Korrelation

Walter Gruber

2025-03-14

Contents

Einleitung

Korrelationen spielen eine zentrale Rolle in der Statistik und in vielen wissenschaftlichen Disziplinen. Sie ermöglichen es uns, Zusammenhänge zwischen zwei oder mehr Variablen zu quantifizieren und zu interpretieren.

Dieses Skriptum frischt Ihre bestehenden Grundkenntnisse über Korrelation auf und vertieft Ihr Verständnis der verschiedenen Korrelationstechniken. Ziel ist es, Ihnen die Werkzeuge an die Hand zu geben, um komplexe

Datenbeziehungen zu analysieren und fundierte Schlussfolgerungen zu ziehen.

Variabilität

Bevor wir uns mit einzelnen Techniken und Verfahren der linearen Modellbildung auseinandersetzen, soll in einem kurzen Exkurs eines der grundlegensten Prinzipien der statistischen Modellbildung wiederholt und diskutiert werden - die *Varianz* von beobachten Werten.

Eigentlich ist es die Variabilität von Merkmalen, die statistische Methoden für die Erklärung von Effekten überhaupt erst auf den Plan ruft. Würden Merkmale wie z.B. Leistung einer Person, Persönlichkeitsmerkmale, Wetter, Produktionsgenauigkeit etc. nicht schwanken/variieren, würden wir heute nicht in diesem Raum sitzen und uns mit statistischen Modellen beschäftigen.

Der Begriff Variabilität ist für uns so alltäglich, dass wir ganz selbstverständlich damit umgehen. Doch was steckt wirklich dahinter? Wie können wir Sie nutzen um komplexere Eigenschaften einer Sache oder eines unerklärlichen Phenomäns auf die Spur zu kommen?

Betrachten wir zunächst einmal ein sehr einfaches Beispiel. In den nachfolgenden Graphen sind (sehr vereinfacht) mehrere Möglichkeiten dargestellt, wie eine Person mitsamt Hund sich entlang einer Straße bewegt.

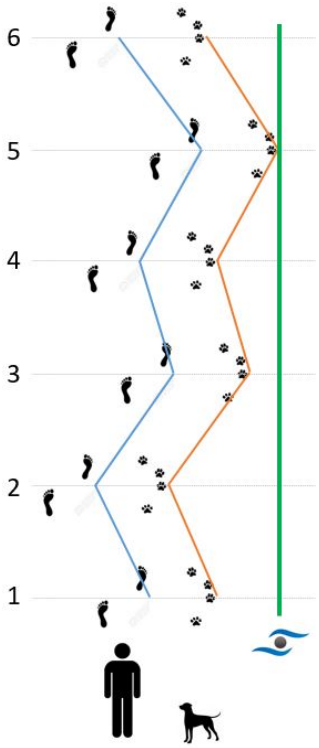


Figure 1: **Abbildung 1**: Gassi gehen mit *Blindenhund*. Die blaue Linie beschreibt den Weg des Hundehalters, die Orange den des Hundes. Die Grüne Line ist die Referenzlinie, von welcher aus der Abstand zur jeweiligen Position (Hund und Mensch) zu sechs Beobachtungszeitpunkten gemessen wurde.

Die Daten der Messungen sind in folgender Tabelle gegeben:

Mensch	Hund	MenschD	HundD	KP	ZMensch	ZHund	ZKP
4	3	0.83	1	0.83	0.71	0.71	0.5
5	4	1.83	2	3.66	1.56	1.42	2.22

Mensch	Hund	MenschD	HundD	KP	ZMensch	ZHund	ZKP
2	1	-1.17	-1	1.17	-1	-0.71	0.71
3	2	-0.17	0	0	-0.15	0	0
2	0	-1.17	-2	2.34	-1	-1.42	1.42
3	2	-0.17	0	0	-0.15	0	0

- In obiger Tabelle zeigen die Spalten *Mensch* und *Hund* jeweils den Abstand zur gedachten Beobachtungslinie pro Beobachtungszeitpunkt.
- Die Spalten *MenschD* und *HundD* ist jeweils die Differenz jeder Beobachtung zum jeweiligen Mittelwert aller Beobachtungen, also $MD_i = M_i - \bar{M}$ und $HD_i = H_i - \bar{H}$ mit $i \in \{1, 6\}$.
- Die Spalte *KP* zeigt das Kreuzprodukt, also $KP_i = MD_i \cdot HD_i$ mit $i \in \{1, 6\}$.
- Die Spalten *ZMensch* und *ZHund* entsprechen den z-transformierten Werten, also $z_i^M = (M_i - \bar{M})/sd(M)$ und $z_i^H = (H_i - \bar{H})/sd(H))$ mit $i \in \{1, 6\}$.
- Die Spalte *ZKP* entspricht dem Kreuzprodukt der z-Transformierten Werte, also $ZKP_i = ZM_i \cdot ZH_i$ mit $i \in \{1, 6\}$.

Statistische Kennwerte für obige Daten (Mittelwert, Varianz und Standardabweichung) sind in folgender Tabelle dargestellt:

	Mensch	Hund	MenschD	HundD	KP	ZMensch	ZHund	ZKP
Mean	3.167	2	-0.003	0	1.333	-0.005	0	0.808
Var	1.367	2	1.367	2	2.052	0.997	1.008	0.756
SD	1.169	1.414	1.169	1.414	1.433	0.998	1.004	0.869

Wenn also ein Hund jeder Bewegung der Person folgt und dabei auch stets denselben Abstand hält, sind deren beobachteten Pfade zwar örtlich gesehen unterschiedlich, aber die Varianz des einen erklärt vollständig die Varianz des anderen Pfades. Mit anderen Worten, die beiden Pfade zeigen eine perfekte Kovariation. Formal wird diese *Kovariation* als *durchschnittliche Summe der Kreuzprodukte* ermittelt, also (M = Mensch, H = Hund):

$cov(M, H) = \frac{\sum_{i=1}^{6}(M_i - \bar{M}) \cdot (H_i - \bar{H})}{N-1} = 1.60$

Setzt man die Beispieldaten in diese Berechnungsvorschrift ein, erhält man für die Summe der Kreuzprodukte 8. Die Kovarianz der beobachteten Werte ist (als durchschnittliche Kreuzproduktsumme) somit $cov(M, H) = 1.6$. Die Korrelation berechnet sich dann ganz einfach zu:

$r(M, H) = \frac{cov(M,H)}{s_M \cdot s_H} = \frac{1.6}{1.169 \cdot 1.414} = 0.968 \simeq 1$

Im vorliegenden Beispiel ergibt sich eine Korrelation $r(M, H) = 0.97$ (also eine positive und perfekte Korrelation).

Der Korrelationskoeffizient hat gegebüber der Kovarianz den Vorteil, dass er durch die Normierung über die Standardabweichungen:

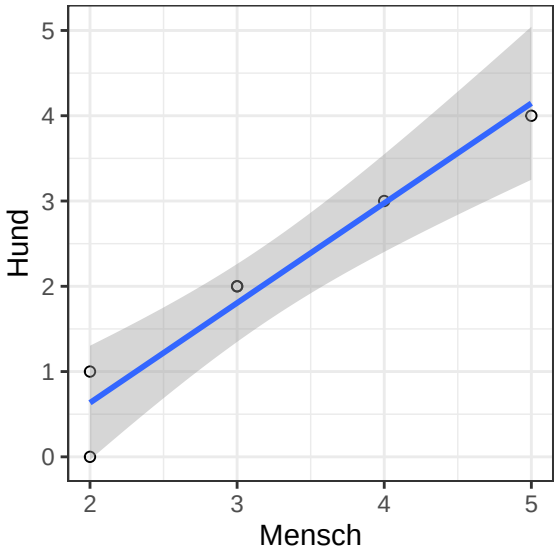
1. einen (einheitenlosen) Wertebereich zwischen $r \in [-1, 1]$ aufweist und damit vergleichbar mit anderen Korrelationswerten wird.
2. Als praktische **Effektgröße** interpretiert werden kann.
3. Das Quadrat des Korrelationskoeffizienten (r^2, auch **Varianzaufklärung**, **Determinationskoeffizient** genannt) Auskunft über die aufgeklärte Varianz gibt.

