

Weitere Korrelationskoeffizienten (1)

Welche Korrelationskoeffizienten man für die Analyse von Zusammenhängen benutzen kann, hängt entscheidend von dem **Skalenniveau** der beteiligten Variablen und der **Fragestellung** ab.

	intervall-skaliert	natürlich dichotom	künstlich dichotom	ordinal-skaliert
intervall-skaliert	P-M-K	P-M-K (punkt-biseriale Korrr.)	polyseriale Korrr. (biseriale Korrelation) P-M-K (punkt-biseriale Korrr.)	polyseriale Korrr. Kendalls τ
natürlich dichotom		P-M-K (φ -Koeffizient)	ν-Koeffizient P-M-K (φ -Koeffizient)	? biseriale Rang-Korr.
künstlich dichotom			polychorische Korrr. (tetrachorische Korrr.) P-M-K (φ -Koeffizient)	polychorische Korrr. biseriale Rang-Korr.
ordinal-skaliert				polychorische Korrr. Kendalls τ

Legende: Blau-Fett

zur Schätzung der P-M-K
der latenten Merkmale

Rot-Fett

zu Bestimmung der Korrelation
der manifesten Merkmale

(schwarz / blau / rot)

alternative Bezeichnung
bzw. Formel

Weitere Korrelationskoeffizienten (2)

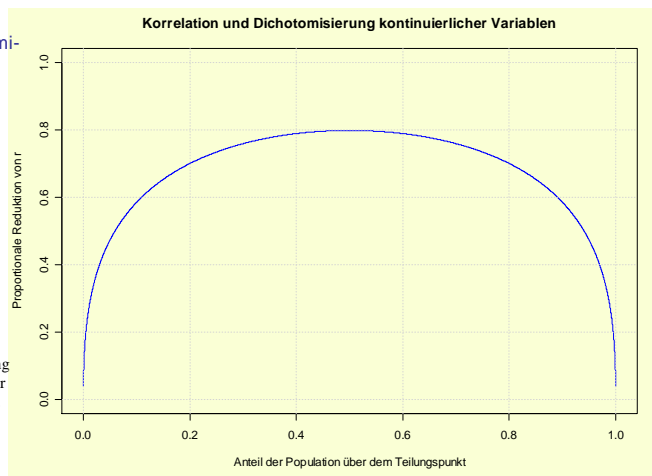
– Polyseriale und polychorische Korrelation (I) –

Werden kontinuierliche Variablen **dichotomisiert**, ergeben sich künstlich dichotome Variablen. Ihre Korrelation mit anderen Variablen wird dadurch in der Regel (aber nicht immer!) vermindert. Die Kategorisierung kontinuierlicher Variablen bedeutet immer Informationsverlust, der bei Dichotomisierung maximal wird. Dabei ergibt sich eine Verringerung der Korrelation auf *höchstens* 80%.

Effekt der Dichotomisierung **einer** von zwei normalverteilten Variablen:

$$\text{Reduktion} = \frac{h}{\sqrt{p \cdot (1-p)}}$$

mit h = Ordinate der Standardnormalverteilung und p = Anteil unter oder über dem Teilungspunkt



Weitere Korrelationskoeffizienten (3.1)

– Polyseriale und polychorische Korrelation (II) –

Da der Effekt der Kategorisierung auf die Größe der Produkt-Moment-Korrelation bei normalverteilten Variablen theoretisch bekannt ist, kann dieses Wissen benutzt werden, um zu **schätzen**, wie groß die Korrelation der Variablen wäre, **wenn** die zugrunde liegenden Variablen **normalverteilt** sind und kontinuierlich gemessen worden wären.

Beispiel P-M- (punkt-biseriale) und polyseriale (biseriale) Korrelation: Angenommen, eine kontinuierliche Variable Y wird im ersten Drittel (33.33% der niedrigsten Werte) geteilt, d.h. $100 \cdot p = 33.33\%$ der Stichprobe haben einen niedrigen und $100 \cdot q = 100 \cdot (p - 1) = 66.67\%$ der Stichprobe einen höheren Y -Wert. Die Produkt-Moment Korrelation zwischen einer kontinuierlichen Variablen X und der künstlich dichotomen Variablen Y sei $r = r_{\text{punkt-biseriale}} = 0.44$.

