Kapitel 7: Zweidimensionale Häufigkeitsverteilungen

1. Regression und Korrelation

Gegenstand der in den Kapiteln 7 und 8 dargestellten Methoden ist der Zusammenhang von zwei (oder mehr) Merkmalen. Dabei sind im wesentlichen folgende Fragestellungen üblich:

- Wie lässt sich der Grad (die Intensität) des Zusammenhangs zwischen zwei (oder mehr) als wechselseitig abhängig (interdependent) angenommenen Merkmalen (bzw. Variablen) messen? Je nach Art der vorliegenden Skalen für X und Y spricht man von Korrelations-, Rangkorrelations-, Kontingenz- oder Assoziationsanalyse oder auch von "Korrelation" im allgemeinen Sinne.
- Besteht ein Zusammenhang zwischen zwei oder mehr Merkmalen dergestalt, dass es möglich ist Y aufgrund von X oder Y aufgrund X_1 und X_2 zu schätzen, d.h. eine Funktion für Y in Abhängigkeit anderer Variablen aus den Daten zu schätzen?

Es ist üblich, die erste Fragestellung als typisch für die Korrelationsanalyse anzusehen, während es bei der Regressionsanalyse um die zweite Fragestellung geht. Es gilt in erster Näherung:

Korrelation = Analyse der Stärke der Interdependenz (wechselseitigen Abhängigkeit) und **Regression** = Analyse der Art der Dependenz (Abhängigkeit einer Variablen von anderen Variablen).

Für eine erste Näherung mag diese Unterscheidung zwischen Regressions- und Korrelationsanalyse, zwei Methoden, die meist in einem Atemzug genannt werden, ausreichen. Bei genauerer Betrachtung (die zunächst noch zurückzustellen ist) zeigt sich, dass der Unterschied zwischen diesen beiden Methoden jedoch weniger in der Zielsetzung liegt, als in den Voraussetzungen, die hinsichtlich der Variablen getroffen werden. Die Unterscheidung zwischen Interdependenz und Dependenz ist im übrigen keineswegs klar. Sie ist deshalb nicht befriedigend. Es ist insbesondere ziemlich abwegig alle multivariaten Verfahren im Sinne dieser Unterscheidung klassifizieren zu wollen, was in manchen Lehrbüchern geschieht.

Die Bestimmung einer Funktion zur Beschreibung des Zusammenhangs zwischen Variablen, bzw. genauer, zur Schätzung einer Variablen Y, aufgrund ihrer Abhängigkeit von anderen Variablen ist Gegenstand der Regressionsanalyse. Eine "abhängige" Variable Y (auch Regressand genannt) wird durch eine oder mehrere "unabhängige" Variablen "erklärt". Einfache Regression bedeutet Y in Abhängigkeit einer unabhängigen Variable (eines Regressors) X zu schätzen. Wird Y durch eine Funktion mehrerer Regressoren X_1, X_2,...,X_p erklärt, so spricht man von multipler Regression. Hinter solchen Betrachtungen kann (muss aber nicht) eine Kausalvorstellung stehen.

Voraussetzung sowohl der Regressions- als auch Korrelationsanalyse ist die Beschreibung eines zweidimensionalen (bivariaten) Datensatzes. In diesem Kapitel soll deshalb zunächst gezeigt werden, wie bi- oder allgemeiner multivariate Datensätze geeignet tabellarisch und (wenn möglich) auch graphisch dargestellt werden können. Es werden dann Maße des Zusammenhangs, also Korrelationskoeffizienten vorgestellt und interpretiert.

2. Darstellung mehrdimensionaler Datensätze

a) Verbundene Beobachtungen, gemeinsame Verteilung

Hinsichtlich der Art der Daten soll im folgenden vorausgesetzt werden:

1. es liegen verbundene Beobachtungen von mehreren Merkmalen vor,
2. diese Merkmale sind metrisch skaliert (mindestens Intervallskala),
3. die Daten können Einzelbeobachtungen, gruppierte oder klassierte Merkmale sein.

Erläuterungen:

zu 1: Was mit verbundenen Beobachtungen gemeint ist, soll durch das einführende Beispiel (Bsp. 7.1) veranschaulicht werden. Im folgenden beschränken wir uns auf zwei Merkmale.

zu 2: Es ist hinsichtlich der Art (Skalen) der Merkmale zu unterscheiden zwischen folgenden Fällen (bei Beschränkung auf zwei Merkmale):

a) zwei auf gleichem Skalenniveau skalierte Merkmale, etwa beide intervallskaliert oder beide nominalskaliert, wie im Bsp. 7.1;

b) Merkmale auf unterschiedlichem Skalenniveau, etwa X nominalskaliert und Y intervallskaliert.

Es soll im folgenden zunächst der Fall a) betrachtet werden.

zu 3: Die Methoden dieses Kapitels werden (im Unterschied zu denen des Kapitels 8) in der Regel nicht für den Fall von Einzelbeobachtungen demonstriert.

Beispiel 7.1:

Bei drei Klausuren A, B und C wurde der Zusammenhang zwischen Geschlecht (Merkmal X) und Klausurleistung (Merkmal Y) untersucht. Es gab jeweils 200 Klausurteilnehmer, darunter 150 Männer und 50 Frauen und jede der Klausuren wurde von 70% der Teilnehmer bestanden (also von 140 Personen) und von 30% nicht bestanden. Man erhielt die folgenden Daten:

x_1 = männlich x_2 = weiblich y_1 = bestanden y_2 = nicht bestanden.

Klausur A

	y_1	y_2	Σ
x_1	105	45	150
x_2	35	15	50
Σ	140	60	200

Klausur B

	y_1	y_2	Σ
x_1	140	10	150
x_2	0	50	50
Σ	140	60	200

Klausur C

	y_1	y_2	Σ
x_1	90	60	150
x_2	50	0	50
Σ	140	60	200

a) Erläutern Sie die Art der Zusammenstellung der Daten und interpretieren Sie die Daten.

b) Worin besteht der Unterschied zwischen verbundenen und unverbundenen Beobachtungen?

Lösung 7.1:

a) Es handelt sich um Vierfeldertafeln, jeweils eine spezielle Form der zweidimensionalen Häufigkeitsverteilung (Kontingenztabelle). Die

Summenzeilen und -spalten stellen die Randverteilungen der Variablen X und Y dar (vgl. Def. 7.3).

b) Man sieht, dass Art und Stärke des Zusammenhangs in den drei Klausuren sehr unterschiedlich sind, obgleich die Randverteilungen gleich sind: das Beispiel soll v.a. zeigen, dass die Kenntnis der (eindimensionalen) Randverteilungen nicht ausreicht, um einen Zusammenhang zu beurteilen. Die Unterschiede der drei Klausuren werden deutlich, wenn man die "Durchfallquoten" von Männern und Frauen bei den drei Klausuren betrachtet (d.h. praktisch die bedingten Verteilungen - vgl. Def. 7.4 - des Merkmals Y untersucht):

	Anteil nicht-bestandener Klausuren in vH		
Klausur	Männer	Frauen	insgesamt
A	30%(45/150)	30%(=15/50)	30%(=60/200)
B	6,7%	100%	30%
C	40%	0%	30%

Im Falle der Klausur A besteht Unabhängigkeit der beiden Variablen X (Geschlecht) und Y (Klausurleistung)(vgl. Def.7.5).

Def. 7.1: Verbundene Beobachtungen

a) im Falle von Einzelbeobachtungen:
Wird jede Einheit v =1,2,...,n mit zwei Merkmalen, d.h. einem Tupel (x_v,y_v), mit drei Merkmalen [einem Tripel (x_v,y_v,z_v)] oder mit p Merkmalen (p-Tupel) beschrieben, so spricht man von verbundenen Beobachtungen (im Rahmen einer zwei-, drei-,..., p-dimensionalen Messung) [im folgenden Beschränkung auf p = 2 Dimensionen].

b) bei gruppierten Daten:
Das Merkmal X habe die Ausprägungen $x_1,x_3,...,x_m$ oder allgemein x_i (i=1,2,...,m) und das Merkmal Y habe die Ausprägungen y_j (j = 1,2,...,k). Dann ist n_{ij} die Anzahl der Einheiten mit den Ausprägungen $X = x_i$ **und** $Y = y_j$ (also die Anzahl gleicher Wertetupel). Wie im Falle der eindimensionalen Häufigkeitsverteilung n(...) eine Funktion ist, die einer Merkmalsausprägung eine absolute Häufigkeit zuordnet, so soll n(...) hier einer **Kombination** von Merkmalsausprägungen eine absolute Häufigkeit zuordnen:

(7.1) $n_{ij} = n(X = x_i \text{ und } Y = y_j)$ $(i = 1,...,m \text{ und } j=1,...,k)$.

Für die relativen Häufigkeiten gilt analog zur eindimensionalen Häufigkeitsverteilung

(7.2) $h_{ij} = n_{ij}/n \quad \text{mit} \quad n = \sum_i \sum_j n_{ij} = \sum_{i,j} n_{ij}$

c) bei klassierten Daten gilt b) analog.

Bemerkungen zu Def. 7.1:

1. Auf die in Gl. 7.2 erscheinende Doppelsumme wird in Def. 7.2 näher eingegangen. Für die mk (m Ausprägungen von X und k Ausprägungen von Y) relativen Häufigkeiten h_{ij} gilt analog zum eindimensionalen Datensatz
   $0 \le h_{ij} \le 1 \quad \text{und} \quad \Sigma\Sigma h_{ij} = 1.$
2. Ein erstes Beispiel für den Fall gruppierter Daten, wenn also gleiche Merkmalsausprägungen von X und Y gehäuft auftreten, ist das Beispiel 7.1, das im übrigen demonstriert, dass die (simultane) Betrachtung zweidimensionaler Messungen (Beobachtungen) nicht identisch ist mit der isolierten Betrachtung zweier eindimensionaler Messungen.
3. Es gibt folgende Darstellungen verbundener Beobachtungen:

	graphisch *)	tabellarisch
Einzelbeobachtungen	Streuungsdiagramm [oder Streudiagramm] (vgl. Beispiel 7.2)	Urliste von n Tupeln (Wertepaaren)
gruppierte und klassierte Daten	dreidimensionales Histogramm **)	Kontingenztafeln (vgl. Def. 7.2)

*) metrische Skala vorausgesetzt.