Letzteres Maß spielt eine wesentliche Rolle sowohl bei der Korrelationsanalyse, als auch bei der multiplen Regression und anderen Verfahren.

Die Bedeutung der Kovarianu sei anhand des verwendeten Beispiels nochmals verdeutlicht:

Im Fall einer perfekten Kovarianz (also 100% Übereinstimmung der Bewegungen von Mensch und Hund), braucht man nur mehr die Bewegung einer Variablen zu wissen (z.B. die des Menschen), um die Bewegungen des Hundes zu bestimmen (erklären). Somit erklärt die Variabilität der Bewegung vom Menschen zu 100 % die Variabilität der Bewegungen des Hundes.

Die Beziehung zwischen zwei (intervallskallierten) Variablen lässt sich am besten mit einem Streudiagramm darstellen:



Die folgende Abbildung zeigt ein weiteres Mensch-Hund Beispiel:

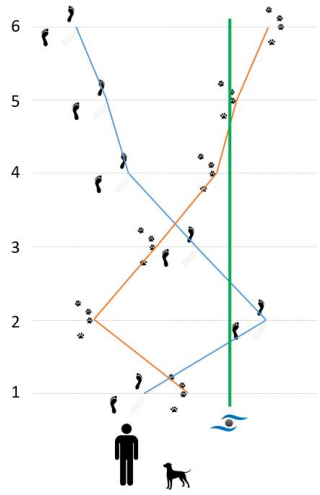
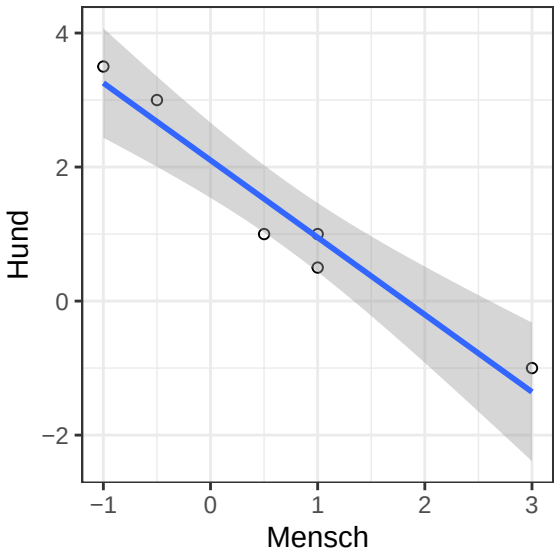


Figure 2: **Abbildung 2**: in diesem Beispiel scheint es sich um eine Hund zu handeln, der bestmöglich das Gegenteil vom Menschen macht. Bestmöglich dahingehen, dass er nicht nur in die genau entgegengesetzte Richtung ausweicht, sondern dabei auch auf den genauen Abstand der Abweichung achtet.

In diesem Fall ist die Korrelation auch perfekt, nur eben in die entgegengesetzte Richtung, was zur Folge hat, dass diese Korrelation den Wert $r(M, H) = -1$ zeigt.



Interessant und der Praxis am ehesten entsprechend, sind jedoch Fälle, in denen zwei Variablen nur teilweise Gemeinsamkeiten aufweisen. Im folgenden Beispiel wäre

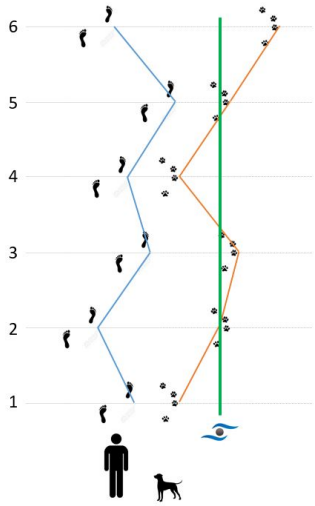
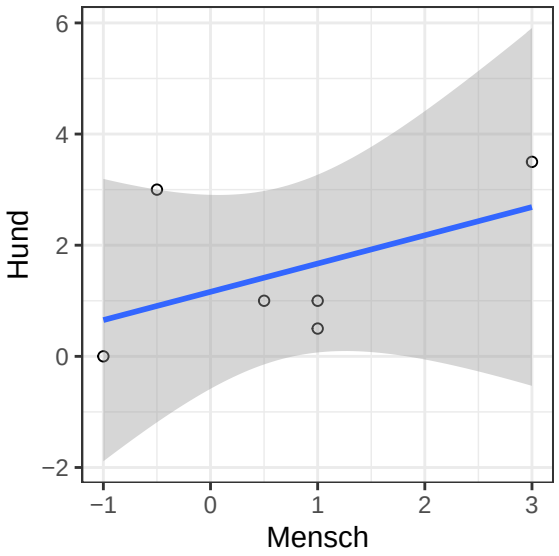


Figure 3: **Abbildung 3**: Hund und Mensch bewegen sich zum Teil unabhängig, zum Teil aber auch synchron. Dies entspricht dann einer Kovarianz, bzw. Korrelation die irgendwo zwischen $r \in [-1, 1]$ liegt (im Beispiel ist $r(M, H) = 0.5$ und somit $r^2 = 0.25$).



Die Korrelation liegt in diesem Beispiel bei $r(M, H) = 0.5$. Daraus lässt sich auch nochmals eine sehr wichtige Erkenntnis bezüglich der geteilten Varianz der beiden Variablen festhalten:

> Bei einer Korrelation von $r(x, y) = 0.5$ entspricht der Determinationskoeffizient $r^2(x, y) = 0.25$. In Prozent ausgedrückt, werden als 25% der Variabilität einer Variablen (z.B. Hund) durch die Variable Mensch erklärt. In welchen Abschnitten der Daten diese gemeinsame Variablität auftritt, lässt sich durch den r nicht bestimmen.

Diese Feststellung führt uns aber zu einer weiteren Betrachtung von Variablitäten:

Würde man davon ausgehen, dass sich der Mensch und Hund bei jedem Messzeitpunkt (1 bis 6) jeweils auf der gleichen Höhe befunden haben, dann wird man eine hohe **negative Korrelation** erhalten. Nimmt man jedoch an, dass der Mensch zum Messzeitpunkt (MZP) 1, der Hund aber bereits auf MZP 2 war, dann verschiebt sich die Spur des Hundes einfach um einen MZP nach oben! Korreliert man nun diese beiden Beobachtungen, würde sich eine nahezu perfekte **positive Korrelation** ergeben!

> Durch schrittweises Verschieben der Werte einer Variablen um eine Einheit (τ_i) mit anschließender Berechnung der Korrelationskoeffizienten ($r_{\tau_i}(x, y)$), erhält man in Abhängigkeit von der Anzahl der Verschiebungen ($i \in [0, N-1]$) maximal N neue Korrelatinskoeffizienten. Man bezeichnet diese Art der Korrelationsberechnung als **Kreuzkorrelation**.

