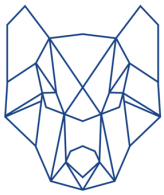
ST. PÖLTEN UNIVERSITY
OF APPLIED SCIENCES

Informatik
& Security



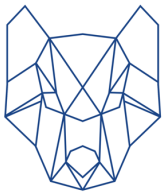
fh st.pölten

Korrelationsanalysen

Angewandte Statistik

Dipl.-Ing. Hubert Schölnast, BSc

Stand: 3. Juli 2023



Inhaltsverzeichnis

1 Pearson-Korrelationskoeffizient (r)

Der Korrelationskoeffizient nach Pearson ist eine Zahl zwischen -1 und 1, die eine Aussage darüber macht, wie stark die einzelnen Werte einer abhängigen Variable von den Werten der unabhängigen Variablen abhängen. Dabei wird ein linearer Zusammenhang angenommen. Die Annahme lautet also:

$$y_i = a \cdot x_i + b + e_i$$

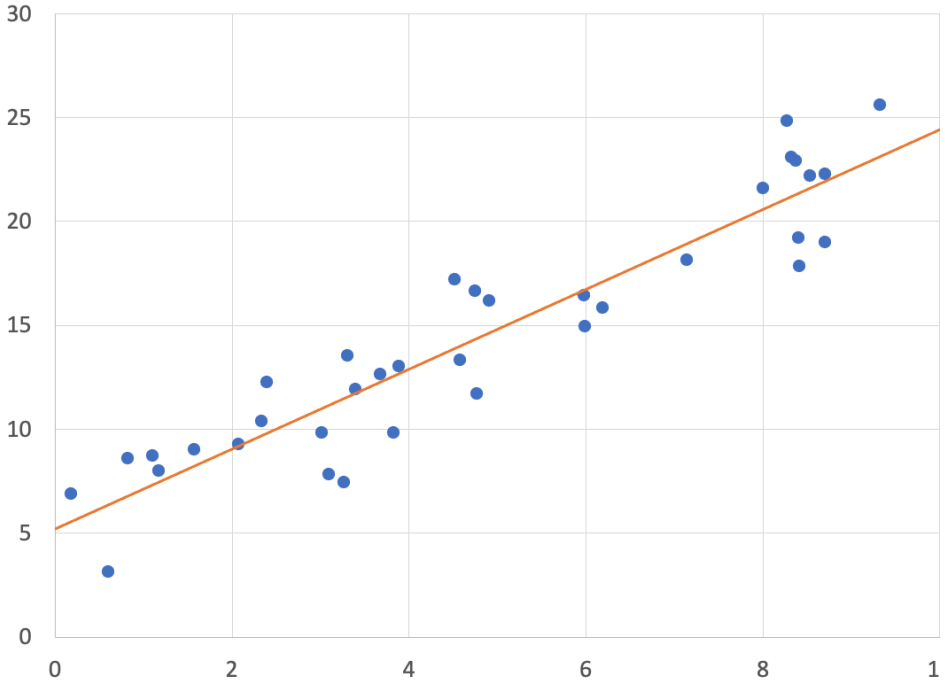
x_i Die einzelnen Werte der unabhängigen Variablen.

y_i Die einzelnen Werte der abhängigen Variablen.

a Die Steigung der Regressionsgeraden.

b Wenn $x = 0$ schneidet die Regressionsgerade die Y-Achse bei $y = b$

e_i Die Regressionsgerade sagt für die abhängige Variable den Wert $a \cdot x_i + b$ voraus. Tatsächlich hat diese Variable aber den Wert y_i. Die Differenz zwischen Vorhersage und tatsächlichem Wert ist der Fehlerwert e_i.



