die Stärke des linearen Zusammenhangs zwischen zwei metrischen Variablen X und Y bestimmt werden. Die Pearson-Korrelation r_{xy} ist folglich eng mit der einfachen linearen Regression verwandt ($Y = \beta_0 + \beta_1 * X + \varepsilon$, mit Achsenabschnitt β_0, Steigung β_1 und Zufallsschwankung ε) und ist insbesondere nicht dazu geeignet, nicht-lineare Zusammenhänge zu untersuchen. Speziell erhält man für das Bestimmtheitsmaß R^2 der einfachen linearen Regressionsanalyse $R^2 = (r_{xy})^2$ (vgl. Abschnitt 5.2 in [4]). Gilt $r_{xy} = -1$ oder $r_{xy} = +1$, so liegt ein perfekter linearer Zusammenhang ($Y = \beta_0 + \beta_1 * X$) mit $R^2 = 1$ vor. In Abbildung 3 sind einige Beispiele für andere Pearson-Korrelationen dargestellt. Wir sehen, dass sich bei einem linearen Zusammenhang zwischen den beiden Variablen ellipsenförmige Punktwolken ergeben sollten. Diese Tatsache kann umgekehrt zur Diagnose genutzt werden, um im Nachhinein festzustellen, ob die Annahme eines linearen Zusammenhangs gerechtfertigt war. Diese Annahme ist insbesondere erfüllt, falls die Werte von X und Y einer gemeinsamen (bivariaten) Normalverteilung folgen, womit auch die Werte von X und Y, separat betrachtet, jeweils normal-

Gegeben sei eine Stichprobe (x_i, y_i) $(i = 1, \ldots, n)$ von paarweisen Beobachtungen zweier metrischer Variablen mit Mittelwerten

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{1}$$

(empirischen) Standardabweichungen

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

und (empirischer) Kovarianz

$$\mathrm{Kov}_{x,y} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \tag{3}$$

Dann gilt für die Bravais-Pearson-Korrelation r_{xy}

$$r_{xy} = \frac{\mathrm{Kov}_{x,y}}{S_x * S_y} \in [-1, 1] \tag{4}$$

Abb. 2: Definition der Bravais-Pearson-Korrelation

verteilt sein müssen. In manchen Anwendungen lässt sich die bivariate Normalverteilung mit Hilfe einer Transformation von einer oder beiden Variablen erreichen.

Es sollte weiter beachtet werden, dass die Pearson-Korrelation sehr leicht durch Ausreißer verfälscht werden kann. Die Einschränkungen der Pearson-Korrelation lassen sich z. B. anhand des Quartetts von Anscombe verdeutlichen [8]. Es handelt sich dabei um vier Datensätze, die zu sehr ähnlichen Pearson-Korrelationen und Regressionsgeraden führen, wobei diese Analysen aber nur für den ersten der vier Datensätze statistisch sinnvoll sind (Abb. 4). Im Fall des zweiten Datensatzes liegt ein nicht-linearer und nicht-monotoner Zusammenhang vor, welcher nicht mit einer Korrelationsanalyse, sondern z. B. mit Hilfe einer nicht-linearen Regressionsanalyse genauer untersucht werden könnte. Der dritte Datensatz enthält einen Ausreißer in y-Richtung. Hier kann ein robuster Korrelationskoeffizient bzw. eine robuste lineare Regressionsanalyse Abhilfe schaffen. Berechnet man für diesen Datensatz die Korrelation z. B. mit dem robusten Minimum Covariance Determinant (MCD)-Schätzer [9], so ergibt sich eine Korrelation von 0,9999966. Beim vierten Datensatz wurde entweder ein schlechtes Versuchsdesign gewählt oder es liegt ein Ausreißer in x-Richtung vor. In diesem Fall könnte die Erhebung weiterer Daten hilfreich sein und Klarheit bringen.