Für das Beispiel ergibt sich eine normale Korrelation von $r(M, H) =$ -1. Die Korrelationen berechnet nach dem Versatzprinzip ergeben folgendes Bild:

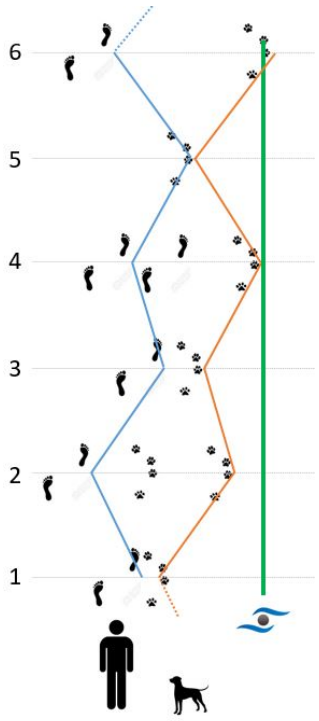
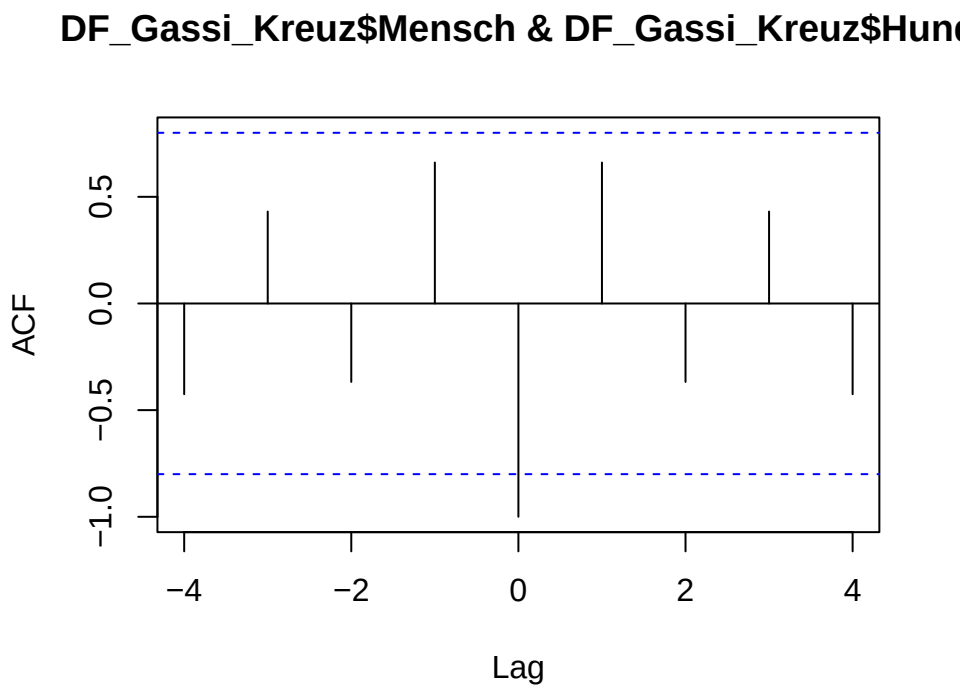


Figure 4: **Abbildung 4**: Hund und Mensch scheinen sich wieder synchron, aber in gegenseitiger Richtung zu bewegen. Man würde also eine negative und hohe Korrelation erwarten. Interressant ist jedoch die Beobachtung, dass ein Versatz der Beobachtungen um eine Einheit zu einem hohen positiven Zusammenhang führen würde!

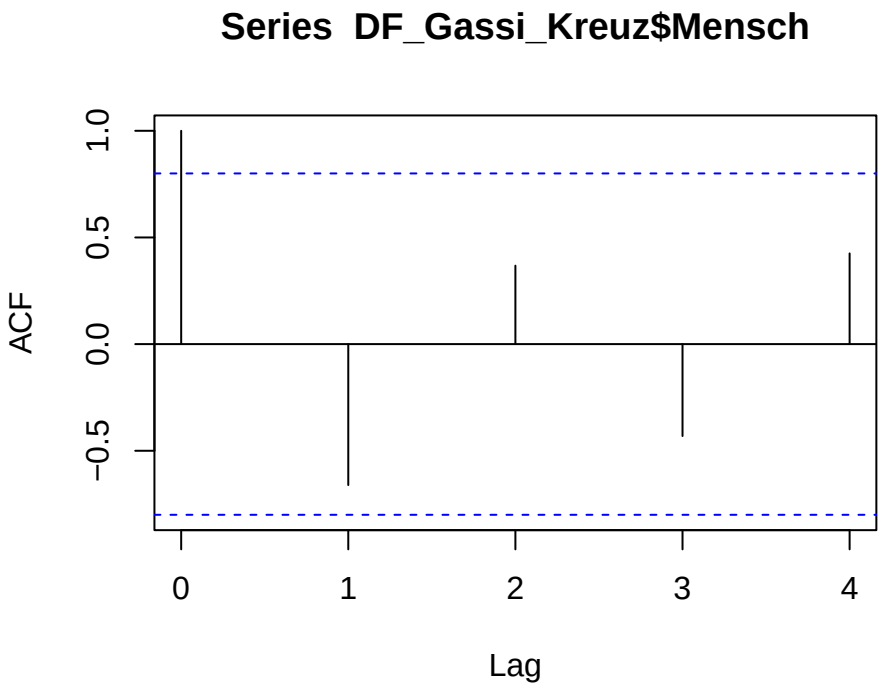


Die Verschiebung τ wurde in diesem Beispiel mit $i = 4$ angegeben, d.h. es wurden die Werte der Variablen Hund um jeweils vier Schritte nach links und vier Schritte nach rechts verschoben. Bei jeder Verschiebung wurde die Korrelation berechnet (im Graphen ist die Verschiebung mit *Lag* auf der x-Achse angegeben). Auf der y-Achse wird der entsprechende Korrelationskoeffient angezeigt.

Tau	CrossCorr
-4	-0.4253
-3	0.431
-2	-0.3678
-1	0.6609
0	-1
1	0.6609
2	-0.3678
3	0.431
4	-0.4253

Die Werte der Tabelle zeigen nochmals den krassen Wechsel der Korrelation zwischen den Werte $\tau = 0$ (also keiner Verschiebung) und $\tau = 1$. Werden die Werte um nur einen Beobachtungspunkt verschoben, ändert sich die Korrelation von einer perfekt negativen, zu einer sehr hohen positiven Korrelation!

Eine weitere wichtige Eigenschaft die mit Hilfe dieser Vorgehensweise geprüft werden kann, ist die der sogenannten **Autokorrelation**. Diese funktioniert im Prinzip wie die eben beschriebene Kreuzkorrelation, mit dem Unterschied, dass eine Variable mit verschobenen "Eigenversionen" korreliert wird. Folgendes Beispiel zeigt das Ergebnis für die Variable Mensch unseres Beispiels:



Die Verschiebung τ wurde in diesem Beispiel mit $i = 4$ angegeben, d.h. es wurden die Werte der Variablen Mensch ebenfalls schrittweise in eine Richtung verschoben. Bei jeder Verschiebung wurde die Korrelation berechnet (im Graphen ist die Verschiebung mit *Lag* auf der x-Achse angegeben). Auf der y-Achse wird der entsprechende Korrelationskoeffient angezeigt.