Dabei werden a und b so gewählt, dass $\sum e_i^2$ möglichst klein ist. Für die beiden Parameter a und b ergeben sich daraus diese Formeln:

$$a = \frac{Cov_{x,y}}{Var_x} \qquad b = \bar{y} - a \cdot \bar{x}$$

Die in diesen Formeln verwendeten Werte werden wie folgt berechnet:

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

$$Var_x = \sum_{i=1}^{n} (x_i - \bar{x})^2 \qquad Var_y = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

$$Cov_{x,y} = \sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y})$$

Unter diesen Bedingungen gibt dann der Korrelationskoeffizient r an, wie gut sich die Gesamtverteilung der Daten denn überhaupt durch eine Gerade beschreiben lässt.

1.1 Berechnung

Der Korrelationskoeffizient wird nach dieser Formel berechnet:

$$r = \frac{Cov_{x,y}}{\sigma_x \cdot \sigma_y}$$

mit

$$\sigma_x = \sqrt{Var_x} \qquad \sigma_y = \sqrt{Var_y}$$

Dabei ist σ_x die Standardabweichung x-Werte und σ_y ist die Standardabweichung der y-Werte. Über dem Bruchstrich steht die Kovarianz dieser beiden Zufallsvariablen.

1.2 Voraussetzung

mindestens Intervallskaliert

Die beiden Variablen, die in die Berechnung einfließen, müssen zumindest intervallskaliert sein. Die Methode ist ungeeignet für ordinal- und nominalskalierte Werte.

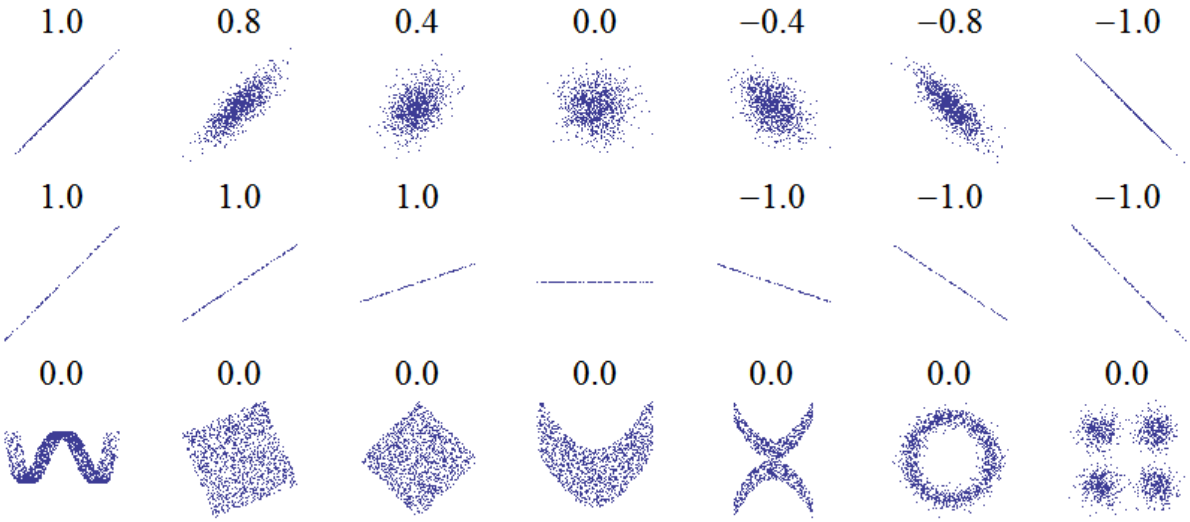
annähernd normalverteilt

Die Werte beider Variablen sollten idealerweise eine Verteilung aufweisen, die einer Normalverteilung entspricht. Wenn das nicht der Fall ist, liegt das oft an Ausreißern, die man vor der Auswertung eliminieren sollte.

Die Hypothese geht von einem linearen Zusammenhang aus

Die Pearson-Korrelation ist nur dann ein sinnvoller Wert, wenn der vermutete Zusammenhang tatsächlich linear ist. Bei anderen Zusammenhängen (exponentiell, quadratisch, logarithmisch usw.) erhält man falsche Werte (zu nahe bei 0).