1) Bestimmung des z-Wertes für $p = 0.3333$:

$$z = \text{qnorm}(0.3333) = -0.431$$

2) Bestimmung der Ordinate h der Dichtekurve für $z = -0.431$ (s. nächste Seite):

$$h = \text{dnorm}(-0.431) = 0.364$$

3) Berechnung des Faktors, um den die P-M-Korrelation durch die Dichotomisierung reduziert wird:

$$\text{Red.Faktor} = \frac{h}{\sqrt{p \cdot (1-p)}}$$
$$\text{Red.Faktor} = \frac{0.364}{\sqrt{0.3333 \cdot (1-0.3333)}} = 0.772$$

4) Schätzung der P-M-Korrelation der latenten kontinuierlichen Variablen:

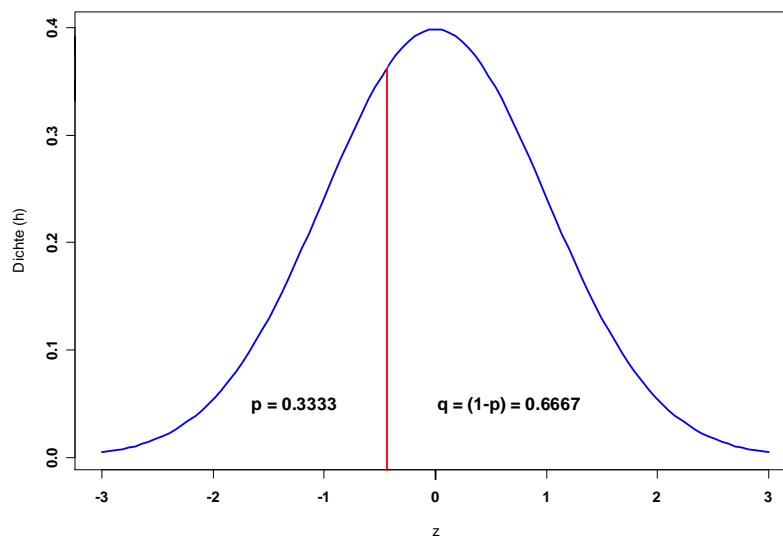
$$r_{\text{polyseriale}} = 0.44 / 0.772 = 0.57$$

Die polyseriale Korrelation der dichotomen Y -Variable mit der kontinuierlichen X -Variable ist $r_{\text{polyseriale}} = 0.57$. Dies wäre die P-M-Korrelation der X - mit der latenten Y -Variablen, **wenn** beide normalverteilt sind.

Weitere Korrelationskoeffizienten (3.2)

– Polyseriale und polychorische Korrelation (III) –

Standardnormalverteilung



Weitere Korrelationskoeffizienten (4)

– Polyseriale und polychorische Korrelation (IV) –

Die Reduktion der Korrelation auf 80% bei Dichotomisierung **einer** Variablen entspricht einer Verringerung der Stichprobengröße um 38%. Die Reduktion wird um so stärker, je weiter der Teilungspunkt vom Mittelwert bzw. Median entfernt ist. Eine Teilung bei 10 oder 90% der Stichprobe bewirkt eine Reduktion der Korrelation auf 58%; eine derartig dichotomisierte Variable ist nicht mehr normalverteilt, auch wenn die kontinuierliche Variable normalverteilt war.

Die Dichotomisierung von **zwei** Variablen reduziert die Korrelation noch stärker: Sie geht dann auf *höchstens* 64% zurück, was einer Verringerung der Stichprobengröße um 60% entspricht. Auch hier ist die Reduktion der Korrelation stärker, wenn die Teilungspunkte weiter vom Mittelwert bzw. Median entfernt liegen.

Unterstellt man hinter **einer künstlich dichotomen Variablen Y** eine normalverteilte kontinuierliche Variable, lässt sich die Produkt-Moment-Korrelation der latenten kontinuierlichen Y mit einer **kontinuierlichen** (mindestens intervallskalierten) normalverteilten **Variablen X** mittels der **polyserialen Korrelation** schätzen. Die polyseriale Korrelation lässt sich auch bei einer Teilung der Variablen in mehr als zwei Kategorien berechnen. Beim speziellen Fall der *Dichotomisierung* bezeichnet man die polyseriale Korrelation auch als **biseriale Korrelation**.