Tau	CrossCorr
0	1
1	-0.6609
2	0.3678
3	-0.431
4	0.4253

Die perfekte positive Korrelation bei einer Verschiebung um den Wert $\tau = 0$ ist bei der Autokorrelation trivial, da es sich ja um einen direkten Vergleich der Variablen mit sich selbst handelt. Bei Lag = 1 wird jedoch ersichtlich, dass sich die Korrelation ändert (auf $r = -0.66$), springt dann wieder auf $r = +0.37$ usw.

Es ist zu beachten, dass dieser Datensatz nur zu Demonstrationszwecken erzeugt wurde. Eine inhaltliche Interpretation wäre im gegebenen Fall nicht angebracht.

Nichts desto trotz sollte durch diese Beispiel gezeigt werden, dass sowohl die Kreuzkorrelation als auch die Autokorrelation vor allem in der Zeitreihenanalyse (und damit auch bei Längsschnittstudien) wichtige Erkenntnisse über die betrachteten Variablen liefern können. Vor allem kann eine vorliegende **Autokorrelation** bei der MLR zu beträchtlichen Einschränkungen der Gültigkeit einen Modells beitragen. Bei den MLR-Methoden werden wir noch über Möglichkeiten sprechen, Autokorrelationen auf statitische Signifikanz zu prüfen (Stichwort: Durbin-Watson).

Korrelationen

Korrelationen sind ein Maß für den statistischen Zusammenhang zweier Datenreihen. Ein Korrelationsmaß impliziert daher auch stochastische Abhängigkeit - ohne jedoch auf kausale Zusammenhänge schließen zu können.

Korrelationen werden i.A. der *deskriptiven Statistik* zugeordnet. Durch eine Reihe von Verfahren, wie z.B. partielle Korrelation, multiple Korrelation oder Faktorenanalyse, kann die einfache Korrelation zweier Variablen auf Beziehungen zwischen zwei Variablen unter Berücksichtigung des Einflusses weiterer Variablen werden.

Korrelationen sind ein unverzichtbares Werkzeug für viele Forschungsgebiete und stehen häufig am Beginn jeder weiteren Datenanalyse, wie z.B.:

- multiple Regression
- Faktorenanalyse
- Clusteranalyse
- Mediator- und Moderator-Analyse

Pearson Produkt Moment Korrelation

Die häufigst verwendete Form der Korrelationsberechnung ist die Pearson-Produkt-Moment Korrelation. Bei dieser Methode wird die Beziehung zwischen zwei metrische Variablen (bzw. eine metrische und eine dichotome Variable) als Kennzahl mit dem Wertebereich $r \in [-1, 1]$ berechnet.

Die Berechnung einer Korrelation ist für sich gesehen an keine Voraussetzungen gebunden. Hingegen fordern eine sinnvolle Interpretationen der berechneten Kennwerte und vor allem die statistischen Tests von Korrelationskoeffizienten folgende inhaltliche und formale Überlegungen:

- **Skalenniveau**: der Korrelationskoeffizient liefert sinnvoll interpretierbare Ergebnisse wenn die Variablen mindestens intervallskaliert sind (oder für eine intervallskalierte und eine dichotome Variable[1]).
- **Endliche Varianz (und Kovarianz)**: bei Erhöhung des Stichprobenumfangs darf sich die Variabilität nicht immer weiter erhöhen, sondern sollte sich stabilisieren. Bei Variablen, die bivariat normalverteilt sind, ist diese Voraussetzung automatisch gegeben. Der Korrelationskoeffizient ist damit auch gleichzeitig der *Maximum-Likelyhood Schäzter* des Korrelationskoeffizienten in der Grundgesamtheit (asymptotisch *erwartungstreu* und *effizient*).

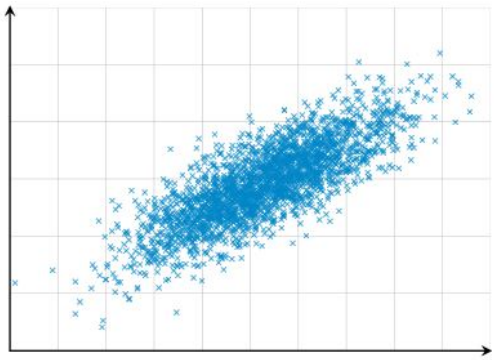


Figure 5: **Abbildung 5**: Endliche Varianz[2]

- **Linearität**: die Korrelation ist ein Maß für **lineare Abhängigkeit**. Abweichungen der Daten von dieser Linearitätsannahme führen zu einer mehr oder weniger starken Verzerrung des Korrelationskoeffizienten, wie in den nachfolgenden Beispielen gezeigt wird:

[1] dieser Spezialfall ist unter biserialer, bzw. punktbiserialer Korrelation bekannt.
[2] die Abbildungen wurden der Website Matheguru entnommen

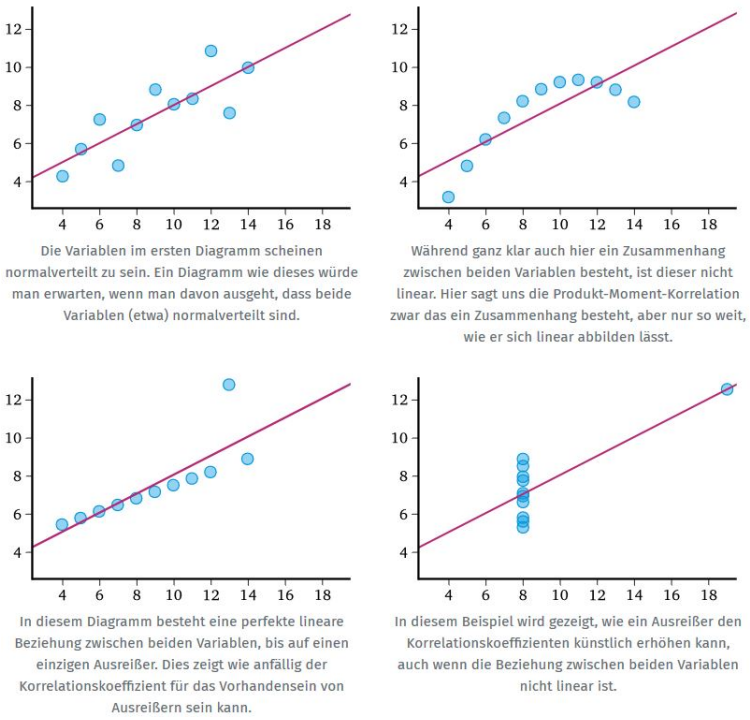


Figure 6: **Abbildung 6**: Linearität und Korrelation

Vor allem zur Prüfung der Signifikanz einer Korrelation sollem man weitere Voraussetzungen überprüfen:

- **Normalverteilung**: Korrelation berechnen sich aus dem Kreuzprodukt von z-standardisierten Werten zweier Variablen. Für diese Berechnung wird der Mittelwert als zentraler Kennwert verwendet, welcher nur dann ein "sinnvoller" Kennwert für die Daten ist, wenn diese zumindest symmetrisch und im besten Fall normalverteilt sind.
- **Homoskedastizität**: bedeutet gleichmäßige Streuung der Daten zweier (exogene und endogene) Variablen. Sind die exogene und die endogene Variable[3] nicht mehr identisch verteilt, d.h. sie ändern ihre Variablität mit zu/abnehmenden Werten einer Variablen, spricht man von **Herteroskedastizität**. Das hat zur Folge hat, dass die KQ[4]-Schätzer nicht mehr effizient sind und der Standardfehler der Koeffizienten verzerrt und nicht konsistent wird.
- **KEINE Ausreißer**: der Korrelationskoeffizient ist nicht robust gegenüber Ausreißern. Dies bedeutet, dass Ausreißer den Korrelationskoeffizienten sowohl künstlich erhöhen als auch künstlich senken können.
- **KEINE Kluster**: es kann vorkommen, dass zwei oder mehr Gruppen eine Korrelation zeigen, die eigentlich getrennt untersucht werden müssten. Dieses Problem wird oft auch mittels **partieller Korrelation** umgangen, bei der mögliche Drittvariablen statistisch konstant gehalten werden.