1.3 Was sagt der Wert aus?



Wert nahe bei 0: Es liegt keine lineare Korrelation vor. Entweder gibt es überhaupt keinen statistischen Zusammenhang, oder einen Zusammenhang, der nicht linear ist.

Negative Werte: Bei Merkmalsträgern, bei denen die unabhängige Variable größere Werte hat, ist bei der abhängigen Variablen tendenziell eher mit kleinen Werten zu rechnen.

Positive Werte: Bei Merkmalsträgern, bei denen die unabhängige Variable größere Werte hat, ist bei der abhängigen Variablen tendenziell eher mit großen Werten zu rechnen.

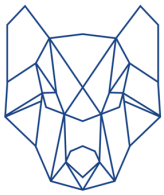
Der genaue Zahlenwert lässt sich besser interpretieren, wenn man daraus das Bestimmtheitsmaß berechnet.

2 Bestimmtheitsmaß (R^2)

Das Bestimmtheitsmaß ist das Quadrat des Pearsonschen Korrelationskoeffizienten. Daher gelten dieselben Vorbedingungen. Das Bestimmtheitsmaß ist eine Zahl zwischen 0 und 1.

2.1 Was sagt der Wert aus?

Bei der Interpretation des Bestimmtheitsmaßes wird davon ausgegangen, dass man die einzelnen Werte der abhängigen Variablen zu einem bestimmten Grad auf den Wert der unabhängigen Variablen zurückführen kann. Man nimmt also an, dass man mit dem Wert der unabhängigen Variablen teilweise erklären kann, warum die abhängige Variable ihren konkreten Wert hat. Die Größe dieses erklärbaren Anteils ist genau das Bestimmtheitsmaß.



Beispiel 1:

Bei tausenden Messtellen wurde die Temperatur in Fahrenheit und in Celsius gemessen. Es wurde dabei ein linearer Zusammenhang mit einem Bestimmtheitsmaß von $R^2 = 1{,}0.$ ermittelt. Das kann wie folgt interpretiert werden: Die Temperatur in Fahrenheit ist zu 100% auf die Temperatur in Celsius zurückzuführen. Es gibt keine anderen Einflüsse.

Beispiel 2:

In einer bestimmten Personengruppe hängt das Körpergewicht linear von der Körpergröße ab, dabei hat das Bestimmtheitsmaß R^2 den Wert $0{,}4$. Das kann wie folgt interpretiert werden: 40% des Körpergewichtes der Personen sind dadurch zu erklären, dass die Personen bestimmte Körpergrößen haben. Die restlichen 60% des Körpergewichtes sind auf andere Einflüsse als die Körpergröße zurückzuführen.

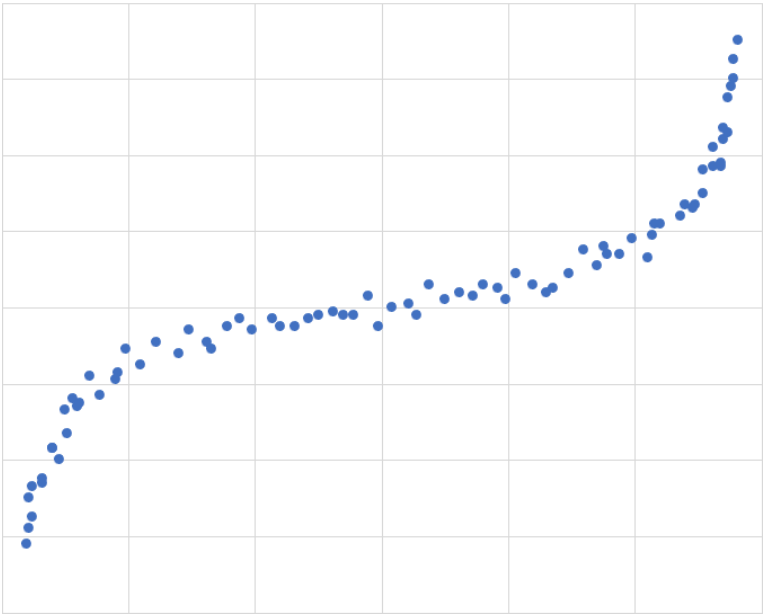
Beispiel 3:

Über viele Jahre hinweg wurde an vielen Sonntagen die Temperatur am Stephansplatz in Wien mit der Zusatzzahl der am selben Tag stattfindenden Lottoziehung verglichen. Es stellte sich heraus, dass das Bestimmtheitsmaß dieses Zusammenhangs R^2 einen Wert hat, der kleiner als $0{,}001$ ist. Das kann wie folgt interpretiert werden: Die gezogene Zusatzzahl scheint praktisch gar nicht durch die Lufttemperatur erklärbar zu sein. Der genaue Wert der Zusatzzahl ist zu praktisch 100% auf andere Einflüsse als die Lufttemperatur zurückzuführen.

3 Rangkorrelationskoeffizient

Der oben beschriebene Pearsonsche Korrelationskoeffizient setzt einen linearen Zusammenhang voraus. Die Art, wie er berechnet wird, führt auch dazu, dass der Pearsonsche Korrelationskoeffizient von Ausreißern stärker beeinflusst wird als von Werten in der Mitte der Verteilung.

Die grundlegende Idee von Rangkorrelationskoeffizienten besteht darin, die Auswertung nicht mit den eigentlichen Daten (Messwerten) zu machen, sondern mit dem Rang der Daten.



Für die Daten aus dem Bild wurde ein Pearsonscher Korrelationskoeffizient von 0,928 berechnet, aber ein Rangkorrelationskoeffizient (hier ist es der Spearmansche Rangkorrelationskoeffizient) ergibt hier 0,994.

Es gibt mehrere Arten, einen Rangkorrelationskoeffizienten zu berechnen. Hier ist der Spearmansche Rangkorrelationskoeffizient beschrieben:

3.1 Berechnung anhand eines Beispiels:

Sie haben von 11 Personen die Länge L (Körpergröße) und deren Gewicht G gemessen. und haben nun eine Tabelle mit 100 Zeilen und den beiden Spalten L und G.

Ermitteln Sie nun die Ränge für die Länge L. Eine Möglichkeit wäre, die Tabelle nach den Werten der Spalte L zu sortieren, und dann eine neue Spalte RL hinzuzufügen, in der von oben nach unten die Zahlen 1, 2, 3 usw. eingefügt werden. Das würde aber nur funktionieren, wenn in L lauter verschiedene Werte stünden. Besser ist es, stattdessen ohne zu sortieren die Spalte RL zu erzeugen und, wenn Sie mit Excel arbeiten, mit der Funktion `RANG.MITTELW(a,b,c)` die Ränge zu ermitteln. Dabei ist `a` ein Verweis auf eine Zelle in der Spalte mit den Längen und in derselben Zeile wie die Formel selbst. Der Wert `b` muss auf die ganze Spalte mit den Längen verweisen und `c` ist ein Wert, der angibt, ob die Ränge aufsteigend oder absteigend vergeben werden sollen. Die Wahl von `c` ist egal, solange Sie auch die Spalte RG, in der die Ränge der Gewichte stehen, mit demselben Wert für `c` erzeugen.

Wie schon angedeutet machen Sie nun dasselbe mit dem Gewicht. Die neue Spalte heißt RG.

Machen Sie nun eine weiter Spalte (die fünfte bisher) mit dem Namen *diff*. Darin berechnen Sie die Differenz der beiden Ränge des jeweiligen Merkmalsträgers (also die Differenz der Werte in den Spalten RL und RG). In einer sechsten Spalte, sie soll *diff2* heißen, berechnen Sie das Quadrat von *diff*.