**) für eine zweidimensionale Häufigkeitsverteilung (je eine Achse für die Variablen X und Y und eine Achse für die Häufigkeiten).

Def. 7.2: Zweidimensionale Häufigkeitsverteilung

Eine zweidimensionale Häufigkeitsverteilung ist eine Zuordnung der gemeinsamen absoluten (n_{ij}) oder relativen (h_{ij}) Häufigkeiten zu den Ausprägungen x_i des Merkmals (der Variablen) X und y_j des Merkmals (der Variablen) Y nach Art nachfolgender Tabelle (Matrix). Bei kategorialen (nominalskalierten) Merkmalen spricht man auch von einer Kontingenztafel.

Zweidimensionale Häufigkeitsverteilung
(relative Häufigkeiten)

Merkmal X	Merkmal Y					
	y_1	y_2	...	y_j	...	y_k
x_1	h_{11}	h_{12}	...	h_{1j}	...	h_{1k}
x_2	h_{21}	h_{22}	...	h_{2j}	...	h_{2k}
.	.	.	.	.	.	.
.	.	.	.	.	.	.
x_i	h_{i1}	h_{i2}	...	h_{ij}	...	h_{ik}
.	.	.	.	.	.	.
.	.	.	.	.	.	.
x_m	h_{m1}	h_{m2}	...	h_{mj}	...	h_{mk}

Zum Sprachgebrauch:

Der Begriff Kontingenztafel wird von vielen Autoren auch bei metrisch skalierten Variablen benutzt. Die absoluten oder relativen Häufigkeiten heißen auch gemeinsame Häufigkeiten und die gesamte Häufigkeitsverteilung auch **gemeinsame Häufigkeitsverteilung.** Die Größen x_i (i=1,2,...,m), bzw. y_j (j=1,2,...,k) können Merkmalsausprägungen (gruppierte Daten) oder Größenklassen (klassierte Daten) der Merkmale X und Y bezeichnen.

Bemerkungen zu Def. 7.2:

1. In der gleichen Art, wie die gemeinsamen relativen Häufigkeiten h_{ij} dargestellt wurden, lassen sich auch die (gemeinsamen) absoluten Häufigkeiten n_{ij} und die (gemeinsamen) relativen oder absoluten Summenhäufigkeiten (H_{ij}, bzw. N_{ij}) darstellen, wobei gilt:

$$(7.3) \qquad N_{ij} = \sum_{x \le i}^{n} \sum_{y \le j}^{n} n_{xy} \quad \text{und} \quad H_{ij} = N_{ij}/n.$$

2. Jeder mehrdimensionalen Häufigkeitsverteilung sind weitere Verteilungen zugeordnet (d.h. sie ergeben sich hieraus), nämlich die Randverteilungen und die bedingten Verteilungen (vgl. die nachfolgenden Definitionen).

Anhand des folgenden Beispiels 7.2 soll das Streuungsdiagramm (oder Streudiagramm, scatter diagram) erklärt werden.

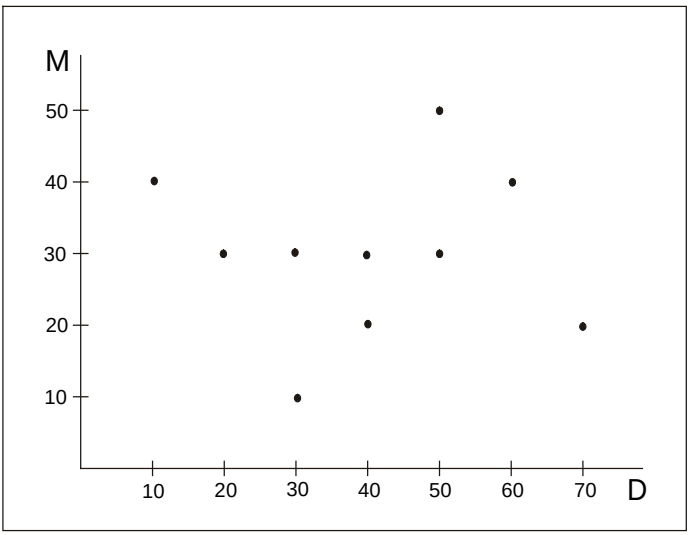
Beispiel 7.2: Streuungsdiagramm

König Egon der XIII, auch der "Labile" genannt, hatte zwei Mätressen, die Pompadur (D) und die Pompamoll (M), die miteinander heftig um die Gunst des Königs konkurrierten. Dass sie jeweils verschiedene Seiten des empfindsamen Gemüts des Königs ansprachen und für ihn deshalb komplementär waren, steht seit der These des berühmten Historikers H in allen Lehrbüchern. H's jüngerer Kollege h glaubt dies jedoch aufgrund einer seinerzeit von der Hofschranze S verfassten Notiz empirisch widerlegen zu können. Aus dieser Notiz geht hervor, wie Egon seine Freizeit (gemessen in Stunden) in den letzten 10 Wochen des Jahres 1789 auf die Damen aufteilte:

D	40	30	20	10	40	30	50	50	60	70
M	30	10	30	40	20	30	50	30	40	20

Zeichnen Sie das Streuungsdiagramm! (Die Aufgabe wird fortgesetzt.)

Abb.7.1: Streuungsdiagramm für das Beispiel 7.2



Lösung 7.2: vgl. Abb. 7.1

Abb. 7.1 zeigt das Streuungsdiagramm für dieses Beispiel. Es wird später gezeigt, dass die beiden Variablen D und M nicht miteinander korrelieren und deshalb die These von H, derzufolge mit einer hohen positiven Korrelation zu rechnen ist, vermutlich nicht aufrechtzuerhalten ist.

b) Aus der gemeinsamen Verteilung abgeleitete Verteilungen

Jeder gemeinsamen Verteilung sind zwei weitere Verteilungstypen zugeordnet, in dem Sinne, dass sie aus der gemeinsamen Verteilung hergeleitet sind:

1. Eine p-dimensionale Verteilung besteht aus einer gemeinsamen Verteilung und p jeweils (p-1)-dimensionalen Randverteilungen (Def. 7.3); im Falle von zwei Dimensionen (p=2) sind es also zwei eindimensionale Randverteilungen.
2. Es gibt ferner jeweils eindimensionale bedingte Verteilungen (Def. 7.4): Jeder Zeile und jeder Spalte der gemeinsamen Verteilung entspricht jeweils eine bedingte Verteilung (die stets eindimensional ist).

Def. 7.3: Randverteilungen, marginal distributions

Da die Ausprägung x_i bei den Kombinationen (x_i,y_1), (x_i,y_2),..., (x_i,y_k) also allen Merkmalskombinationen der i-ten Zeile der zweidimensionalen Häufigkeitsverteilung (Kontingenztabelle) vorliegt, ist die Randhäufigkeit $h_{i.}$ definiert als Zeilensumme

$$(7.4) \quad h_{i.} = \sum_{j=1}^{j=k} h_{ij} = \sum_{j} h_{ij} = h(X=x_i).$$

Die als Summen von Zeilen gebildeten Randhäufigkeiten $h_{1.}$, $h_{2.}$,...,$h_{m.}$ stellen die Randverteilung $h_x(x)$ der Variablen X dar.
Entsprechend bilden die als Summen von Spalten definierten Randhäufigkeiten $h_{.1}$,$h_{.2}$,...,$h_{.k}$ die Randverteilung $h_y(y)$ des Merkmals (der Variablen) Y, wobei gilt:

$$(7.5) \quad h_{.j} = \sum_{i=1}^{i=m} h_{ij} = \sum_{i} h_{ij} = h(Y=y_j).$$

Die Randverteilungen ausgedrückt in absoluten Häufigkeiten $n_x(x)$ mit den über k Spalten summierten absoluten Häufigkeiten einer Zeile

$$(7.4a) \quad n_{i.} = n_{i1} + n_{i2} + ... + n_{ik}$$

und die Randverteilung $n_y(y)$ mit den k absoluten Häufigkeiten $n_{.j}$ sind entsprechend definiert.

Die beiden Randverteilungen (in relativen Häufigkeiten) sind in der folgenden Tabelle besonders durch Einrahmung markiert:

Merkmal X	Merkmal Y: y_1	y_2	...	y_j	...	y_k	$\Sigma: h_x(x)$
x_1	h_{11}	h_{12}	...	h_{1j}	...	h_{1k}	$h_{1.}$
x_2	h_{21}	h_{22}	...	h_{2j}	...	h_{2k}	$h_{2.}$
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
x_i	h_{i1}	h_{i2}	...	h_{ij}	...	h_{ik}	$h_{i.}$
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
x_m	h_{m1}	h_{m2}	...	h_{mj}	...	h_{mk}	$h_{m.}$
$\Sigma: h_y(y)$	$h_{.1}$	$h_{.2}$	...	$h_{.j}$	...	$h_{.k}$	1

Die Summenspalte $h_x(x)$ ist die Randverteilung von X und die Summenzeile $h_y(y)$ ist die Randverteilung von Y.

Auch bei den Randverteilungen kann man - wie bei jeder Verteilung - unterscheiden zwischen der Häufigkeitsfunktion (-verteilung) und der **Verteilungsfunktion (Summenverteilung, kumulierte Häufigkeiten**) und das jeweils bei absoluten und bei relativen Häufigkeiten. Es ist ferner offensichtlich, dass alle Kennzahlen eindimensionaler Häufigkeitsverteilungen (z.B. Mittelwerte, Streuungsmaße, Schiefemaße usw.) auch für die Randverteilungen berechnet werden können (vgl. Abschn. 3).