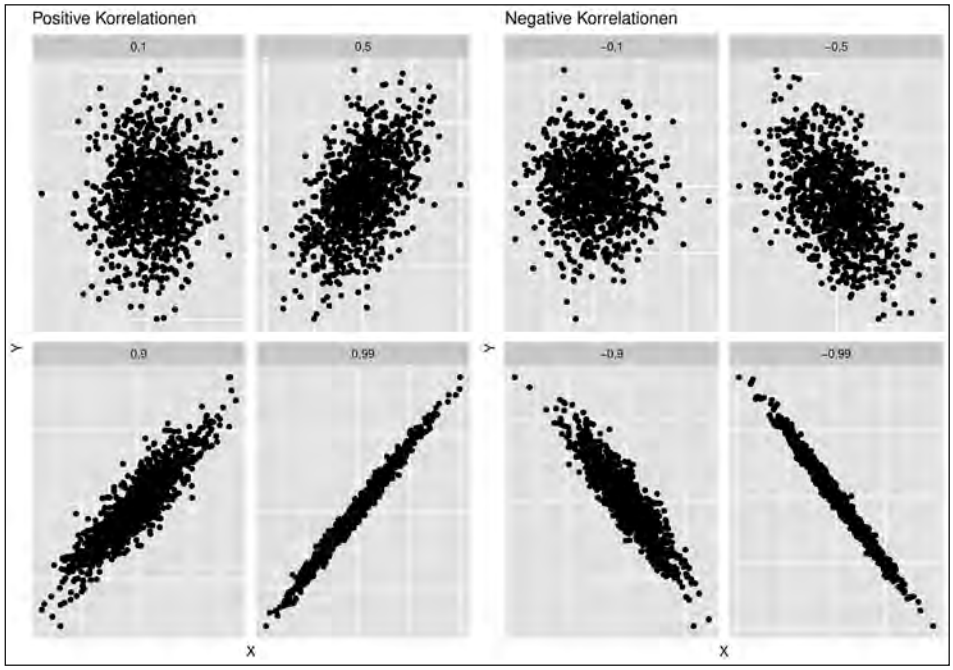

Abb. 3: Beispiel für Pearson-Korrelationen (erstellt mit der Statistiksoftware R [5] und den R Paketen MKdescr [6] und ggplot2 [7])

Rangkorrelation nach Spearman

Da der Korrelationskoeffizient ρ von Spearman (kurz: Spearmans ρ) auf Rängen basiert, kann dieser nicht nur bei metrischen, sondern auch bei ordinalen Variablen angewendet werden. Für die Berechnung müssen die beobachteten Werte für X und Y zunächst aufsteigend angeordnet werden. Der kleinste Wert von x_1, x_2, …, x_n bekommt den Rang 1, der größte Wert den Rang n. Entsprechend verfährt man mit y_1, y_2, …, y_n. Treten gleich große Werte auf (sog. Bindungen), wird für alle diese Werte der Mittelwert der Ränge verwendet. Anstelle der Beobachtungspaare (x_i, y_i) (i = 1, 2, …, n) werden die Rangpaare (R_{xi}, R_{yi}) für die Berechnungen herangezogen und in die Gleichungen (1) – (4) von Abbildung 2 eingesetzt (vgl. Abschnitt 5.4 in [4]). Hierbei wird vorausgesetzt, dass ein linearer Zusammenhang für die Ränge der Beobachtungen vorliegt. Überträgt man diese Voraussetzung auf die ursprünglichen Beobachtungen, bedeutet dies, dass die Spearman-Korrelation angewendet werden kann, wenn es einen monotonen Zusammenhang zwischen den beiden Variablen X und Y gibt. Bei ordinalen Variablen sollten außerdem die Abstände zwischen den möglichen Werten der Variablen möglichst gleich groß (äquidistant) sein, da bei der Berechnung des Korrelationskoeffizienten auch die Differenzen der Ränge eingehen. Dies trifft z. B. auf viele medizinische Scores zu. Ergibt sich ρ = -1 oder ρ = +1, so liegt ein streng monotoner Zusammenhang vor. Durch die Betrachtung der Ränge besitzt die Spearman-Korrelation, wie auch andere Rangstatistiken, eine gewisse Robustheit gegenüber Ausreißern [10]. Wenden wir Spearmans ρ auf den dritten Anscombe-Datensatz (Abb. 4) an, so ergibt sich eine deutlich höhere Korrelation als im Fall der Pearson-Korrelation mit ρ = 0,991.