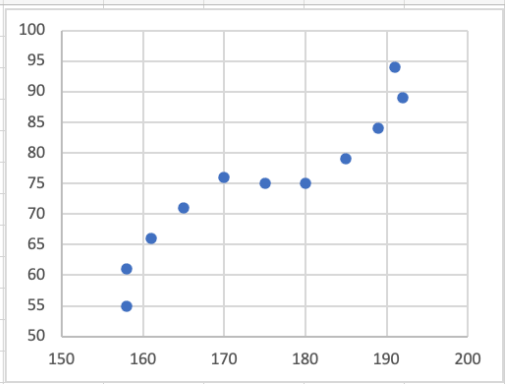
Addieren Sie nun alle Werte in der Spalte *diff2*, das ergibt die Summe S. Außerdem brauchen Sie noch die Anzahl der Merkmalsträger n. In unserem Beispiel ist $n = 11$.

Rechnen Sie nun:

$$\rho = 1 - \frac{6 \cdot S}{n \cdot (n^2 - 1)}$$

L	G
191	94
158	55
185	79
180	75
161	66
158	61
175	75
165	71
189	84
170	76
192	89

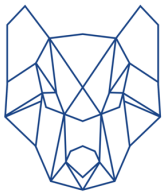
	A	B	C	D	E	F	G	H
1	L	G	RL	RG	diff	diff2	Summe	9
2	191	94	10	11	1	1	Anzahl	11
3	158	55	1,5	1	-0,5	0,25	rho	**0,95909**
4	185	79	8	8	0	0		
5	180	75	7	5,5	-1,5	2,25		
6	161	66	3	3	0	0		
7	158	61	1,5	2	0,5	0,25		
8	175	75	6	5,5	-0,5	0,25		
9	165	71	4	4	0	0		
10	189	84	9	9	0	0		
11	170	76	5	7	2	4		
12	192	89	11	10	-1	1		




```
Formel C2:  =RANG.MITTELW(A2;A$2:A$12;1)
Formel D2:  =RANG.MITTELW(B2;B$2:B$12;1)
Formel E2:  =D2-C2
Formel F2:  =E2^2
Formel H1:  =SUMME(F:F)
Formel H2:  =ANZAHL(F:F)
Formel H3:  =1-(6*H1)/(H2*(H2^2-1))
```

3.2 Was sagt der Wert aus?

Der Spearmansche Rangkorrelationskoeffizient ist gleich zu interpretieren wie der Pearson-Korrelationskoeffizient.



3.3 Voraussetzung

Der Spearmansche Rangkorrelationskoeffizient setzt keine lineare Korrelation aus, sondern nur eine monotone Korrelation.

Lineare Korrelation: Im Streudiagramm muss man eine Form erahnen können, die an eine gerade Linie erinnert. Das heißt: Wenn man eine glatte Linie zeichnet, die sich möglichst gut entlang der Datenpunkte entlangschlängelt, sollte diese Line überall dieselbe Steigung haben, also eine Gerade sein.

Monotone Korrelation: Es genügt, wenn die Steigung der Näherungskurve immer positiv oder immer negativ ist.

4 Kontingenztabelle = Kreuztabelle

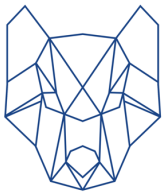
Das ist eine Tabelle, in der die absoluten oder relativen Häufigkeiten des gemeinsamen Auftretens zweier verschiedener Merkmale notiert sind. Ein Beispiel ist die Häufigkeit der Kombination von Augenfarbe und Haarfarbe. Kontingenztabellen sind für sich alleine schon nützlich, sie sind aber auch die Voraussetzung, um den Chi-Quadrat-Koeffizienten zu berechnen (siehe nächster Abschnitt).

	schwarz	braun	rot	blond
dunkelbraun	111	200	43	12
hellbraun	25	108	23	17
blau	33	140	28	157
grün	8	48	20	27

Diese Tabelle ergänzen Sie um Zeilen- und Spaltensummen

	schwarz	braun	rot	blond	→ Σ
dunkelbraun	111	200	43	12	366
hellbraun	25	108	23	17	173
blau	33	140	28	157	358
grün	8	48	20	27	103
↓ Σ	177	496	114	213	1000