Def. 7.4: bedingte Verteilung

Die durch Gl. 7.6 definierten bedingten relativen Häufigkeiten $h_{i\,|\,j}$ stellen die bedingte Häufigkeitsfunktion (-verteilung) von X, gegeben $Y = y_j$ dar, also die Spalte j als eine Verteilung:

$$(7.6) \quad h_{i\,|\,j} = \frac{h_{ij}}{h_{.j}} = \frac{n_{ij}}{n_{.j}} = h(x|Y=y_j) \; .$$

Analog ist die bedingte Häufigkeitsfunktion (-verteilung) von Y definiert durch die relativen Häufigkeiten der Ausprägung $y_1, y_2, ..., y_k$ (allgemein: y_j) "gegeben $X = x_i$" (oder: bedingt durch x_i, oder: wenn $X = x_i$)

$$(7.7) \quad h_{j\,|\,i} = \frac{h_{ij}}{h_{.i}} = \frac{n_{ij}}{n_{i.}} = h(y|X=x_i) \; .$$

Es ist eine Zeile (die i-te) als Häufigkeitsverteilung.

Bemerkungen zu Def. 7.4:

1. Es ist in diesem Fall nur üblich, von *relativen* (nicht absoluten) Häufigkeiten auszugehen. Es gilt, wie leicht zu zeigen ist:

$$\text{(7.6a)} \quad \sum_{i=1}^{i=m} h_{i\,|\,j} = 1 \qquad \text{und} \qquad \text{(7.7a)} \quad \sum_{j=1}^{j=k} h_{j\,|\,i} = 1 \quad .$$

2. Mit Gl. 7.6 werden aus den k Spalten (nach den Ausprägungen des Merkmals Y, j=1,2,...,k) und mit Gl. 7.7 aus den m Zeilen (i=1,2,...,m) der Matrix der gemeinsamen Verteilung jeweils Häufigkeitsverteilungen gebildet.

Def. 7.5: Unabhängigkeit

Unabhängigkeit lässt sich auf zwei Arten definieren:

1. Sind die k bedingten Verteilungen $h_{i\,|\,j}$ des Merkmals X bei allen Ausprägungen y_j (j = 1,2,...,k) des Merkmals Y identisch (und damit auch gleich der [unbedingten] Randverteilung von X), so sind X und Y unabhängig (analog gilt: Gleichheit der m bedingten Verteilungen $h_{j\,|\,i}$ bei allen Ausprägungen von X des Merkmals Y, bedeutet Unabhängigkeit von X und Y).
2. Im Falle der Unabhängigkeit ergeben sich die absoluten, bzw. relativen gemeinsamen Häufigkeiten aus den entsprechenden Häufigkeiten der Randverteilungen gemäß

$$\text{(7.8)} \quad n_{ij} = \frac{n_{i.}\, n_{.j}}{n} \qquad \text{bzw.} \qquad \text{(7.8a)} \quad h_{ij} = h_{i.} h_{.j} \quad .$$

Folgerungen:

1. Unabhängigkeit ist die stärkere Forderung als die später zu besprechende (vgl. Def. 7.7) Unkorreliertheit:

Satz 7.1:

Unabhängigkeit impliziert Unkorreliertheit aber nicht umgekehrt, d.h. Unkorreliertheit kann bestehen, obgleich die Variablen X und Y nicht unabhängig sind.

Beweis: siehe Bem. Nr.9 zu Def. 7.6.

2. Aus Teil 1 der Def. 7.5 folgt, dass bei Unabhängigkeit jeweils alle bedingten Verteilungen mit der Randverteilung identisch sind. Gilt Identität der k bedingten Verteilungen des Merkmals X, so sind auch die bedingten Verteilungen des Merkmals Y

für alle m Ausprägungen von X identisch, d.h. Unabhängigkeit ist eine symmetrische Relation: ist X unabhängig von Y, so ist auch Y unabhängig von X.

3. Beide Arten der Definition der Unabhängigkeit sind äquivalent, d.h. eine Art der Definition lässt sich jeweils aus der anderen folgern. Aus Folgerung 2 ergibt sich $n_{ij} / n_{i.} = n_{.j} / n$ und hieraus folgt Gl. 7.8.

4. Umgekehrt gilt: Aus Gl. 7.8a folgt

$$h_{i \mid j} = \frac{h_{ij}}{h_{.j}} = \frac{h_{i.}\, h_{.j}}{h_{.j}} = \frac{h_{ik}}{h_{.k}} = h_{i \mid k} = \frac{h_{i.}\, h_{.k}}{h_{.k}} = h_{i.} ,$$

d.h. bei Unabhängigkeit ist Gleichheit der bedingten Verteilungen von X gegeben (X bedingt durch $Y = y_j$ und X bedingt durch $Y = y_k$).

Beispiel 7.3:

Einer fehlgeschlagenen Intrige bei Hofe hat es Graf Giselher von Gelsenkirchen zu verdanken, dass er in einem Burgverlies schmachtet. Statt vor dem Verwaltungsgericht Gelsenkirchen zu klagen, (diese neuzeitliche Denkweise war Giselher noch vollkommen fremd) machte er sich daran, die meterdicke Wand zu durchbohren. Es gibt Tage, an denen er y=1, y=2 und y=3 Zentimeter der Wand wegschaben konnte. Über den Zeitaufwand X des Schabens (in Stunden) und die Zentimeterleistung Y des Verdünnens der Wand bestehen für 13 Tage Aufzeichnungen des Grafen:

X: Arbeitszeit	Y: Leistung 1cm	2cm	3cm
6Std.	1	2	0
8Std.	1	3	1
10Std.	1	2	2

Man bestimme sowohl die beiden Randverteilungen als auch die bedingten Verteilungen sowie die absoluten gemeinsamen Summenhäufigkeiten. Sind die beiden Merkmale X und Y unabhängig?

Lösung 7.3:

Randverteilungen			
x_i	$n_{i.}$	y_j	$n_{.j}$
6	3	1	3
8	5	2	7
10	5	3	3

Bedingte Verteilungen							
	y=1	**y=2**	**y=3**		**x=6**	**x=8**	**x=10**
x=6	1/3	2/7	0	**y=1**	1/3	0,2	0,2
x=8	1/3	3/7	1/3	**y=2**	2/3	0,6	0,4
x=10	1/3	2/7	2/3	**y=3**	0	0,2	0,4
von x (bedingt durch y)				von y (bedingt durch x)			

X und Y sind nicht unabhängig. Es genügt, ein Tabellenfeld der gemeinsamen Verteilung zu überprüfen. Nach Gl. 7.8 müsste für n_{12} gelten $n_{12} = n_{1.}\, n_{.2} /n = (3 \cdot 7)/13 = 1{,}615$ statt $n_{12} = 2$. Auch kein anderes Feld der gemeinsamen Verteilung erfüllt Gl. 7.8. So ist etwa $n_{22} = 3$ statt $(5 \cdot 7)/13 = 2{,}692$. Dass keine Unabhängigkeit gegeben ist, kann man auch daran erkennen, dass die bedingten Verteilungen verschieden sind. In der Aufgabe war auch verlangt, die absoluten Summenhäufigkeiten zu bestimmen. Die Größe N_{ij} ist die absolute Häufigkeit für $x \le x_i$ und $y \le y_j$

	y=1 cm	y=2 cm	y=3 cm
x=6 Std.	1	3	3
x=8 Std.	2	7	8
x=10 Std.	3	10	13

Die Häufigkeit 7 in dieser Tabelle bedeutet, dass es 7 Tage gibt, an denen Giselher bis zu höchstens 8 Stunden arbeitet (also 6 oder 8 Stunden) und dabei bis zu höchstens 2 cm der Wand abschabt.

3. Kennzahlen zur Beschreibung einer zweidimensionalen Verteilung

Einer gemeinsamen Verteilung h(x,y) der (mindestens intervallskalierten Variablen X und Y) sind jeweils eindimensionale Randverteilungen und bedingte Verteilungen zugeordnet. Jede der genannten Verteilungen lässt sich durch Kenngrößen (Maßzahlen, Parameter) beschreiben, die

- eindimensionalen Verteilungen (Randverteilungen, bedingte Verteilungen) durch Mittelwerte (auch Mediane), Varianzen etc.;
- zweidimensionale gemeinsame Verteilung durch die Kovarianz und den Korrelationskoeffizienten.

a) Kennzahlen der eindimensionalen Verteilungen und Regressionslinie

Übersicht 7.1 stellt die Zusammenhänge zwischen den Verteilungen und den im folgenden zu behandelnden beschreibenden Kennzahlen (Parameter) dar.
Es sollen zunächst die Parameter der eindimensionalen Verteilungen betrachtet werden (Nr. 1 und 2) und im nächsten Abschnitt die der gemeinsamen Verteilung:

1) Mittelwert und Varianz der Randverteilungen

Mittelwert $\bar{x}$ der Randverteilung $h_x(x)$

(7.9) $\bar{x} = \Sigma x_i h_{i.} = \sum_i x_i h_{ij}$

und die Varianz

(7.10) $s_x^2 = \Sigma x_i^2 h_{i.} - \bar{x}^2$.

Die entsprechenden Parameter der Randverteilung $h_y(y)$ sind analog definiert.

2) Parameter der bedingten Verteilungen

a) Die wichtigsten Parameter der bedingten (Häufigkeits-) Verteilungen sind die **bedingten Mittelwerte**

(7.11) $\bar{x}|y = \bar{x}(y_j) = \sum_{i=1}^{i=m} x_i h_{i|j}$

(7.12) $\bar{y}|x = \bar{y}(x_i) = \sum_{j=1}^{j=k} y_j h_{j|i}$.

b) Seltener ist die Berechnung der **bedingten Varianzen**

(7.12a) $s_x^2(y_j) = \sum_{i=1}^{i=m} [x_i - \bar{x}(y_j)]^2 h_{i|j} = \sum_{i=1}^{i=m} x_i^2 h_{i|j} - [\bar{x}(y_j)]^2$

und $s_y^2(x_i)$ analog, bzw. der bedingten Standardabweichungen $s_x(y_j)$ und
$s_y(x_i)$.