Im Fall von **zwei künstlich dichotomen Variablen** kann die Produkt-Moment-Korrelation der zugrunde liegenden (latenten) kontinuierlichen Variablen mittels der **polychorischen Korrelation** geschätzt werden, die auch bei einer Teilung in mehr als zwei Kategorien benutzt werden kann. Beim speziellen Fall der *Dichotomisierung* bezeichnet man die polychorische Korrelation auch als **tetrachorische Korrelation**.

Weitere Korrelationskoeffizienten (5)

– Polyseriale und polychorische Korrelation (V) –

Beispiel: Gewalteinrichtung (GE) und Selbstkontrolle (SK)

1. Beide kontinuierlich: P-M-Korr. $r = -.57$

2. SK Teilung ≤ 50.0

niedrig: $n = 96$, hoch: $n = 204$

P-M-Korr. (punkt-biseriale) $r = -.50$

polyseriale (biseriale) Korr. $r = -.65$

	SK ≤ 50	SK > 50	Σ
GE ≥ 50	36	13	49
GE < 50	60	191	251
Σ	96	204	300

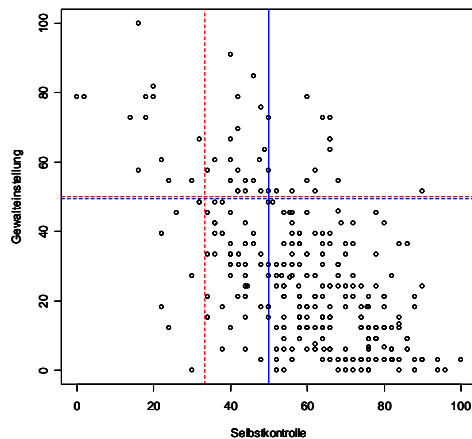
P-M-Korr (phi): $r = -.39$

polychorische (tetrachor.) Korr. $r = -.65$

	SK ≤ 33	SK > 33	Σ
GE ≥ 50	13	36	49
GE < 50	9	242	251
Σ	22	278	300

P-M-Korr (phi): $r = -.33$

polychorische (tetrachor.) Korr. $r = -.62$



Gewalteinrichtung: $\bar{x} = 27.6$, $\hat{\sigma} = 21.2$,
kurt = 0.18, skew = 0.86

Selbstkontrolle: $\bar{x} = 58.6$, $\hat{\sigma} = 18.0$,
kurt = -0.03, skew = -0.37

Weitere Korrelationskoeffizienten (6)

– Polyseriale und polychorische Korrelation (VI) –

Im obigen Beispiel sind die Produkt-Moment-Korrelationen ($r_{\text{punkt-biserial}}$ und φ) größer als erwartet, da die Variablen nicht wirklich normalverteilt sind – deshalb überschätzen in diesem Fall auch die polyseriale Korrelation und die polychorischen Korrelationen den Zusammenhang.

Die polyseriale und polychorische Korrelation eignen sich auch dann, wenn Variablen in **mehr als zwei Kategorien** geteilt werden. Es ist dabei (wie auch bei der Dichotomisierung) zu beachten, dass häufig die Variablen nicht erst nachträglich kategorisiert werden, sondern in den meisten Fällen schon in dichotomisierter oder anderweitig kategorisierter Form erfasst wurden. Hierbei lässt sich dann die Annahme der Normalverteilung der zugrunde liegenden latenten kontinuierlichen Variablen nicht mehr prüfen, sondern kann nur mehr oder weniger plausibel angenommen werden.

Eine besondere Form von kategorisiert erhobenen Daten stellen Variablen mit **Ordinalskalenniveau** dar. Hierfür ist die **polychorische Korrelation** dann häufig ebenfalls das angemessene Korrelationsmaß, zumindest dann, wenn durch die kategorisierte Form der Variablen ein eigentlich kontinuierliches Konstrukt gemessen werden sollte.

Bei ordinalskalierten Variablen stellt die Produkt-Moment-Korrelation auch dann kein angemessenes Korrelationsmaß dar, wenn man nur den Zusammenhang der *beobachteten* Werte beschreiben möchte. Hier müssen dann spezielle **Rangkorrelations-Koeffizienten** benutzt werden (s.u.).