Im Feld in der rechten unteren Ecke steht dann die Gesamtanzahl der Elemente. Die bisherige Tabelle enthält die absoluten Häufigkeiten. Teilen Sie nun alle Werte durch die Gesamtanzahl der Elemente (hier: 1000):



	schwarz	braun	rot	blond	→ Σ
dunkelbraun	0,111	0,200	0,043	0,012	0,366
hellbraun	0,025	0,108	0,023	0,017	0,173
blau	0,033	0,140	0,028	0,157	0,358
grün	0,008	0,048	0,020	0,027	0,103
↓ Σ	0,177	0,496	0,114	0,213	1

4.1 Berechnen von bedingten Häufigkeiten

Sie wollen wissen, wie groß der Anteil der rothaarigen Menschen unter den grünäugigen ist? Ganz einfach: Teilen Sie den Wert der rothaarigen grünäugigen Menschen (im Beispiel: 0,020) durch die Zeilensumme der grünäugigen Menschen (hier: 0,103). Das Ergebnis ist die gesuchte Zahl: $\frac{0{,}020}{0{,}103} = 0{,}194\cdots = 19{,}4\%$

4.2 Erkennen von Abhängigkeiten

Um zu erkennen, ob es einen Zusammenhang zwischen der Haarfarbe und der Augenfarbe gibt, multiplizieren Sie eine beliebige Zeilensumme (z.B. 0,358 für die Zeile mit den blauen Augen) mit einer beliebigen Spaltensumme (z.B. mit 0,177 für die Spalte mit den schwarzen Haaren). Das Ergebnis dieser Multiplikation ist in diesem Beispiel 0,063366. Das ist der Erwartungswert für die relative Häufigkeit, wenn die Haarfarbe von der Augenfarbe unabhängig wäre. Tatsächlich beträgt die relative Häufigkeit für blauäugige schwarzhaarige Menschen aber 0,033. Der tatsächliche Wert beträgt also nur ungefähr die Hälfte des berechneten Wertes. Und daraus kann man schließen, dass es eine Abhängigkeit zwischen Haarfarbe und Augenfarbe gibt. Nur wenn in allen Zellen die Produkte aus Zeilen- und Spaltensumme mit den eingetragenen Werten übereinstimmen, sind die beiden Variablen unabhängig.

Allerdings führen schon kleine zufällige Schwankungen bei der Stichprobenziehung zu kleinen Abweichungen, weswegen diese Methode noch verfeinert werden muss.

5 Chi-Quadrat-Koeffizient und normierter Kontingenzkoeffizient

5.1 Chi-Quadrat-Koeffizient

Der Chi-Quadrat-Koeffizient ist die Verfeinerung der soeben beschriebenen Methode. In 4.2 wurde erklärt, wie der Erwartungswert für die unabhängige Häufigkeit von blauäugigen schwarzhaarigen Menschen berechnet wird. Dafür haben wir die Zeilensumme und die Spaltensumme aus der Tabelle mit den relativen Häufigkeiten miteinander multipliziert. Das kann man natürlich für jede Zelle machen und erhält dann eine neue Tabelle mit allen Erwartungswerten für unabhängige Verteilungen:

	schwarz	braun	rot	blond	→ Σ
dunkelbraun	0,064782	0,181536	0,041724	0,077958	0,366
hellbraun	0,030621	0,085808	0,019722	0,036849	0,173
blau	0,063366	0,177568	0,040812	0,076254	0,358
grün	0,018231	0,051088	0,011742	0,021939	0,103
↓ Σ	0,177	0,496	0,114	0,213	1

Nun berechnet man für jede Zeile folgenden Ausdruck, der **Chi-Quadrat-Statistik** heißt:

$$X^2 = \frac{(echt - theor)^2}{theor}$$

Man berechnet also die Differenz aus dem echten Wert und dem theoretischen, quadriert diese Differenz und teils sie dann durch den theoretischen Wert. Für schwarzhaarige Menschen mit dunkelbraunen Augen ergibt das diese Chi-Quadrat-Statistik:

$$X^2_{schwarz,dunkelbraun} = \frac{(0{,}111 - 0{,}064782)^2}{0{,}064782} = 0{,}03267\cdots$$