Mit den bedingten Varianzen wird die Streuung der Beobachtungen um die Regressionslinie (vgl. Def. 7.6) gemessen. Sie ist Teil der internen Streuung in einer Varianzzerlegung (vgl. Satz 7.4).

Übersicht 7.1

a) Zusammenhänge zwischen den Verteilungen

Ausgangspunkt: **Zweidimensionale** gemeinsame Verteilung, d.h.
eine Matrix mit m Zeilen für die Ausprägungen
x_i (i = 1,,2,...,m) und k Spalten für y_j (j = 1,2,...,k)
daraus abgeleitete **eindimensionale** Verteilungen

zwei Randverteilungen für X und Y: $h_x(x)$ mit den Häufigkeiten $h_{i.}$ und $h_y(y)$ mit den Häufigkeiten $h_{.j}$	m bedingte Verteilungen von Y (bedingt durch m Ausprägungen von X) und k bedingte Verteilungen von X (bedingt durch Y)

b) Beschreibende Kennzahlen* **der**

zweidimensionalen Verteilung	**eindimensionalen Verteilungen**	
Kovarianz s_{xy} und Korrelationskoeffizient r_{xy}	Randverteilungen: Mittelwerte und Varianzen	bedingte Verteilungen: bedingte Mittelwerte (Regressionslinie)

*) Nur die am häufigsten verwendeten Maßzahlen (Kennzahlen). Man kann natürlich auch z.B. die Schiefe der Randverteilungen bestimmen oder bedingte Varianzen berechnen.

Def. 7.6: empirische Regressionslinie

Die lineare Verbindung der bedingten Mittelwerte $\bar{x}|y$ ist die Regressionslinie (empirische Regressionslinie) der Variablen X. Entsprechend ist die lineare Verbindung der Punkte P(x, $\bar{y}|x$) die Regressionslinie der Variablen Y.

Der Begriff Regressions"linie" soll deutlich machen, dass die Punkte nicht notwendig auf einer Geraden liegen müssen. Es sind also Regressionslinie und Regressionsgerade (Kap. 8) zu unterscheiden.

Selbst wenn die Regressionslinie eine Gerade ist, muss sie nicht identisch mit der Regressionsgeraden sein.

Beispiel 7.4:

Man bestimme und zeichne die Regressionslinien für das Bsp. 7.3!

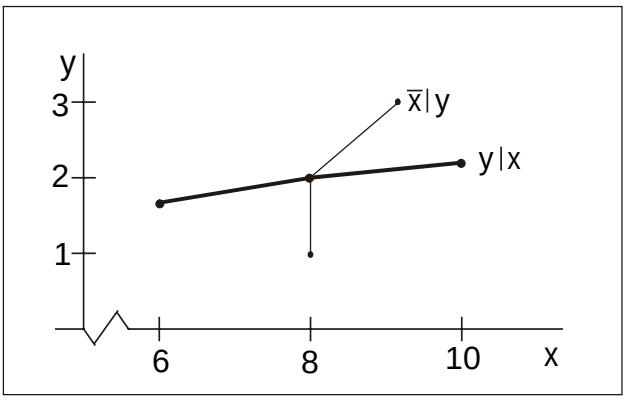
Lösung 7.4:

Bedingte Mittelwerte

von y:		von x:	
$\bar{y}\|x=6$	$= 1{,}67$	$\bar{x}\|y=1$	$= 8$
$\bar{y}\|x=8$	$= 2$	$\bar{x}\|y=2$	$= 8$
$\bar{y}\|x=10$	$= 2{,}2$	$\bar{x}\|y=3$	$= 9{,}33$

Zur graphischen Darstellung der Regressionslinien vgl. Abb. 7.2

Abb. 7.2: Regressionslinien (Beispiel 7.4 bzw. 7.3)



Beispiel 7.5:

Der Student S glaubt wieder einmal eine Recht-Klausur ganz astrein gelöst zu haben. Mit seiner Selbsteinschätzung (Variable X), die mehr oder weniger gefühlsmäßig und zufällig, weniger aus tiefer juristischer Einsicht erfolgt, liegt er zwar oft in der Tendenz ganz richtig. Die genaue Klausurnote (Y) erscheint ihm aber fast immer rätselhaft und unerklärlich. So wie es ihm geht, ergeht es jedoch auch seinen 35 Mitstudenten. Dass die Noten bei den Rechtklausuren irgendwie mysteriös sind, scheinen inzwischen fast alle zu glauben, wie die folgende Gegenüberstellung von X und Y für alle 36 Studenten zeigt:

		Variable Y				
		1	2	3	4	5
Variable X	1	1	2	3	2	0
	2	1	2	2	1	0
	3	0	1	2	2	1
	4	0	0	3	4	3
	5	0	2	1	1	2

Bestimmen Sie die empirischen Regressionslinien!

Lösung 7.5:

$(\overline{y}|x=1) = 2{,}75$ $(\overline{x}|y=1) = 1{,}5$
$(\overline{y}|x=2) = 2{,}5$ $(\overline{x}|y=2) = 2{,}714 = 19/7$
$(\overline{y}|x=3) = 3{,}5$ $(\overline{x}|y=3) = 2{,}727 = 30/11$
$(\overline{y}|x=4) = 4$ $(\overline{x}|y=4) = 3{,}1$
$(\overline{y}|x=5) = 3{,}5$ $(\overline{x}|y=5) = 4{,}167 = 25/6$

Beispiel 7.6:

Gegeben sei die folgende zweidimensionale Häufigkeitsverteilung (Angabe mit relativen Häufigkeiten) für welche die Regressionslinien sowie die Parameter der Randverteilungen zu bestimmen sind:

	y=2	y=3	y=4
x=2	0,2	0,5	0,1
x=3	0,1	0,1	0

Lösung 7.6:

Mittelwerte $\overline{x} = 2{,}2$ und $\overline{y} = 2{,}8$; Varianzen: $s_x^2 = 0{,}16$ und $s_y^2 = 0{,}36$. Die bedingten Mittelwerte von x lauten $\overline{x}|y=2 = 2{,}33$, $\overline{x}|y=3 = 2{,}167$ und $\overline{x}|y=4 = 2$. Diejenigen von y lauten $\overline{y}|x=2 = 2{,}875$ und $\overline{y}|x=3 = 2{,}5$. Beide Regressions**linien** sind Geraden (sie sind übrigens, wie später gezeigt wird, identisch mit den Regressions**geraden**).

b) Kovarianz und Korrelationskoeffizient

Die Kovarianz ist Ausdruck des Zusammenhangs zwischen zwei metrisch skalierten Variablen. Als Maß für den Grad (die Intensität) des Zusammenhangs ist sie im Unterschied zur Korrelation jedoch nicht geeignet, weil ihr Betrag von der Maßeinheit der Variablen X und Y abhängt. Über Kovarianz und Korrelation im Falle einer klassierten Verteilung vgl. Abschn. 3e dieses Kapitels.

1. Kovarianz

Def. 7.7: Kovarianz

Die Kovarianz ist als beschreibende Kennzahl einer zweidimensionalen Verteilung definiert als

$$(7.13)\quad s_{xy} = \frac{1}{n} \Sigma (x_v - \bar{x})(y_v - \bar{y}) \quad \text{bei n Einzelbeobachtungen}$$

bzw. bei gruppierten Daten mit absoluten gemeinsamen Häufigkeiten

$$(7.14)\quad s_{xy} = \frac{1}{n} \sum_{i=1}^{m} \sum_{j=1}^{k} (x_i - \bar{x})(y_j - \bar{y}) n_{ij}$$

und mit relativen Häufigkeiten

$$(7.14a)\quad s_{xy} = \Sigma\Sigma (x_i - \bar{x})(y_j - \bar{y}) h_{ij}.$$

Zur Interpretation der Kovarianz (Bemerkungen zu Def. 7.7)

1. a) Die Kovarianz heisst auch zentrales **Produktmoment**: "zentral", weil die Mittelwerte $\bar{x}$ und $\bar{y}$ jeweils abgezogen werden von x_v bzw. y_v und "Produkt", weil diese Abweichungen miteinander multipliziert werden. Starke Abweichungen vom jeweiligen Mittelwert beeinflussen die Kovarianz stark, so dass diese empfindlich ist gegenüber Ausreißern.

   b) Ähnlich wie bei der Varianz kann auch hier unterschieden werden zwischen der Berechnung der Kovarianz

   - als beschreibende Statistik einer Stichprobe gem. Gl. 7.13, bzw. 7.14 und
   - als Schätzwert für die Kovarianz der Grundgesamtheit aufgrund der Stichprobe: dann ist durch n-1 statt durch n zu dividieren.

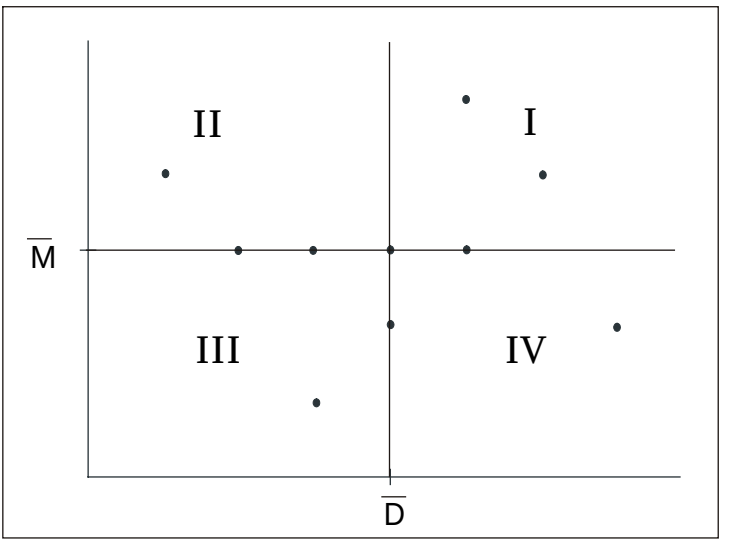
2. Es ist unmittelbar zu sehen, dass die Varianz als Spezialfall der Kovarianz aufgefaßt werden kann $s_{xx} = s_x^2$. Die Kovarianz ist außerdem symmetrisch, d.h. es gilt $s_{xy} = s_{yx}$.