Beispiel Pearson Korrelation

Im folgenden, fiktiven Beispiel werden die Zusammenhänge von Klausurperformanz (EP), Intelligenz (IQ), Vorbereitungszeit (VZ) und Prüfungsangst (PA) korreliert. Der Code zum Laden der Daten sowie die Daten selbst sind in nachfolgender Ausgabe/Tabelle dargestellt:

[3]eine *exogene* Variable ist eine erklärende Variable, die mit der Störgröße unkorreliert ist (sogenannte Exogenität). Eine *endogene* Variable in einem multiplen Regressionsmodell ist eine erklärende Variable, die entweder aufgrund einer ausgelassenen Variablen, eines Messfehlers oder wegen Simulatanität mit der Störgröße korreliert ist (sogenannte Endogenität).

[4]KQ steht für kleinste Quadrate (auch MLS - Minimum Least Square, oder OLS - Ordinary Least Square) und ist eine einfache Schätzung über minimierte quadradische Abstände der Residuen (Fehler) zu einem Modell (Mittelwert, Gerade, etc.)

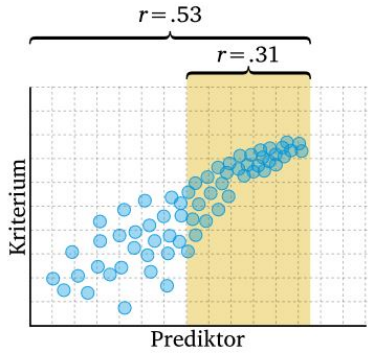


Figure 7: **Abbildung 7**: Variablität(einschränkung) und Korrelation

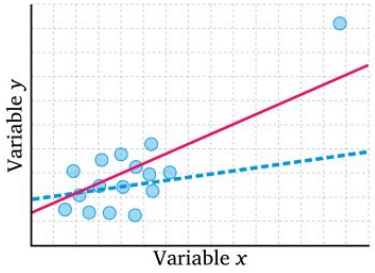


Figure 8: **Abbildung 8**: Einfluss von Ausreißer bei linearer Modellbildung

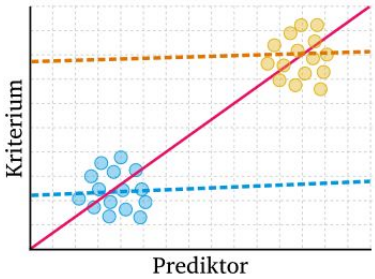


Figure 9: **Abbildung 9**: Kluster und deren Auswirkung bei linearer Modellierung

```r
load("Daten/CorrBsp1.Rda")
```

EP	IQ	VZ	PA
74	109	16	117
67	96	18	122
72	106	13	108
66	89	12	97
63	93	14	98
67	102	15	106

Aufgabe Korrelation

Kopiere den obigen Code zum Laden der Daten in eine R-Script-Datei. Führe nun folgende Aufgaben aus:

1. Ermittle mit einer geeigneten Funktion die Korrelationen und prüfe diese auch auf statistische Signifikanz.
2. Zeichne einen Korrelationsplot mit dem Paket *corrplot*.
3. Berechne die Teststärke der Korrelation $r(IQ, EP)$ (*Hinweis*: verwende die Funktion *pwr.r.test* des Pakets *pwr*).
4. Verwende diese Funktion (*pwr.r.test*) um für eine Korrelation $r(x, y) = 0.21$ den optimalen Stichprobenumfang zu berechnen.
5. Prüfe mit Hilfe der Funktion *mvn* aus dem Paket *MVN* die Voraussetzung der bivariaten Normalverteilung der Variablenpaare (EP,IQ), (EP, VZ) und (EP,PA).
6. Berechne die durchschnittliche Korrelation von $r_1(EP, IQ)$, $r_1(EP, VZ)$ und $r_1(EP, PA)$. Beachte, dass zur Berechnung von durchschnittlichen Korrelationswerten eine Fisher-Z-Transformation notwendig ist (*Hinweis*: verwende die *fisherz()* und *fisherz2r()* des Pakets *psych*).
7. Prüfe, ob der Unterschied der Korrelationskoeffizienten $r(EP, IQ) = 0.47$ und $r(EP, VZ) = 0.36$ statistisch signifikant ist. Verwende die Funktion *paired.r()* aus dem Paket *psych*.

Lösung Aufgabe

Kausalität

Eine relevante (statistisch signifikante) Korrelation liefert keinen Beleg für die Kausalität. Vor allem in der Medizin und Psychologie suchen Forscher nach Kriterien für Kausalität. Es existieren mehrere Ansätze zur Erklärung der Ursächlichkeit einer Korrelation (siehe z.B. die 9 Bradford-Hill-Kriterien).

Partial- Semipartialkorrelation

Die *partielle Korrelation* ist die bivariate Korrelation zweier Variablen, welche mittels linearer Regression vom Einfluss einer Drittvariablen bereinigt wurden.

Eine *Semipartialkorrelation* ist ein Zusammenhang zwischen einer residualisierten und einer nicht-residualisierten Variable.

Beispiel Partial- Semipartialkorrelation

Folgendes Beispiel verdeutlicht die Wirkungsweise einer Partial- und Semipartialkorrelation. Kopier den folgenden Code in ein R-Script und führe diesen dann aus. Diskutiere die Ergebnisse.

```r
examData  <- read.delim("Daten/Exam Anxiety.dat", header = TRUE)
examData2 <- examData[, c("Exam", "Anxiety", "Revise")]

# Normale Korrelation
  pander::pander(round(cor(examData2),2))
# Partielle Korrelation
```

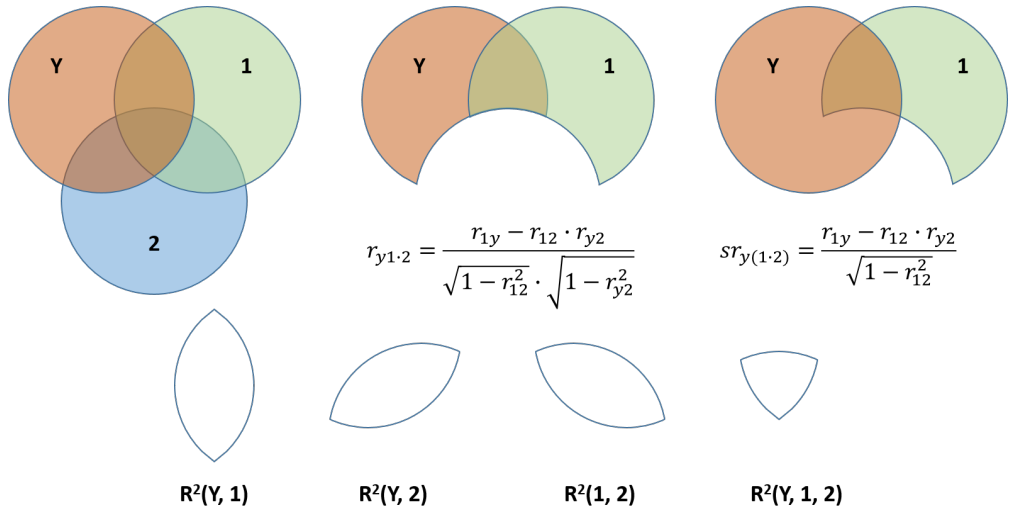


Figure 10: **Abbildung 10**: Partial und Semipartialkorrelation in einem Venn-Diagramm dargestellt

```
  # library(ppcor)
  pander::pander(round(ppcor::pcor(examData2)$estimate,2))
# Partialkorrelation mit Linearen Modell
  Mod1            <- lm(Exam ~ Revise, data = examData2)
  Res_Exam_Rev    <- residuals(Mod1)
  Mod2            <- lm(Anxiety ~ Revise, data = examData2)
  Res_Anx_Rev     <- residuals(Mod2)
  pr_Exam_Anx_Rev <- round(cor(Res_Exam_Rev, Res_Anx_Rev), 2)
```