Weitere Korrelationskoeffizienten (7)

– Rangkorrelation (I) –

Ein klassisches Maß der Korrelation **ordinalskalierter Variablen** ist **Spearman's ρ** (sprich: rho). Werden die Daten vor der Berechnung der Korrelation in Ränge transformiert, ist Spearman's ρ identisch mit der Produkt-Moment-Korrelation der rangtransformierten Werte.

Für eine Rangtransformation werden alle Werte einer Variablen aufsteigend sortiert und mit aufeinanderfolgenden ganzen Zahlen mit einem Rangplatz versehen. Haben allerdings mehrere Personen den gleichen Wert, bekommen sie den jeweils mittleren Rangwert zugewiesen.

Beispiel: Die sortierten Werte 5, 9, 9, 9, 12, 17, 17, 25 werden in die Ränge 1, (2+3+4)/3, 5, (6+7)/2, 8 und somit in die Rangwerte 1, 3, 3, 3, 5, 6.5, 6.5, 8 transformiert. Die Ränge 3 und 6.5 stellen dabei sogenannte **verbundene Ränge** (*ties*) dar.

Kendalls τ (sprich: tau) gilt als die bessere Wahl, vor allem weil es sich bei verbundenen Rängen, die häufig vorkommen, besser für Signifikanztests eignet. Es ist ein Maß für den monotonen Zusammenhang zwischen zwei Variablen. Im Fall unverbundener Ränge ist Kendalls τ die standardisierte Differenz der Anzahl der Rangplätze einer Variablen Y , die *kleiner* als die nachfolgenden (Proversionen P) und die *größer* als die nachfolgenden (Inversionen I) sind, wenn die Fälle anhand der Werte einer Variablen X aufsteigend sortiert werden:

$$\tau = \frac{P - I}{n \cdot (n - 1) / 2} = \frac{S}{S_{\max}}$$

mit P = Proversionen, I = Inversionen, n = Fallzahl
und S_{\max} = Anzahl aller Rangplatzvergleiche

Werden die Werte anhand der Y -Variablen (aufsteigend) sortiert, werden die Rangplätze der X -Variablen paarweise verglichen, um P und I zu bestimmen (das Ergebnis ist gleich).

Bei gebunden Rängen funktioniert diese Methode nicht und die Formel für τ ist dann anders.

Weitere Korrelationskoeffizienten (8)

– Rangkorrelation (II) –

Beispiel: Für sieben Personen liegen die folgenden *ordinal skalierten* Messwerte der Variablen X und Y vor:

	1	2	3	4	5	6	7
X	10	14	2	7	19	15	9
Y	11	17	5	1	18	16	13

Die (**fälschlicherweise!**) berechnete **Produkt-Moment Korrelation** der Werte ist **$r = 0.849$** .

Nach **Rangtransformation** beider Variablen und aufsteigender Sortierung aller Fälle anhand der X-Variablen stellen sich die Werte wie folgt dar:

	3	4	7	1	2	6	5
X	1	2	3	4	5	6	7
Y	2	1	4	3	6	5	7

Die Produkt-Moment-Korrelation der rangtransformierten Werte ist **Spearman's $\rho = 0.893$** .

Um Kendalls τ zu berechnen, werden zunächst P und I bestimmt. Hierbei werden die Rangplätze aller Y-Werte mit den folgenden Rangplätzen verglichen. Insgesamt sind dabei $P = 18$ Werte kleiner als die jeweils vorhergehenden (2, 4), (2, 3) (2, 6), (2, 5), (2, 7); (1, 4), (1, 3), (1, 6), (1, 5), (1, 7); (4, 6), (4, 5), (4, 7); (3, 6), (3, 5), (3, 7); (6, 7); (5, 7) und $I = 3$ Werte größer als der jeweils vorhergehende (2, 1), (4, 3), (6, 5) (dies Verfahren gilt nicht bei verbundenen Rängen!).

Mit $(18 - 3) / (7 \cdot 6 \cdot 0.5)$ ergibt sich die Rangkorrelation von X und Y: **Kendalls $\tau = 0.714$** .

Die **polychorische Korrelation** von X und Y ist dem gegenüber **$r_{\text{polychorisch}} = 0.912$** .