3. Weil ein Produkt von Abweichungen gebildet wird, kann die Kovarianz positiv oder negativ sein. Es ist üblich, sich die Bedeutung der Mittelung über Abweichungsprodukte bei der Kovarianz anhand der Abb. 7.3 zu verdeutlichen.

Daten von Beispiel 7.2: D = Pompadur und M = Pompamoll; die arithmetischen Mittel sind $\overline{D} = 40$ und $\overline{M} = 30$. Das Streuungsdiagramm zeigt Unkorreliertheit der Variablen an.

- Bei Beobachtungen im Quadrant I sind die Abweichungen $(x_v-\overline{x})$ und $(y_v-\overline{y})$ **beide** jeweils positiv, so dass $(x_v-\overline{x})(y_v-\overline{y})>0$;
- Quadrant III $(x_v-\overline{x})(y_v-\overline{y}) > 0$ (weil beide Abweichungen negativ sind);
- Gegenläufig verhalten sich dagegen X und Y (d.h. negative Korrelation) in den Quadranten II und IV, weshalb dann das Produkt $(x_v-\overline{x})(y_v-\overline{y}) < 0$ also negativ ist.

Abb. 7.3: Streuungsdiagramm (Bsp. 7.2) zur Veranschaulichung des Konzepts der Kovarianz



Liegen die Punkte im Streuungsdiagramm hauptsächlich in den Quadranten I und III, so liegt eine positive Korrelation vor, liegen sie vorwiegend in den Quadranten II und IV, so ist die Korrelation negativ. Man kann ein Korrelationsmaß definieren, indem man einfach die Anzahl n_k (k = konkordant) der Punkte im Quadranten I und III mit der Anzahl n_d (d=diskonkordant) der Punkte im Quadrant II und IV vergleicht: **Fechners Korrelationskoeffizient** $r_F = (n_k-n_d)/(n_k+n_d)$.

4. Die Kovarianz ist betragsmäßig nicht beschränkt. Sie ist deshalb nicht geeignet für die Messung des Zusammenhangs zwischen zwei Vari-

ablen X und Y. Der Korrelationskoeffizient (vgl. Def. 7.8) ist dagegen die auf den Wertebereich von -1 bis +1 normierte Kovarianz.

5. Die Kovarianz ist nicht invariant gegenüber linearen Transformationen. Man kann leicht zeigen, dass die Kovarianz der transformierten Variablen $x^* = a + bx$ und $y^* = c + dy$ wie folgt mit der Kovarianz zwischen x und y zusammenhängt:

(7.15) $s_{x^*y^*} = bds_{xy}$ (Kovarianz bei Lineartransformation)

6. Auch für die Kovarianz gilt der **Verschiebungssatz**:

(7.13a) $s_{xy} = \frac{1}{n}\Sigma x_v y_v - \bar{x}\cdot\bar{y}$

bzw. bei gruppierten Daten mit absoluten Häufigkeiten n_{ij}

(7.14a) $s_{xy} = \frac{1}{n}\Sigma\Sigma x_i y_j n_{ij} - \bar{x}\cdot\bar{y}$

und relativen gemeinsamen Häufigkeiten h_{ij}

(7.14b) $s_{xy} = \Sigma\Sigma x_i y_j h_{ij} - \bar{x}\cdot\bar{y}$

oder

(7.16) $s_{xy} = \overline{xy} - \bar{x}\cdot\bar{y}.$

Hierin ist $\overline{xy}$ der Mittelwert des Produkts der x und y Werte und $\bar{x}\cdot\bar{y}$ ist das Produkt der Mittelwerte.

7. Die damit gegebene Beziehung zwischen dem Anfangsproduktmoment xy,¾ und dem zentralen Produktmoment s_{xy} führt auch wegen der Schwerpunkteigenschaft des arithmetischen Mittels zu folgenden Darstellungen der Kovarianz:

(7.17) $s_{xy} = \frac{1}{n}\Sigma(x_v - \bar{x})y_v = \frac{1}{n}\Sigma x_v(y_v - \bar{y}).$

Es genügt also, wenn eine der beiden Variablen zentriert ist.

8. Ein einfach zu zeigender Satz ist:

Satz 7.2:

Verschwindet eine der Varianzen (etwa $s_x^2 = 0$), so ist auch die Kovarianz Null (also $s_{xy} = 0$). Die Umkehrung des Satzes gilt nicht, d.h. $s_{xy} = 0$ ist verträglich mit $s_x^2 > 0$ und $s_y^2 > 0$.

Äquivalent ist die folgende Formulierung:

Die Kovarianz einer Variablen X mit einer Konstanten k ist stets Null, also $s_{xk} = 0$.
Beweis: Dieser Satz ergibt sich unmittelbar aus Gl. 7.17 und ist auch einfach zu interpretieren:

$s_x^2 = 0$ bedeutet, dass alle x-Werte gleich sind, d.h. alle Punkte des Streudiagramms liegen auf einer zur Ordinaten parallelen Geraden. Es liegt dann praktisch eine zu einer eindimensionalen Verteilung degenerierte zweidimensionale Verteilung vor. Soll X die Ursache für Y sein, was impliziert, dass eine nicht unerhebliche Kovarianz zwischen X und Y besteht, so muss X in einem gewissen Maße variieren. Man kann z.B. auch nicht die Ursächlichkeit eines Düngemittels für einen mehr oder weniger großen Ernteertrag nachweisen, d.h. den Zusammenhang "je mehr Dünger desto mehr Ernte" aufzeigen, wenn auf allen Äckern die gleiche Menge gedüngt wird. Man darf andererseits auch nicht folgern: die Variablen X und Y sind umso mehr korreliert, je stärker X und Y streuen. Eine hohe Korrelation ist mit den verschiedensten Varianzen (sofern diese nicht Null sind) verträglich. Das ergibt sich unmittelbar aus der Invarianz des Korrelationskoeffizienten gegenüber linearen Transformationen (vgl. Bemerkung 2 zu Def. 7.8).

9. Unabhängigkeit führt zu einer Kovarianz von Null, denn wenn Gl. 7.8 gilt, dann ist

$$\overline{xy} = \frac{1}{n^2}\sum_i \sum_j x_i n_{i.} y_j n_{.j} = \bar{x}\cdot\bar{y} \text{ und damit } s_{xy} = 0.$$

Die Umkehrung gilt jedoch nicht: die Kovarianz kann sehr wohl verschwinden, obgleich keine Unabhängigkeit besteht (vgl. Beispiele 7.7 und 7.8). Es gilt also der bereits erwähnte

Satz 7.1:

Unabhängigkeit führt zum Verschwinden der Kovarianz (und damit zu Unkorreliertheit). Die Umkehrung gilt jedoch nicht.

Beweis: s.o. und Bemerkungen zu Def. 7.5.

10. Ist Y linear abhängig von X (und damit auch umgekehrt X von Y), etwa dergestalt, dass $y_v = a + bx_v$, so folgt aus Gl. 7.13a $s_{xy} = bs_x^2$.
Ferner ist $s_y^2 = b^2 s_x^2$, so dass bei linearer Abhängigkeit gilt $(s_{xy})^2 = s_x^2 s_y^2$
Wie Def. 7.8. zeigt, bedeutet dies, dass der Korrelationskoeffizient dann den Wert 1 annimmt. In allen anderen Fällen gilt der folgende Satz:

Satz 7.3:

(7.18) $0 \le (s_{xy})^2 \le s_x^2 s_y^2$

Dieser Zusammenhang ist auch bekannt als Schwarzsche - Ungleichung oder Cauchy-Schwarzsche-Ungleichung.

Beweis:

Offenbar gilt wegen der Quadrierung

$0 \le \frac{1}{n}\Sigma[(y_v-\overline{y}) - (s_{xy}/s_y^2)(x_v-\overline{x})]^2$.

Die rechte Seite ergibt den Ausdruck $s_y^2 - 2s_{xy}^2 / s_x^2 + s_{xy}^2 / s_x^2$, der nicht-negativ sein muss, also $0 \le s_y^2 - s_{xy}^2 / s_x^2$. Das ergibt umgeformt Gl. 7.18.

Die Ungleichung lässt sich auch zeigen, wenn man y durch eine Regressionsgerade (vgl. Kapitel 8) erklärt: Es gilt dann $y_v = a + bx_v + u_v$ mit $u = 0$, so dass man erhält: $s_{xy} = bs_x^2 + s_{xu}$ und $s_y^2 = b^2 s_x^2 + s_u^2$. Da im Modell der Regressionsanalyse gilt $s_{xu} = 0$ und $s_u^2 > 0$, ist die Ungleichung 7.18 erfüllt.

11. Der folgende Zusammenhang zwischen der Kovarianz und der mittleren Differenz zwischen den Merkmalswerten ist für manche Betrachtungen (z.B. Rangkorrelation) von Interesse:

Satz 7.4:

Mit der Differenz $d_v = x_v - y_v$ des X-Werts und des Y-Werts der v-ten Einheit (Beobachtung) gilt für die Kovarianz:

(7.19) $s_{xy} = ½[(s_x^2 + s_y^2) + (\overline{x}^2 - \overline{y}^2) - (\Sigma d_v^2)/n]$

Beweis:

$\Sigma d_v^2 = \Sigma(x_v - y_v)^2 = \Sigma x_v^2 - 2\Sigma x_v y_v + \Sigma y_v^2$. Damit ist
$\Sigma x_v y_v = ½(\Sigma x_v^2 + \Sigma y_v^2 - \Sigma d_v^2)$ so dass
$s_{xy} = n^{-1}\Sigma x_v y_v - \overline{x}\cdot\overline{y} = ½(n^{-1}\Sigma x_v^2 + n^{-1}\Sigma y_v^2 - n^{-1}\Sigma d_v^2 - 2\overline{x}\cdot\overline{y})$, womit man nach einigen Umformungen Gl. 7.19 erhält.