In diesem Code wurde zur Veranschaulichung der Wirkungsweise einer Partial/Semipartialkorrelation eine lineare Regression verwendet. Was dabei genau passiert sei durch nachfolgende Abbildung nochmals veranschaulicht:

1. Examperformance wird durch Revisiontime vorhergesagt. Die Residuen sind jener Anteil an Variabilität der Examperformanz, der nicht durch Revsiontime vorhergesagt werden können[5]. Diese über die durch Revisiontime erklärbare Variabilität von Examperformance kann zurückgeführt werden auf:
   - andere erklärende Merkmale, bzw.
   - Messfehler
2. Anxiety wird durch Revisiontime vorhergesagt. Auch hier gilt wieder, dass die Residuen der Variabilität von Anxiete, bereinigt von Revisiontime entsprechen.
3. Die Korrelation der Residuen entspricht nun genau der Partialkorrelation $r_{Y1\cdot2}$

Bei der Semipartialkorrelation bereinigt man nun nicht beide Variablen, sonder eben nur einen Teil (z.B. wird nur die Anxiety von Revisiontime bereiningt).

Kopiere den nachfolgenden Code in ein R-Script und führe diesen aus. Diskutiere die Ergebnisse!

```
# Semipartielle Korrelation
  pander::pander(round(ppcor::spcor(examData2)$estimate,2))
# Semipartialkorrelation mit Linearen Modell
  sr_Exam_Anx_Rev <- round(cor(examData2$Exam, Res_Anx_Rev), 2)
```

[5] anderenfalls würden ja alle beobachteten Werte auf der Gerade liegen!

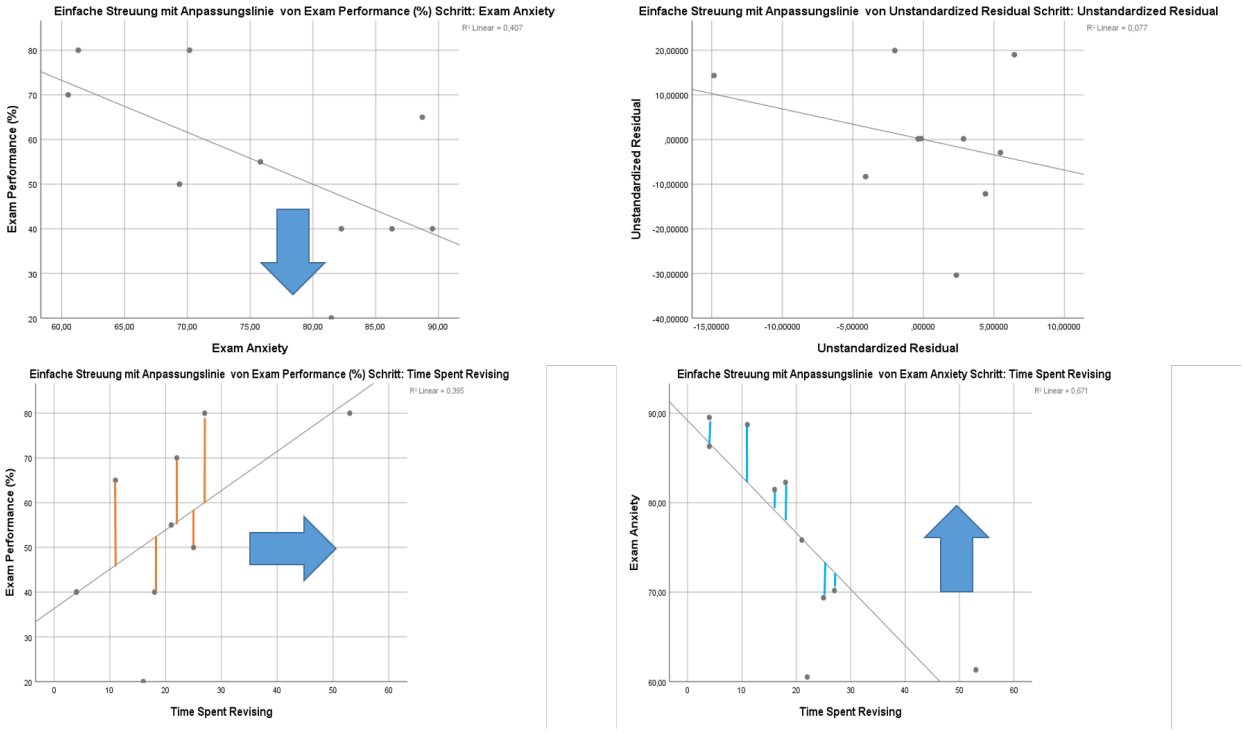


Figure 11: **Abbildung 11**: Partialkorrelation als lineares Regressionsmodell. Für die Beziehung Examperformance und Exam Anxiety soll der Effekt von Revisiotime berücksichtigt werden. Die roten Linien entsprechen den Residuen der Regression Revisiontime mit Exam Performance. Die blauen den Residuen der Regression Revisiontime mit Exam Anxiety. Der linke obere Graph stellt die Beziehung von Anxiety und Examperfomance bereinigt von Revisiontime dar. Details siehe nachfolgendemn Text.

Korrelationstechniken

Neben dem Pearson-Produkt-Moment-Korrelationskoeffizienten r existieren noch etliche weitere Korrelationskoeffizienten und Zusammenhangsmaße. Die meisten hiervon sind Sonderfälle der Pearson-Produkt-Moment-Korrelation. Nachfolgende Tabelle zeigt, wann welcher Koeffizient berechnet werden soll. Die Verwendung unterschiedlicher Korrelationsberechnungen ist i.A. abhängig vom Skalenniveau der beteiligten Variablen.

			Nomialskaliert		
			dichotom		
	Intervallskaliert	Ordinalskaliert	künstlich	natürlich	polytom
Intervallskaliert	■ Pearson Produkt-Moment-Korrelation	■ Spearman's Rho ■ Kendall's Tau ■ polychorische Korrelation	■ punktbiseriale Korrelation ■ biseriale Korrelation	■ punktbiseriale Korrelation	■ η-Koeffizient
Ordinalskaliert		■ Spearman's Rho ■ Kendall's Tau ■ polychorische Korrelation	■ biseriale Rangkorrelation ■ polychorische Korrelation	■ biseriale Rangkorrelation	■ Cramér's V
Nomialskaliert (künstlich dichotom)			■ Punkttetrachorische Korrelation (φ-Koeffizient) ■ Tetrachorische Korrelation	■ Punkttetrachorische Korrelation (φ-Koeffizient) ■ ν-Koeffizient	■ Cramér's V
Nomialskaliert (natürlich dichotom)				■ Punkttetrachorische Korrelation (φ-Koeffizient) ■ Yule's Y	■ Cramér's V
Nomialskaliert (polytom)					■ Cramér's V

Figure 12: **Abbildung 12**: verschiedene Korrelationskoeffizienten

Spearman und Kendall

Für die Berechnung des Pearson-Korrelationskoeffizienten (r) ist das Vorliegen von kontinuierlichen Variablen erforderlich. Bei **ordinalskalierten Daten** wird eine der folgenden Rangkorrelation berechnet:

1. **Spearman** r_s: Spearman-Rangkorrelation setzt voraus, dass Ränge gleichabständig sind[6] und keine Ausreißer vorliegen.
2. **Kendall** τ: Ränge müssen nicht gleichabständig sein und Ausreißer beeinflussen diesen Korrelationskoeffizienten weit weniger als z.B. den den r_s.