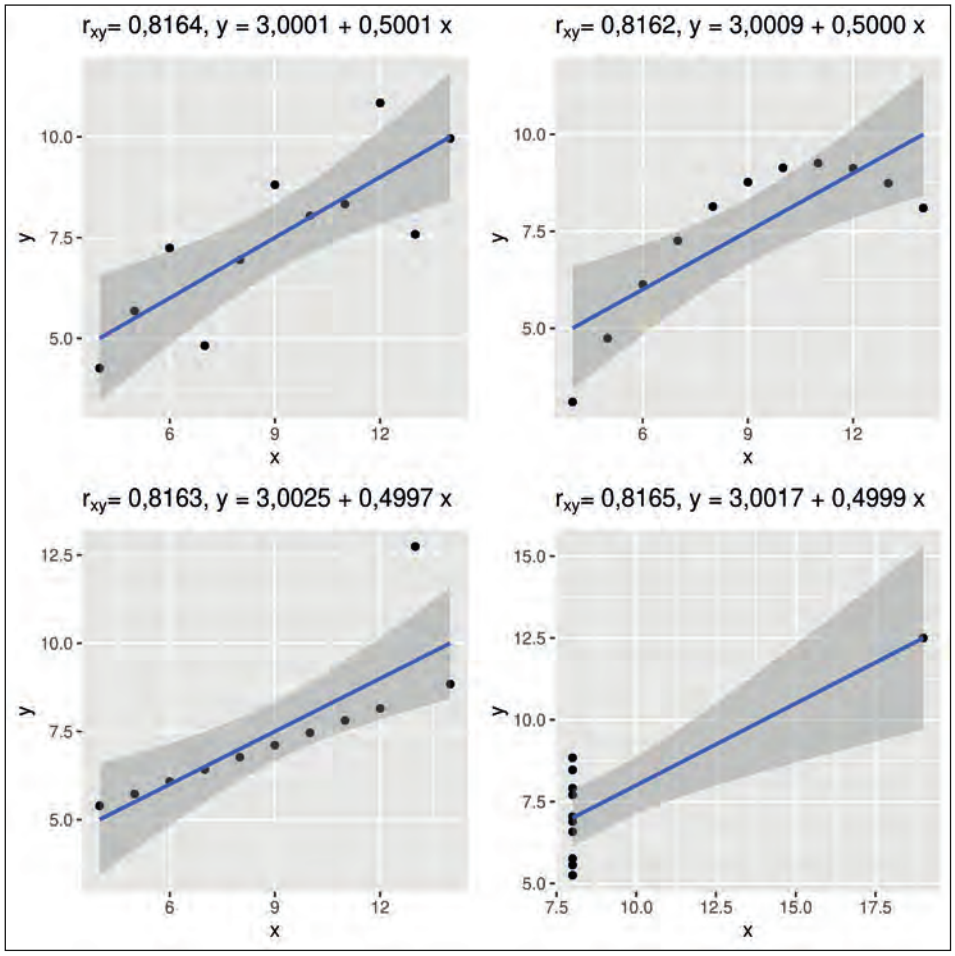

Abb. 4: Das Quartett von Anscombe (erstellt mit den R Paketen datasets [5] und ggplot2 [7])

Rangkorrelation nach Kendall

Für die Berechnung des Korrelationskoeffizienten τ von Kendall (kurz: Kendalls τ) werden erneut die Ränge der Beobachtungen verwendet. Im Unterschied zu Spearmans ρ werden aber keine Rangdifferenzen benötigt, so dass die Kendall-Korrelation auch bei ordinalen Variablen mit ungleichmäßigen Abständen zwischen den möglichen Werten der Variablen anwendbar ist. Dies kann etwa bei den Antwortmöglichkeiten von Fragebögen der Fall sein. Für die Details zur Berechnung von Kendalls τ verweisen wir auf Abschnitt 3.2.5 von [11]. Es gilt, wie im Fall von Spearmans ρ, dass im Fall $\tau = -1$ oder $\tau = +1$ ein streng monotoner Zusammenhang vorliegt. Analog zur Spearman-Korrelation besitzt auch die Kendall-Korrelation eine gewisse Robustheit gegenüber Ausreißern. Für den dritten Anscombe-Datensatz (Abb. 4) erhalten wir $\tau = 0{,}964$. Dies deckt sich mit der Beobachtung, dass die Kendall-Korrelation oftmals etwas kleiner als die Spearman-Korrelation ist.

Konfidenzintervalle der Korrelationskoeffizienten

Bei der Berechnung von Konfidenzintervallen (und statistischen Tests) für die vorgestellten Korrelationskoeffizienten kommen in der Regel approximative Ansätze zum Einsatz, die entweder auf der Normalverteilung basieren (vgl. Abschnitt 7.8 in [11]) oder Resampling-Verfahren (Permutation, Bootstrap) nutzen (vgl. [12]).