Zu weiteren Bemerkungen über die Kovarianz vgl. auch Gl. 7.21ff.

Beispiel 7.7:

Man zeige, dass im Beispiel 7.2 zwar die Kovarianz verschwindet, gleichwohl aber keine Unabhängigkeit vorliegt!

Lösung 7.7:

Für die Variablen D (Pompadur) und M (Pompamoll) erhält man die folgenden Werte: ΣD = 400, ΣM = 300 und ΣMD = 12000, so dass man mit n = 10 erhält s_{DM} = 12000/10 - 40·30 = 0. Um zu zeigen ob M und D unabhängig sind ist aus den Daten eine zweidimensionale Verteilung herzuleiten. Man erhält:

	M=10	20	30	40	50	Σ
D=10	0	0	0	1	0	1
20	0	0	1	0	0	1
30	1	0	1	0	0	2
40	0	1	1	0	0	2
50	0	0	1	0	1	2
60	0	0	0	1	0	1
70	0	1	0	0	0	1
Σ	1	2	4	2	1	10

Es ist unschwer zu erkennen, dass die Häufigkeiten dieser Tabelle nicht mit denen übereinstimmen, die sich bei Unabhängigkeit gem. Gl. 7.8 ergäben: so ist z.B. die relative Häufigkeit für die Kombination D = 40 und M = 30 bei Unabhängigkeit 0,2·0,4 = 0,08, statt des empirischen Werts von 0,1. Bei Unabhängigkeit dürfte auch kein Tabellenfeld eine absolute Häufigkeit von Null ausweisen.
Beispiel 7.8 ist ein weiteres Beispiel dafür, dass die Kovarianz verschwinden kann, obgleich keine Unabhängigkeit vorliegt.

Beispiel 7.8:

Gegeben sei die folgende zweidimensionale Häufigkeitsverteilung (relative Häufigkeiten):

	Y=-2	Y=0	Y=1	Σ
X=4	1/8	1/4	1/8	1/2
X=5	3/16	1/16	1/4	1/2
Σ	5/16	5/16	3/8	1

Man zeige, dass hier (wie im Bsp. 7.7) die Kovarianz zwischen X und Y verschwindet, gleichwohl aber keine Unabhängigkeit vorliegt.

Lösung 7.8:

Bei Unabhängigkeit müßte die relative Häufigkeit h_{12} den Wert ½(5/16) = 0,15625 und nicht 1/8 = 0,125 annehmen. Es genügt, ein Tabellenfeld zu überprüfen, um festzustellen, dass keine Unabhängigkeit vorliegt.

2. Korrelationskoeffizient

Die meisten Bemerkungen zur Kovarianz gelten auch für den Korrelationskoeffizienten, der nur eine auf den Wertebereich von -1 bis +1 normierte Kovarianz darstellt.

Def. 7.8: Korrelationskoeffizient

Der Korrelationskoeffizient nach Bravais-Pearson (auch Produkt-Moment-Korrelationskoeffizient oder im folgenden einfach Korrelationskoeffizient genannt) ist das Verhältnis aus Kovarianz (vgl. Def. 7.7) und dem Produkt der Standardabweichungen

$$(7.20)\quad r_{xy} = \frac{s_{xy}}{s_x s_y} \quad .$$

Bem: Aus den unterschiedlichen Darstellungsmöglichkeiten der Kovarianz (Def. 7.7) und der Varianz ergeben sich auch unterschiedliche Berechnungsformeln für den Korrelationskoeffizienten.

Interpretation und Eigenschaften

1. Aus der Schwarzschen Ungleichung (Satz 7.3, Gl. 7.18) folgt die Einschrankung

   $$(7.20a)\quad -1 \le r_{xy} \le +1 \; .$$

   Der Korrelationskoeffizient ist also die durch das Produkt der Standardabweichungen (oder: das geometrische Mittel der Varianzen) auf den Wertebereich von -1 bis +1 normierte Kovarianz. Die Grenzen r = -1 und r = +1 werden gem. Bem. 10 zu Def. 7.7 erreicht, wenn y eine Lineartransformation von x ist, wenn also gilt y = a + bx. Die Punkte des Streuungsdiagramms liegen dann genau auf der Geraden y = a + bx. Das Vorzeichen des Korrelationskoeffizienten wird allein durch die Kovarianz bestimmt, weil der normierende Nenner $s_x s_y$ stets positiv ist.

2. Durch die Normierung auf den Wertebereich von -1 bis +1 ist der Korrelationskoeffizient (anders als die Kovarianz) maßstabsunabhängig, d.h. er ist unabhängig davon, in welcher Maßeinheit die Variablen X und Y gemessen sind und invariant gegenüber linearen Transformationen.
   Insbesondere gilt bei $x^* = a + bx$ und $y^* = c + dy$ für die Korrelation zwischen x^* und y^* in Relation zur ursprünglichen Korrelation

$$r_{x^*y^*} = \begin{cases} + r_{xy} & \text{wenn } bd > 0 \\ - r_{xy} & \text{wenn } bd < 0 \end{cases}$$

Der Korrelationskoeffizient r_{xy} ist die Kovarianz zwischen den standardisierten Variablen $x^* = (x - \bar{x}) / s_x$ und $y^* = (y - \bar{y}) / s_y$. Die Standardisierung ist eine spezielle Lineartransformation.

3. Die unter Nr. 2 genannte Eigenschaft bedeutet, dass r ein Maß des linearen Zusammenhangs ist. Mit $|r| = 1$ ist ein perfekter **linearer** Zusammenhang gegeben (die Punkte des Streuungsdiagramms liegen genau auf einer Geraden). Je nachdem, wie stark r betragsmäßig vom Wert 1 abweicht, weichen die Punkte mehr oder weniger von den in Kap. 8 behandelten Regressionsgeraden ab. Der Wert $r = 0$ bedeutet nicht, dass kein Zusammenhang zwischen den Variablen X und Y besteht, sondern nur, dass kein **linearer** Zusammenhang besteht (vgl. Beispiel 7.9). Er bedeutet insbesondere auch nicht notwendig Unabhängigkeit; denn nach Satz 7.1 impliziert Unabhängigkeit Unkorreliertheit aber nicht umgekehrt.

4. Gem. Satz 7.2 gilt: Das Verschwinden einer Varianz (etwa $s_x^2 = 0$) impliziert $r_{xy} = 0$. Das bedeutet: Wenn auch nur eine der beiden Variablen X und Y konstant ist, dann können sie auch nicht miteinander korreliert sein.

   Um also eine Korrelation zwischen zwei Variablen X und Y feststellen zu können, müssen beide Variablen streuen. Daraus kann jedoch nicht geschlossen werden, dass die Korrelation umso größer ist, je stärker die beiden Variablen streuen. Denn man kann durch eine Lineartransformation von X zu X^*, etwa $x^* = a + bx$ die Standardabweichung vergrößern (ver-b-fachen), ohne dass sich die Korrelation wegen der Invarianz gegenüber Lineartransformationen ändert ($r_{x^*y} = r_{xy}$).

Weitere Zusammenhänge zwischen Varianzen, Kovarianzen und Korrelationskoeffizienten

Für die folgende Darstellung ist es nützlich, die Notation etwas zu vereinfachen: statt mit s_{xy} soll die Kovarianz mit C(X,Y) bezeichnet werden; entsprechend ist $V(X) = s_x^2$ und

$R(X,Y) = r_{xy}$. Dann gilt:

(7.21) $C(X,Y+Z) = C(X,Y) + C(X,Z)$
(7.22) $V(X+Y) = V(X) + V(Y) + 2C(X,Y) = C(X,X+Y) + C(Y,X+Y)$
(7.23) $C(X+Y,X-Y) = V(X) - V(Y)$

Danach kann X+Y und X-Y unkorreliert sein, obgleich X mit Y korreliert ist.
Aus $C(X,Y-X) = C(X,Y) - V(X)$ folgt, wenn $V(X) = V(Y)$

(7.24) $R(X,Y-X) = -\sqrt{½(1-r_{xy})} \leq 0$ und

(7.25) $R(X,Y+X) = \sqrt{½(1+r_{xy})} \geq 0.$

Beispiel 7.9

Auf ihrer fünftägigen Erkundung des noch wenig erforschten, bisher für unbewohnt gehaltenen, aber bereits gut kartographierten Planeten "Amar" (persisch: Statistik) maßen zwei Astronauten jeweils an drei Zeitpunkten täglich Längengrad (X) und Breitengrad (Y). Dabei erhielten sie die folgenden Messwerte:

Tag	X	Y
Montag	8	5
	9	4
	12	3
Dienstag	15	4
	16	5
	17	8
Mittwoch	16	11
	15	12
	12	13
Donnerstag	9	12
	8	11
	7	8

Zu ihrem großen Staunen korrelierten X und Y nicht mit r = +1 miteinander, obgleich die beiden Astronauten von einer vorgegebenen Linie nicht abgewichen sind. Wie ist das zu erklären?