Bei den Kendall-Koeffizienten unterscheidet man noch drei unterschiedliche Maße[7]:

[6] diese Voraussetzung ist eher selten erfüllt. Sie ist gleichzusetzen mit der Annahme, dass in einem Skirennen der erste, zweite, dritte, etc. Platz genaus die gleichen Zeitabstände aufweisen. Ist diese nicht gegeben, sollte Kendalls τ verwendet werden.

[7] Details zu den unterschiedlichen Kendalls-τ sind der Literatur zu entnehmen. Weitere Betrachtungen beziehen sich auf das Kendalls-τ_b

- Kendalls τ_a: Rangbindungen werden nicht berücksichtigt.
- Kendalls τ_b: Rangbindungen werden berücksichtigt.
- Kendalls τ_c: für nicht quadratische Kontingeztafeln.

Zur Veranschaulichung der verschiedenen rangbasierten Korrelationsmaße sind folgende Aufgaben zu bearbeiten:

1. Berechne zuerst nochmal die Pearson-Korrelation $r(EP, IQ)$ des bereits geladenen Datensatzes und rechne dann eine Spearman Korrelation. Verwende nun die Funktion *cor()* des Basispakets. Vergleiche die Ergebnisse!
2. Vererwende die Funktion *rank()* um den Variablen *EP* und *IQ* Ränge zuzuordnen. Speichere die Ergebnisse in *EP_Ranks* und *IQ_Ranks* und berechnen Sie anschließend eine Pearson-Korrelation. Vergleiche die Ergebnisse mit dem vorherigen Pearson-r.

Lösung

Biseriale Korrelation

Biseriale Korrelationen kommen zur Anwendung, wenn ein Merkmal **Intervall-** oder **Ordinalskaliert** und das zweite Merkmal **dichotom Nominalskaliert** ist. Für das Nominalskalierte Merkmal unterscheidet man noch zwischen:

- **Echt dichotome Variable**: natürlich vorkommende Gruppenteilung wie z.B. wahr/falsch, männlich/weiblich, etc. Der Zusammenhang einer solchen mit einer intervallskalierten Variablen wird durch die **punktbiseriale Korrelation** beschrieben.
- **Künstlich dichotome Variable**: wird eine kontinuierliche Variable in zwei Gruppen aufgeteilt, wie z.B. zwei Altersgruppen (jung, alt), oder hohe Leistungsfähigkeit vs. niedrige Leistungsfähigkeit, etc., dann spricht man von einer künstlich dichotomen Variablen. Zusammenhänge dieser mit einer intervallskalierten Variablen werden durch die **biseriale Korrelation** beschrieben.

Kopier den folgenden Code in dein R-Script und bearbeite folgende Aufgabenstellungen:

1. Verwende die Funktion *dicho()* des Pakets sjmisc um alle Variablen über den Median zu dichotomisieren (*Hinweise*: ersetze die *XXX* im Code mit den entsprechenden Werten).
2. Berechne die biseriale Korrelation der Variablen *IQ* und der Exam-Performance-Gruppe (*EP_Grp*).

```r
# Zuerst wird die Examensperformanz über den Median in zwei Gruppen geteilt
# library(sjmisc)
DF_Biserial <- DF_Korr[order(EP), ]
DF_Biserial <- sjmisc::dicho(XXX,
                             dich.by = "XXX",
                             as.num = FALSE,
                             var.label = "Grp",
                             val.labels = c("low", "high"),
                             append = TRUE,
                             suffix = "_Grp")

biserial(x = XXX, y = XXX)
```

Lösung

Phi-Koeffizient

Korrelationen zwischen echt-dichotomen Variablen (männlich/weiblich, etc.) können mit dem Phi-Koeffizienten berechnet werden. Um den Phi-Koeffizienten zu berechnen, werden Häufigkeiten in Form einer Vier-Felder-Tafel benötigt.

Folgendes einfaches Beispiel zeigt die Berechnung des Phi-Koeffizienten sowie dessen Äquivalenz mit einer Pearson-Korrelation:

	Nicht bestanden	Bestanden	
Männlich	2 (a)	1 (b)	3 (a+b)
Weiblich	1 (c)	5 (d)	6 (c+d)
	3 (a+c)	6 (b+d)	9 (a+b+c+d)

$$\phi = \frac{(a \cdot d) - (b \cdot c)}{\sqrt{(a+b) \cdot (c+d) \cdot (a+c) \cdot (b+d)}}$$

$$\chi^2 = \frac{n \cdot [(a \cdot d) - (b \cdot c)]^2}{(a+b) \cdot (c+d) \cdot (a+c) \cdot (b+d)}$$

Figure 13: **Abbildung 13**: Beispiel und Berechnung des Phi-Koeffizienten

```
Geschlecht <- c(1,  1,  0,  0,  1,  0,  1,  1,  1)
Bestanden  <- c(1,  1,  1,  0,  0,  0,  1,  1,  1)

VFT        <- table(Geschlecht, Bestanden)
pander::pander(VFT)

cor(Geschlecht, Bestanden)

phi(VFT)
CST <- chisq.test(VFT, correct = FALSE)
pander::pander(CST)
qchisq(p=.95,df=1) # kritischer Chi-Square-Wert bei einem Freiheitsgrad und Alpha = 5%
```

Die Anzahl der Freiheitsgrade beträgt in diesem Fall immer eins, da wir es mit zwei dichotomen Merkmalen zu tun haben.

Tetrachorische Korrelation

Soll der Zusammenhang zwischen zwei künstlich-dichotomen Variablen berechnet werden, die aus stetigen, normalverteilten latente Variablen abgeleitet wurden (z.B. Intelligenz und Examperformanz in Statistik), verwendet man die tetrachorische Korrelation.

Auf Details zur Berechnung der Kenngröße wird hier verzichtet. Grundlage ist wiederum eine Vier-Felder-Tafel (wie beim Phi-Koeffizienten), wobei die tetrachorische Korrelation nicht so stark von der Randverteilung der Vier-Felder-Tafel abhängt. Zu beachten ist jedoch, dass die Zellbesetzung von b und c nicht 0 sein darf! Nachfolgendes Beispiel sollte die Anwendung der tetrachorischen Korrelation verdeutlichen. Lade dazu den folgenden Code und führe diesen Zeilenweise aus. Diskutiere die Ergebnisse!

```
# library (psych)
load("Daten/TetraCorrBsp.Rda")
TCC_Res <- tetrachoric(DF_TCC)
```

```r
pander(TCC_Res$rho)
```

Polychorische Korrelation

Um Korrelationen zwischen ordinalen Daten zu beschreiben, verwendet man die polychorische Korrelation. Dabei schätzt man die Korrelation zwischen zwei (ordinalen) Merkmalen, die in mehr als zwei geordnete Kategorien unterteilt sind. Die Berechnung ist überaus komplex und wird hier nicht dargestellt.