Das macht man für alle Zellen und erhält in unserem Beispiel diese Werte (alle Zahlen auf 6 Nachkommastellen gerundet)

	schwarz	braun	rot	blond
dunkelbraun	0,032974	0,001878	0,000039	0,055805
hellbraun	0,001032	0,005739	0,000545	0,010692
blau	0,014552	0,007948	0,004022	0,085503
grün	0,005742	0,000187	0,005808	0,001167

Die Summe all dieser Chi-Quadrat-Statistiken ist der Chi-Quadrat-Koeffizient X^2. Er beträgt in unserem Beispiel $0{,}233632$. Hätten wir nicht mit den relativen Häufigkeiten gerechnet, sondern mit den absoluten, hätten wir $X^2 = 233{,}632$ erhalten.

Der Wertebereich liegt zwischen 0 und ∞. Ein Wert von genau 0 würde aussagen, dass die beiden Variablen Haarfarbe und Augenfarbe voneinander absolut unabhängig sind, aber das kommt bei Werten, die man aus Stichproben erhält, praktisch nie vor. Jeder andere Wert ist aber schwierig zu interpretieren, wenn der Wertebereich bis unendlich geht.

Abhilfe schafft der normierte Kontingenzkoeffizient:

5.2 Normierter Kontingenzkoeffizient

5.2.1 Kontingenzkoeffizient (noch nicht normiert):

Der Kontingenzkoeffizient C wird wie folgt berechnet:

$$C = \sqrt{\frac{X^2}{n + X^2}}$$

Dabei ist n die Summe aller Häufigkeiten. Wenn wir mit den absoluten Häufigkeiten gerechnet hätten, hätte X^2 den Wert $233{,}632$ und $n = 1000$. Wir haben aber mit den relativen Häufigkeiten gerechnet und haben $X^2 = 0{,}233632$ und $n = 1$. Das Ergebnis für C ist in beiden Fällen dasselbe: $C = 0{,}435184$. Aber auch dieser Wert kann noch nicht so einfach interpretiert werden. Er muss erst noch normiert werden:

5.2.2 Normierung:

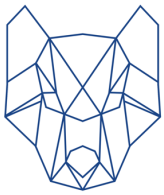
Der höchste Wert, den C annehmen kann, ist nicht 1, sondern dieser Wert:

$$C_{max} = \sqrt{\frac{m-1}{m}}$$

Dabei ist die ganze Zahl m das Minimum aus der Anzahl der Spalten und der Anzahl der Zeilen der Kontingenztabelle. Bei einer Tabelle mit 3 Zeilen und 50 Spalten wäre $m = 3$. Bei einer Tabelle mit 9 Zeilen und 7 Spalten wäre $m = 7$, also immer der kleinere der beiden Werte.

In unserem Beispiel haben wir 4 Zeilen und 4 Spalten, daher ist $m = 4$ und

$$C_{max} = \sqrt{\frac{4-1}{4}} = \sqrt{\frac{3}{4}} = \frac{\sqrt{3}}{2} = 0{,}8660\cdots$$



Wenn wir jetzt C durch C_{max} dividieren, erhalten wir endlich einen Wert, der garantiert zwischen 0 und 1 liegt. Das ist der normierte Kontingenzkoeffizient:

$$C_{norm} = \frac{C}{C_{max}}$$

In unserem Beispiel:

$$C_{norm} = \frac{0{,}435184}{0{,}8660\cdots} = 0{,}5025\cdots$$

Der Wert liegt also bei ca. 50%.

5.3 Interpretation

Wenn der normierte Kontingenzkoeffizient nahe bei 0 liegt, sind die verglichenen Variablen unabhängig. Ein Wert nahe bei 1 signalisiert eine sehr hohe Abhängigkeit. Ein Wert dazwischen ist schwierig zu interpretieren, ohne Aussagen über die Signifikanz der Abhängigkeit zu machen.