Lösung 7.9:

Man kann leicht nachrechnen, dass die Korrelation zwischen X und Y den Wert Null annimmt. Es liegt ein funktionaler, aber kein linearer Zusammenhang vor. Eine graphische Darstellung zeigt, dass die Beobachtungen auf einem Kreis im x,y-Koordinatensystem liegen (mit dem Mittelpunkt [$\overline{x}$ = 12 und $\overline{y}$ = 8] sowie dem Radius 5), so dass der funktionale Zusammenhang $(x - 12)^2 + (y - 8)^2 = 5^2$ gegeben ist.

c) Scheinkorrelation

Zu den bekanntesten Fehlinterpretationen der Korrelation gehört die falsche (sachlich nicht gerechtfertigte) kausale Interpretation. Wenn X und Y miteinander korrelieren so kann dies bedeuten, dass:

- X die Ursache von Y ist,
- Y die Ursache von X ist, wobei dies vom ersten Fall mit dem Korrelationskoeffizienten nicht zu unterscheiden ist;
- X und Y rein zufällig in einer entsprechend kleinen Stichprobe miteinander korrelieren, in der Grundgesamtheit jedoch nicht (ein Aspekt der Induktiven Statistik, der in der Deskriptiven Statistik nicht zu behandeln ist);
- Messfehler bei der Beobachtung der Variablen auftreten (wären X und Y fehlerfrei beobachtet, dann würden sie nicht miteinander korrelieren);
- X und Y nur deshalb miteinander korrelieren, weil sie gemeinsam abhängig sind von einer dritten Variablen Z (Scheinkorrelation) und mit Z (nicht direkt unter einander) in einer Kausalbeziehung stehen.

Wegen dieser Nichteindeutigkeit ist die Meinung sehr verbreitet, dass Korrelation und Kausalität nichts miteinander zu tun hätten. Das bedeutet allerdings, das Kind mit dem Bade auszuschütten. Hierauf soll an späterer Stelle eingegangen werden (Abschn. 5).

Def. 7.9: Scheinkorrelation, spurious correlation

Sind zwei Variablen X und Y nur deshalb hoch korreliert, weil sie gemeinsam abhängig sind von einer dritten Variablen Z, so spricht man von Scheinkorrelation.

Bemerkungen zu Def. 7.9:

1. Das beliebteste Beispiel für eine Scheinkorrelation ist die Korrelation zwischen Störchennestern und Geburten. Jeder weiß, dass dies nicht kausal interpretiert werden kann, also kein direkter Kausalzusammenhang besteht.
   Die "dahinterstehende" Variable ist die Urbanisierung und wirtschaftliche Entwicklung, der Übergang von der Agrar- zur Industriegesellschaft, was zum einen dazu geführt hat, dass den Fröschen (und damit auch Störchen) der Lebensraum genommen wurde und zum anderen dazu, dass die Familiengrößen kleiner wurden. Da es in diesem Fall sehr offensichtlich ist, dass nur eine Scheinkorrelation und keine "echte" (d.h. kausal zu interpretierende) Korrelation vorliegt, spricht man auch von **nonsense correlation** (vgl. auch Beispiel 7.10).

Ein weiteres Beispiel ist die Korrelation zwischen der Anzahl der Feuerwehrlöschzüge (X) und der Größe des Brandschadens (Y). Der dritte Faktor (Z) ist die Größe des Brandes (z.B. die Flammenmenge). Die falsche, kausale Interpretation würde lauten: je mehr Feuerwehrlöschzüge bei einem Brand eingesetzt werden, desto größer ist der Brandschaden; also ist der Feuerwehreinsatz die Ursache des Brandschadens.

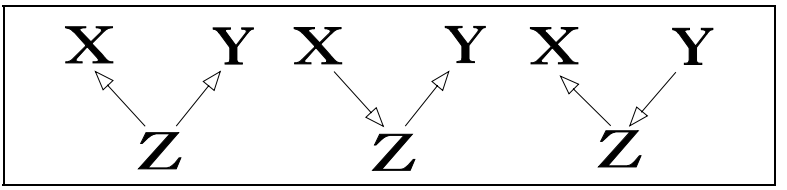
2. Es gibt aber auch viele Fälle einer Scheinkorrelation, bei denen es weniger offensichtlich ist, dass eine Kausalinterpretation nicht zulässig ist. Das ist sehr häufig bei der Korrelation von Zeitreihen, die einen gemeinsamen Trend haben, der Fall (die trendbereinigten Zeitreihen X^* und Y^* korrelieren dann weniger miteinander als die noch trendbehafteten Ursprungswerte X und Y). Sehr häufig tritt das in der Wirtschaftsstatistik bei der Korrelation mit Sozialproduktsgrößen oder allen wertmäßigen (und damit von der Inflation tangierten) Größen auf.
   Ein scherzhaftes Beispiel hierfür ist die hohe Korrelation zwischen den Preisen von kubanischem Rum und den Gehältern von amerikanischen Priestern. Die (in US-$ ausgedrückten) Rumpreise sind nicht gestiegen, weil die Priester so viel Rum nachfragten (eine Kausalinterpretation), sondern weil die Preise - wie die Gehälter - der Entwertung des US-$ unterlagen.

3. Häufig entsteht Scheinkorrelation auch durch Aggregation von Daten (Beispiel 7.11). Bei Disaggregation zeigt sich, dass sich die Korrelation verringert, d.h. dass sie bei Bezugnahme auf homogenere Gesamtheiten nicht gilt.

4. Bei nicht metrisch skalierten Merkmalen zeigt sich Scheinkorrelation durch ein Verschwinden (oder eine Verringerung) der Korrelation zwischen X und Y wenn der Einfluß der dritten Variablen Z "ausgeschaltet" wird d.h. der partiellen Korrelation (vgl. Kap. 8 u. 9).

5. Das Wirken einer dritten Variable Z geschieht bei Scheinkorrelation meist nach Art von Abb. 7.4 linkes Bild. Das Pfeilschema soll andeuten, dass X und Y gemeinsam "verursacht" werden von Z. Hinsichtlich der formalen Zusammenhänge zwischen den Korrelationskoeffizienten sind aber die drei Situationen der Abb. 7.4 nicht unterscheidbar. Auf Abb. 7.4 wird im abschließenden Abschn. 5 noch einmal eingegangen.

Abb. 7.4: Die drei Fälle von Scheinkorrelation zwischen X und Y (wobei jeweils gilt $r_{xy} = r_{xz}r_{zy}$)



Beispiel 7.10:

Während des italienischen Feldzuges im Zweiten Weltkrieg wurde eine positive Korrelation zwischen der Anzahl X der Propaganda-Flugblätter, die über den deutschen Linien abgeworfen wurden, und der Größe des von den Alliierten eroberten Gebietes (Y) bei einer bestimmten Stärke der Offensive (Z) festgestellt. Es sei

$r_{xy} = 0{,}8 \quad r_{xz} = 0{,}94 \quad$ und $\quad r_{yz} = 0{,}85$!

Beweist der hohe Wert von r_{xy}, dass man einen Krieg allein durch möglichst viele Propa-ganda-Flugblätter gewinnen kann?

Lösung 7.10:

Es besteht sicher kein direkter Kausalzusammenhang zwischen der Propaganda X und dem Kriegserfolg (Y). Die relativ hohe Korrelation von $r_{xy} = 0{,}8$ ist also nicht kausal im Sinne von $X \rightarrow Y$ oder gar $Y \rightarrow X$ zu deuten. Als dritter Faktor ist die Stärke der Offensive (Z) zu betrachten. Es ist wohl meist anzunehmen, dass je stärker die Offensive ist, desto eher ist sie erfolgreich ($Z \rightarrow Y$) und desto mehr ist sie auch mit entsprechenden Propagandafeldzügen verbunden ($Z \rightarrow X$). Für die Korrelation gilt in der Tat hier $r_{xy} = r_{xz}\, r_{zy} = 0{,}799$ (Die Zahlenangaben waren ja auch fiktiv). Das Ergebnis besagt, dass die partielle Korrelation $r_{xy.z}$ (vgl. Kap. 8) verschwindet.

Beispiel 7.11:

Gegeben seien die folgenden Daten über die Schuhgröße (X) und das Monatseinkommen in 1000 DM (Y):

x	35	36	37	38	41	42	43	44
y	2,8	2,5	2,0	2,5	4,5	5,5	5,2	3,8

Die Korrelation beträgt $r \approx 0{,}8$. Heisst dies, dass man deshalb mehr verdient, weil "man auf großem Fuß lebt"?

Lösung 7.11:

Es liegt ein typischer Fall von Scheinkorrelation vor. Angenommen, bei den ersten vier Personen handelt es sich um Frauen, die in der Regel eine kleinere Schuhgröße haben als Männer und häufig auch weniger verdienen. Die zweiten vier Personen seien Männer. Dann erhält man für die ersten vier Personen (also für die Frauen) für die Korrelation zwischen X und Y $r_{xy} = -0{,}54495$ und für die zweiten vier Personen (also die Männer)

r_{xy} = -0,40801, bei den beiden Gruppen zusammen aber r_{xy} = 0,8. Man beachte auch, dass sich das Vorzeichen ändert!

d) Bestimmtheitsmaß

Ist $y_v = z_v + u_v$, wobei Z irgendeine Funktion von X ist [allgemein $z_v = z(x_v)$], etwa eine lineare Regressionsfunktion $z_v = a + bx_v$ (s. Kap. 8), so gilt wegen Gl. 7.22 für die Varianz von Y:

$$s_v^2 = s_z^2 + s_u^2 + 2s_{zu}.$$

Verschwindet die Kovarianz s_{zu}, wie im Falle einer Regressions**geraden**, so lässt sich die Varianz von Y darstellen als Summe einer systematischen (durch X "erklärten") Varianz s_z^2 und einer (nicht erklärten) Residualvarianz s_u^2. Es gilt also bei einer mit der systematischen Komponente Z unkorrelierten Variable U stets die folgende **Varianzzerlegung**:

(7.26) $s_v^2 = s_z^2 + s_u^2$

Gesamtvarianz	erklärte Varianz	Residualvarianz

und nach Division durch s_v^2

$$(7.27)\quad 1 = \frac{s_z^2}{s_y^2} + \frac{s_u^2}{s_y^2} = B_{yx} + U_{yx},$$

worin B_{yx} die Bestimmtheit (von y durch x) und U_{yx} die Unbestimmtheit (von y durch x) darstellt.