Polychorische Korrelationen werden unter anderem verwendet, um konfirmatorische Faktorenanalysen mit ordinalen Daten zu berechnen. Es ist mit Programmen wie R, AMOS, LISREL oder MPlus auch möglich, exploratorische Faktorenanalysen mit polychorischen Korrelationen durchzuführen.

Betrachte folgendes Beispiel:

```r
library(polycor)
library(mvtnorm)
set.seed(12345)
data <- rmvnorm(1000, c(0, 0), matrix(c(1, .5, .5, 1), 2, 2))
# Pearson correlation of those data points.
cor(data) # 0.5264
# And the Spearman correlation
cor(data, method="spearman") #0.5043
# Now let's try with polychoric correlation.
# First we need some cutoffs to break the data into cells ideally with breakpoints that avoid nearly
x <- cut(data[,1], c(-Inf, .75, Inf))
y <- cut(data[,2], c(-Inf, -1, .5, 1.5, Inf))
# Then take the polychoric correlation.
# The default method uses the faster 2-step method and only returns the correlation.
polychor(x, y)
# You don't need all the multivariate random normal overhead, polychor() works on any crosstab.
tab <- table(x,y)
polychor(tab)
# Let's try the same data with some different cutoffs and see how it does
x2 <- cut(data[,1], c(-Inf, -2, -1, 0, 1, 2, Inf))
y2 <- cut(data[,2], c(-Inf, -2, -1, 0, 1, 2, Inf))
table(x2,y2)
polychor(x2,y2,ML=TRUE,std.err=TRUE)
```

Lösungen

Aufgabe Korrelation Lsg

```r
# library(Hmisc) für Hmisc::rcorr
# library(corrplot) für corrplot
# library(pwr)

# 1. Ermitteln Sie mit einer geeigneten Funktion die Korrelationen und prüfen Sie
#    diese auch auf statistische Signifikanz.
  CorRes <- Hmisc::rcorr(as.matrix(DF_Korr), type="pearson") # type can be pearson or spearman
  pander::pander(round(CorRes$r, 2))
  pander::pander(round(CorRes$P , 2))
# 2. Zeichnen Sie einen Korrelationsplot mit dem Paket *corrplot*.
  corrplot::corrplot(cor(DF_Korr), type="upper", method = "ellipse")
# 3. Berechnen Sie die Teststärke der Korrelation $r(IQ, EP)$
```

```r
#    (Hinweis: verwenden Sie die Funktion *pwr::pwr.r.test* des Pakets *pwr*).
  N    <- dim(DF_Korr)[1]
  PwrA <- pwr::pwr.r.test(n = N, r = 0.47, sig.level = 0.05, alternative = 'two.sided')
  pander::pander(data.frame(Kennwerte = unlist(PwrA)))
# 4. Verwenden diese Funktion (*pwr::pwr.r.test*) um für eine Korrelation $r(x,y) = 0.21$
#    den optimalen Stichprobenumfang zu berechnen.
  OptN <- pwr::pwr.r.test(r = 0.21, sig.level = 0.05, power = 0.95, alternative = 'greater')
  pander::pander(data.frame(Kennwerte = unlist(OptN)))
# 5. Prüfen Sie mit Hilfe der Funktion *mvn* aus dem Paket *MVN*
#    die Voraussetzung der bivariaten Normalverteilung der
#    Variablenpaare (EP,IQ), (EP, VZ) und (EP,PA).
  # library(MVN)
  # mvn(DF_Korr[, c("EP","IQ")], multivariatePlot = "persp")
  # mvn(DF_Korr[, c("EP","VZ")], multivariatePlot = "persp")
  # mvn(DF_Korr[, c("EP","PA")], multivariatePlot = "persp")

# 6. Berechnen Sie die durchschnittliche Korrelation von $r_1(EP,IQ)$, $r_1(EP,VZ)$ und $r_1(EP,PA)$.
    round(fisherz2r(mean(c(fisherz(.47), fisherz(.36), fisherz(-.25)))), 2)
# 7. Prüfen Sie, on der Unterschied der Korrelationskoeffizienten $r(EP,IQ) = 0.47$ und $r(EP,VZ) = 0
#    statistisch signifikant ist. Verwenden Sie die Funktion *psych::paired.r()* aus dem Paket *psych
  # library(psych)
  psych::paired.r(xy        = .47,
                  xz        = .36,
                  n         = N,
                  twotailed = TRUE)
```

zurück zu Aufgabe

Aufgabe Spearman Lsg

```r
  # library(Hmisc)

  # 1. Berechnen Sie zuerst nochmal die Pearson-Korrelation $r(EQ,IQ)$ des bereits geladenen Datensat
  #    und rechnen Sie dann eine Spearman Korrelation. Verwenden Sie nun die Funktion cor() des Basis
  #    Vergleichen Sie die Ergebnisse!

  EP <- DF_Korr$EP
  IQ <- DF_Korr$IQ

  CorPearson  <- round(cor(EP, IQ, method = "pearson"), 2)
  CorSpearman <- round(cor(EP, IQ, method = "spearman"), 2)
  CorKendall  <- round(cor(EP, IQ, method = "kendall"), 2)

  pander::pander(data.frame(Pearson  = CorPearson,
                            Spearman = CorSpearman,
                            Kendall  = CorKendall))
  # Alternativ kann auch cor.test() verwendet werden, dabei werden die Tests auf
  # Signigikanz gleich mitgerechnet.
  # cor.test(x = EP,
  #          y = IQ,
  #          alternative = 'two.sided',
  #          method = 'pearson')
  # cor.test(x = EP,
```

```r
    #           y = IQ,
    #           alternative = 'two.sided',
    #           method = 'spearman')
    # cor.test(x = EP,
    #           y = IQ,
    #           alternative = 'two.sided',
    #           method = 'kendall')
    # Bemerkung: cor.test() mit Kendall bringt Warnung bezüglich der Rangbindungen.
    #            Alternativ kann man daher die Funktion Kendall() des Paketes Kendal verwenden:
      # library(Kendall)
        # Kendall(x = EP,
        #          y = IQ)
    # 2. Verwenden Sie die Funktion rank() um den Variablen EP und IQ Ränge zuzuordnen.
    #   Speichern Sie die Ergebnisse in EP_Ranks und IQ_Ranks und berechnen Sie anschließend
    #   eine Pearson-Korrelation. Vergleichen Sie die Ergebnisse mit dem vorherigen Pearson-r.
    EP_Ranks <- rank(EP, na.last = TRUE,
                     ties.method = c("average"))
    IQ_Ranks <- rank(IQ, na.last = TRUE,
                     ties.method = c("average"))
    round(cor(EP_Ranks, IQ_Ranks, method = "pearson"), 2)
```

zurück zu Aufgabe

Aufgabe Biserial Lsg

```r
    # Zuerst wird die Examensperformanz über den Median in zwei Gruppen geteilt
    # library(sjmisc)
    DF_Biserial <- DF_Korr[order(EP), ]
    DF_Biserial <- sjmisc::dicho(DF_Biserial,
                         dich.by = "median",
                         as.num = FALSE,
                         var.label = "Grp",
                         val.labels = c("low", "high"),
                         append = TRUE,
                         suffix = "_Grp")
    IQ          <- DF_Biserial$IQ

    biserial(x = IQ, y = DF_Biserial$EP_Grp)

# https://www.r-bloggers.com/2021/02/how-does-polychoric-correlation-work-aka-ordinal-to-ordinal-correl
```

zurück zu Aufgabe