Stärke der Korrelation

Für die Beschreibung der beobachteten Korrelation ist es hilfreich, die möglichen Werte in Kategorien einzuteilen. Hierfür gibt es verschiedene Möglichkeiten, die auch vom Anwendungsbereich abhängen. Eine mögliche und aus unserer Sicht für medizinische Anwendungen sinnvolle Einteilung findet sich in Tab. 1 [13].

Beispiel: Ischämiezeit und Troponin

Wir untersuchen den Zusammenhang zwischen Ischämiezeit und Troponin auf Basis der Daten von 140 Patient:innen (eigene Daten). Wir erwarten einen monotonen wachsenden Zusammenhang, da eine längere Ischämiezeit zu größeren Zellschädigungen führen sollte, welche durch einen Anstieg der Troponinwerte angezeigt werden sollte. Das Streudiagramm der Originaldaten in Abbildung 5 bestätigt die Vermutung, wobei wir auch vereinzelte Ausreißer sowie eine deutliche Zunahme der Streuung der Troponinwerte mit der Ischämiezeit (Heteroskedastizität) sehen. Im Fall einer Heteroskedastizität kann oftmals mit Hilfe einer log-Transformation eine Varianzstabilisierung und Annäherung an die Normalverteilung (Symmetrisierung) erreicht werden. Dies gelingt auch hier recht gut, wie das rechte Streudiagramm in Abbildung 5 zeigt. Wir erhalten eine Annäherung der Punktewolke an eine Ellipsenform. Zur Berechnung der 95 %-Konfidenzintervalle (CI95) für die Korrelationskoeffizienten verwenden wir die Bootstrap-Bc_a-Methode mit 9999 Wiederholungen (vgl. Supplement von [12] – auch über diesen QR-Code einsehbar:).

Wir erhalten eine Spearman-Korrelation von $\rho = 0{,}566$ (CI95: 0,426–0,677) und eine Kendall-Korrelation von $\tau = 0{,}406$ (CI95: 0,297–0,491), wobei bei den Rängen die log-Transformation als streng monotone Transformation zu keiner Veränderung der Ergebnisse führt. Für die Pearson-Korrelation erhalten wir mit Hilfe der log10-Transformation $r_{xy} = 0{,}570$ (CI95: 0,429–0,669) in sehr guter Übereinstimmung mit der Spearman-Korrelation. Im Fall des robusten MCD-Schätzers erhalten wir ebenfalls ein ähnliches Ergebnis mit einer Korrelation von 0,563 (CI95: 0,273–0,771). Da die Untergrenzen der 95 %-Konfidenzintervalle jeweils größer 0 sind, dürfen wir die entsprechenden Korrelationen auch als signifikant größer als 0 bezeichnen. Wir erhalten demnach eine signifikant positive, moderate Korrelation zwischen der Ischämiezeit und der Troponin- bzw. log10-Troponin-Konzentration, was sich mit unserer Ausgangshypothese deckt.

Zusammenfassung

Die Pearson-Korrelation sollte, um Fehlinterpretationen zu vermeiden, nur zum Einsatz kommen, wenn ein Datensatz sicher frei von Ausreißern ist und wenn bereits vor der Analyse bekannt bzw. zumindest plausibel ist, dass ein linearer Zusammenhang zwischen den beiden Variablen vorliegt. Dies ist insbesondere erfüllt, falls eine bivariate Normalverteilung vorliegt. Die Annahme eines linearen Zusammen-

Bezeichnung der Korrelation	Positive Korrelationen	Negative Korrelationen
sehr stark	0,9 bis 1,0	-1,0 bis -0,9
stark	0,7 bis 0,9	-0,9 bis -0,7
moderat	0,4 bis 0,7	-0,7 bis -0,4
schwach	0,1 bis 0,4	-0,4 bis -0,1
vernachlässigbar	0,0 bis 0,1	-0,1 bis -0,0