Weitere Korrelationskoeffizienten (9)

– Der v-Koeffizient –

Ist eine der Variablen **natürlich** dichotom und die andere **künstlich dichotom**, kann die Produkt-Moment-Korrelation (hier häufig als φ -Koeffizient bezeichnet) berechnet werden, sofern es darum geht, den Zusammenhang der beobachteten Variablen zu ermitteln. Lautet jedoch die Frage, wie sich zwei Gruppen hinsichtlich einer eigentlich als kontinuierlich gedachten Variable unterscheiden, sollte analog zur polyserialen Korrelation ein Koeffizient benutzt werden, dessen Größe nicht von der Verteilung der Fälle auf die beiden Gruppen und dem Teilungspunkt der latenten kontinuierlichen (als normalverteilt unterstellten) Variablen abhängt.

In diesem Fall, der in kriminologischen Fragestellungen häufig auftritt (z.B. die Ausprägung eines Risikofaktors nach Geschlecht), kann der **Koeffizient v** (sprich: nü) benutzt werden:

$$v = \frac{\Delta}{\sqrt{\Delta^2 + \frac{1}{p_{1\bullet} \cdot (1 - p_{1\bullet})}}} \quad \text{mit} \quad \Delta = \Phi^{-1} \left[\frac{p_{00}}{p_{0\bullet}} \right] - \Phi^{-1} \left[\frac{p_{10}}{p_{1\bullet}} \right]$$

Hierbei stellt Φ^{-1} den z-Wert für die jeweilige Wahrscheinlichkeit und $p_{0\bullet}$ und $p_{1\bullet}$ die relativen Häufigkeiten der beiden Kategorien des natürlichen Merkmals sowie p_{00} und p_{10} die relativen Häufigkeiten der ersten Kategorie der künstlich dichotomen Variable für die erste und zweite Kategorie des natürlichen Merkmals dar.

Beispiel:

	weiblich	männlich
Risiko niedrig	.15	.10
Risiko hoch	.25	.50
$p_{x\bullet}$.40	.60

$$\Delta = \text{qnorm}(.15 / .40) - \text{qnorm}(.10 / .60) = .649$$

$$v = \frac{.649}{\sqrt{.649^2 + \frac{1}{.60 \cdot (1 - .60)}}} = .303$$

Die entsprechende Produkt-Moment-Korrelation beträgt **$r = \varphi = .236$** .

Weitere Korrelationskoeffizienten (10)

– Probleme künstlich dichotomer / kategorialer Daten –

Häufig werden Daten, die eigentlich kontinuierlich sind, aus pragmatischen Gründen in kategorialer bzw. dichotomer Form **erhoben**. Die Analyse dichotomer Daten ist jedoch aufwendiger und weniger zuverlässig; Normalverteilungsannahmen, die Voraussetzung für angemessenere Koeffizienten wie polyseriale oder polychorische Korrelationen sind, sind dann nicht überprüfbar. Deshalb sollte soweit wie möglich das höchste Skalenniveau schon bei der Datenerhebung benutzt werden!

Es ist beliebt, kontinuierliche Daten **nachträglich** zu kategorisieren, weil kategoriale bzw. dichotome Daten Praktikern leichter zu vermitteln seien. Andere Argumente sind, dass die Daten nicht normalverteilt sind und durch Kategorisierung die Abweichungen von der Normalverteilung geringer würden – das ist eine irrige Vorstellung! Kategorisierung reduziert nicht nur die Chance, signifikante Ergebnisse zu finden, sondern kann die Ergebnisse auch deutlich verfälschen!

Daten, die theoretisch kontinuierlich sind, sollten niemals ohne Not kategorisiert erfasst werden! Es gibt auch nur selten gute Gründe dafür, Daten für Analysezwecke zu kategorisieren – mit der Kategorisierung geht immer ein Informationsverlust einher!

Es kann sinnvoll sein, Daten für **Präsentationszwecke** zu kategorisieren oder dichotomisieren. Die der Präsentation und Argumentation zugrunde liegenden **Analysen** sollten jedoch immer mit dem vollständigen Material durchgeführt werden.

Folgende Literatur ist hierzu zu empfehlen:

- Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement*, 7, 249-253.
- Farrington, D.P. & Loeber, R. (2000). Some benefits of dichotomization in psychiatric and criminological research. *Criminal Behaviour and Mental Health*, 10, 100-122.
- MacCallum, R.C., Zhang, S., Preacher, K.J. & Rucker, D.D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods*, 7, 19-40.