Tab. 1: Mögliche Bezeichnungen für die Stärke einer beobachteten Korrelation

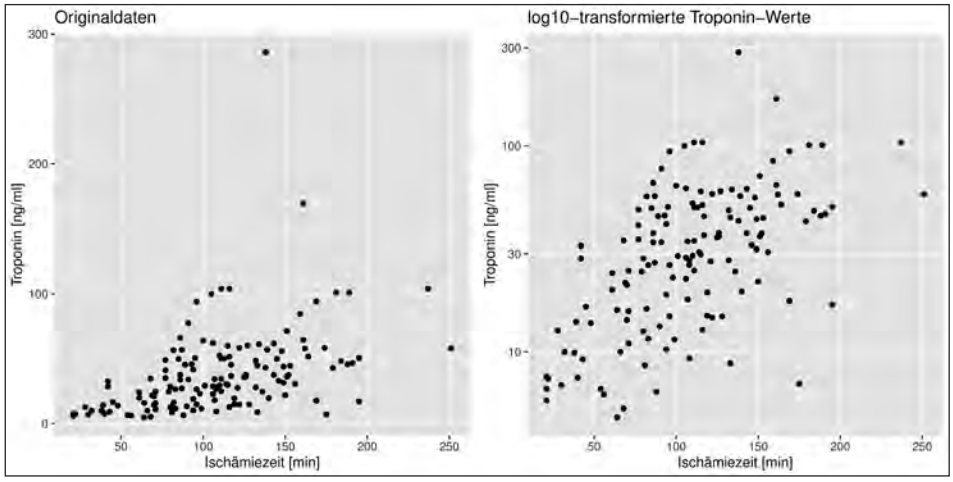

Abb. 5: Troponinkonzentration in ng/ml in Abhängigkeit von der Ischämiezeit in min (erstellt mit R Paket ggplot2 [7])

hangs kann bei ausreichend großen Datensätzen gut mit Hilfe eines Streudiagramms überprüft werden, welches eine ellipsenförmige Punktwolke zeigen sollte. Leider haben wir den Eindruck, dass die oben genannten Einschränkungen für die Pearson-Korrelation nicht allseits bekannt sind bzw. beachtet werden. Daher sollte man Pearson-Korrelationen, die berichtet werden, immer mit großer Vorsicht betrachten. Die Rangkorrelationen von Spearman und Kendall reagieren im Unterschied zur Pearson-Korrelation deutlich weniger sensibel auf Ausreißer und sind außerdem allgemeiner bei monotonen Zusammenhängen einsetzbar. Hierbei sollte die Kendall-Korrelation vor allem bei ordinalen Variablen mit ungleichmäßigen Abständen zwischen den möglichen Werten der Variablen zum Einsatz kommen. In allen anderen Fällen ist in der Regel die Spearman-Korrelation vorzuziehen. Aufgrund der oben angeführten Argumente empfehlen wir in medizinischen Anwendungen verstärkt Rangkorrelationen anstelle der Pearson-Korrelation zu verwenden.

LITERATUR

1. Simon HA. Spurious Correlation: A Causal Interpretation. Journal of the American Statistical Association 49, 1954; no.267:467–479.

2. Matthews R. Storks Deliver Babies (p= 0.008). Teaching Statistics 2000; 22: 36-38.

3. Seite „Cum hoc ergo propter hoc". In: Wikipedia – Die freie Enzyklopädie. Bearbeitungsstand: 30. August 2022; 04:57 UTC. URL: https://de.wikipedia.org/w/index.php?title=Cum_hoc_ergo_propter_hoc&oldid=225752253 (Abgerufen: 10. September 2022, 07:43 UTC)

4. Weiß C. Basiswissen Medizinische Statistik. 2019; 7. Auflage, Springer-Verlag.

5. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing 2022; Vienna, Austria. URL https://www.R-project.org/.

6. Kohl M. MKdescr: Descriptive Statistics. R package version 0.7.; 2021.

7. Wickham H. ggplot2: Elegant Graphics for Data Analysis. 2016; Springer-Verlag New York.

8. Anscombe FJ. Graphs in statistical analysis. The American Statistician 1973; 27:17-21.

9. Rousseeuw PJ, Leroy AM. Robust Regression and Outlier Detection. 1987; Wiley.

10. Rieder H. A general robustness property of rank correlations. Communications in Statistics - Theory and Methods, 1982; 11(3):233–234.

11. Hedderich J, Sachs L. Angewandte Statistik. Methodensammlung mit R. 2020; 17. Auflage, Springer-Verlag.

12. Kohl M, Münch F. Statistik Teil 3: Konfidenzintervalle. Kardiotechnik 2022; 31(3):95-98; doi: 10.47624/kt.031.QQOV9624.

13. Schober P, Boer C, Schwarte L. Correlation Coefficients: Appropriate Use and Interpretation. Anesthesia & Analgesia 2018; 126(5):1763-68.