Def. 7.10: Bestimmtheitsmaß

Der Ausdruck

$$(7.28)\quad B_{yx} = \frac{s_z^2}{s_y^2}$$

heißt Bestimmtheitsmaß (oder Bestimmtheit, coefficient of determination) und $U_{yx} = 1 - B_{yx}$ heißt Unbestimmtheit (oder coefficient of alienation). s_z^2 ist die erklärte Varianz von y und s_v^2 die Gesamtvarianz von y.

Bemerkungen zu Def. 7.10:

1. Da B_{yx} ein Varianzanteil ist, gilt $0 \le B_{yx} \le 1$. Die Bestimmtheit ist also nichtnegativ. Sie ist nicht notwendig symmetrisch, d.h. es muss nicht $B_{yx} = B_{xy}$ erfüllt sein.

2. Das Bestimmtheitsmaß liefert einen allgemeineren Zugang zur Messung des Zusammenhangs zweier metrisch skalierter Variablen als der Korrelationskoeffizient, der nur den Grad des **linearen** Zusammenhangs misst.

   Im Falle einer **linearen** Beziehung zwischen y und x gilt:

   1. Symmetrie: $B_{yx} = B_{xy}$
   2. Zusammenhang mit Korrelation: $B_{yx} = r^2_{xy}$,

   was bei nichtlinearer Beziehung nicht gewährleistet ist.

   Bei nichtlinearer Korrelation ist die Berechnung von r_{xy} (gem. Def. 7.8) wenig sinnvoll. Es kann aber $+\sqrt{B_{yx}}$ oder $+\sqrt{B_{xy}}$ als Korrelationskoeffizient berechnet werden. Diese Korrelation ist dann keine Produkt-Moment-Korrelation, sie kann nicht negativ sein und ist auch nicht notwendig symmetrisch. Ist der Zusammenhang nichtlinear, so macht es i.d.R. auch nicht viel Sinn zwischen positiver und negativer Korrelation zu unterscheiden.

3. Zum Begriff "**erklärte**" Varianz und zum Konzept der "Ursache" (X als Ursache für Y):
   "Erklärung" im Sinne der Statistik besteht stets in der Zerlegung der Varianz einer zu erklärenden Variable Y in einen Varianzanteil B_{yx}, der auf eine bekannte Variationsquelle X zurückgeführt werden kann und einen restlichen Anteil U_{yx}, der nicht auf eine explizit berücksichtigte, identifizierbare, "erklärende" Variationsquelle U zurückzuführen ist.

   Für diese "Erklärung" ist es nicht unwichtig, ob Beobachtungs- und Experimentdaten vorliegen: Nur im letzten Fall besteht a priori eine klare Unterscheidung, welche Einflußfaktoren zu X und welche zu U zu rechnen sind, weil eine Größe X bewußt und kontrolliert (bei Konstanz anderer Größen) variiert werden kann.

4. Im Unterschied zum Korrelationsverhältnis, das auf die Regressions**linie** Bezug nimmt, ist im Falle des Bestimmtheitsmaßes eine Regressions**funktion**, z.B. eine lineare Regressionsfunktion zu schätzen. Da dies erst im Kap. 8 gezeigt wird, soll hier kein Rechenbeispiel für das Bestimmtheitsmaß betrachtet werden.

e) Korrelationsverhältnis und Korrelationskoeffizient bei klassierter Verteilung

Es mag sein, dass bei Nichtlinearität des Zusammenhangs zwischen zwei metrisch skalierten Variablen X und Y der Produktmoment-Korrelationskoeffizient r_{xy} den Grad des Zusammenhangs unterschätzt und dass durch Berechnung eines nichtlinearen Korrelationskoeffizienten (vgl. Bem. Nr. 2 zu Def. 7.10) ein höherer Grad des Zusammenhangs festgestellt werden kann (wenn z.B. der Zusammenhang durch Annahme einer parabolischen statt linearen Beziehung besser erfaßt wird). Im Falle gruppierter oder klassierter Daten beschreibt die Regressionslinie den Zusammenhang zwischen X und Y in einer Weise, wie es durch keine (wie immer geartete) nichtlineare Regressionsfunktion besser geschehen könnte. Das Korrelationsverhältnis η(eta) misst die Güte der Anpassung der Regressions**linie** an die Daten und beruht auf der folgenden Varianzzerlegung.

Satz 7.5: Varianzzerlegung

Mit den m bedingten Mittelwerten $\bar{y}(x_i) = \bar{y}|x=x_i$ gem. Gl. 7.12 lassen sich die Abweichungen der n_i Beobachtungen in der i-ten Klasse bezüglich des Merkmals X (bzw. in den Fällen, in denen $X = x_i$ ist) vom Gesamtmittelwert $\bar{y}$ darstellen als

$$y_{pi} - \bar{y} = y_{pi} - \bar{y}(x_i) + \bar{y}(x_i) - \bar{y}$$

mit i = 1,2,...,m Ausprägungen oder Klassen bezüglich X und p = 1,2,...,n_i Beobachtungen in der i-ten Klasse (wenn X die Ausprägung x_i hat).

Damit gilt auch

$$(7.29)\quad \Sigma\Sigma y_{pi} - \bar{y} = \Sigma\Sigma[y_{pi} - \bar{y}(x_i)] + \Sigma n_{i.}[\bar{y}(x_i) - \bar{y}].$$

Summierung über alle m Ausprägungen oder Klassen der Variablen X und Quadrierungen der Abweichungen liefert

$$(7.30)\quad \Sigma\Sigma(y_{pi} - \bar{y})^2 = \Sigma n_{i.} s_y^2(x_i) + \Sigma n_i[\bar{y}(x_i) - \bar{y}]^2$$
$$SAQ_{tot} = SAQ_{int} + SAQ_{ext}$$

wobei die interne Varianz $SAQ_{int}/n = \Sigma h_{i.} s_y^2(x_i)$ ein gewogenes Mittel der bedingten Varianzen (gem. Gl. 7.12a) darstellt.

In Gl. 7.30 ist SAQ die Summe der Abweichungsquadrate. Die gesamte Summe der Abweichungsquadrate (SAQ_{tot}) lässt sich also in eine interne

und eine externe Summe der Abweichungsquadrate (SAQ_{int} und SAQ_{ext}) zerlegen. Dividiert man die linke und rechte Seite von Gl. 7.30 durch n, so erhält man die Varianzzerlegung:

$$(7.30a) \quad V(y_{tot}) = V(y_{int}) + V(y_{ext}) \qquad (V = SAQ/n = \text{Varianz}).$$

Beweis: Zu zeigen ist allein der Übergang von Gl. 7.29 zu Gl. 7.30. Er ergibt sich durch Ausmultiplizieren unter Berücksichtigung von $s_y^2(x_i) = n_i^{-1} \Sigma y_{pi}^2 - [\bar{y}(x_i)]^2$ und aus der Schwerpunkteigenschaft des arithmetischen Mittels. Die Zerlegung der Varianz (Gl. 7.30a) bzw. der Summe der Abweichungsquadrate (SAQ, Gl. 7.30) wird auch in den Beispielen 7.12 und 7.13 demonstriert.

Def. 7.11: Korrelationsverhältnis

Das Korrelationsverhältnis η (eta) ist der Ausdruck

$$(7.31) \quad \eta_{yx} = +\sqrt{\frac{SAQ_{ext}}{SAQ_{tot}}}$$

mit den Summen der Abweichungsquadrate SAQ gem. Gl. 7.30.

Bemerkungen zu Def. 7.11:

1. Das Korrelationsverhältnis ist nicht notwendig symmetrisch, d.h. in der Regel sind η_{yx} und η_{xy} nicht gleich (vgl. Beispiele 7.12 und 7.13). Auch kann mit η nicht zwischen positiver und negativer Korrelation unterschieden werden, denn η ist (wie das Bestimmtheitsmaß) aus einem Varianzanteil abgeleitet: η_{yx}^2 ist der Anteil der durch die Regres-sionslinie erklärten Varianz von Y und entsprechend wird in η_{xy}^2 eine Zerlegung der Varianz von X vorgenommen.

2. η ist i.d.R. verschieden vom (linearen) Korrelationskoeffizienten r_{xy} bei gruppierten (bzw. analog mit $\bar{x}_i$, $\bar{y}_j$ klassierten) Daten:

$$(7.32) \quad r_{xy} = \frac{n \Sigma\Sigma x_i y_i n_{ij} - (\Sigma x_i n_{i.})(\Sigma y_j n_{.j})}{\sqrt{[\Sigma(x_i-\bar{x})^2 n_{i.}]\,[\Sigma(y_j-\bar{y})^2 n_{.j}]}}$$

r_{xy}: linearer Korrelationskoeffizient bei gruppierten bzw. klassierten Daten

Im Zähler von Gl. 7.32 steht die n^2-fache Kovarianz und im Nenner die Wurzel aus dem Produkt der n-fachen Varianzen von X und Y. Die Berechnung des Korrelationskoeffizienten wird in Bsp. 7.12 und 7.13 demonstriert.

Bei klassierten Daten ist anstelle von x_i und y_j der jeweilige Klassenmittelwert der i-ten Klasse bezüglich X, bzw. der j-ten Klasse bezüglich Y einzusetzen.

2. Das Korrelationsverhältnis ist nicht unabhängig von der Anzahl der Klassen bzw. "Gruppen" (unterschiedliche Werte x_i bzw. y_j). Werden nur wenige Klassen unterschieden z.B. aus Gründen der Rechenvereinfachung im Beispiel 7.14, so ist die Berechnung von η wenig sinnvoll. Andererseits gilt: Bei vielen schwach besetzten Klassen können die bedingten Mittelwerte stark schwanken, weil in jeder Klasse nur wenige Beobachtungen vorliegen.

Beispiel 7.12:

Man verifiziere die Varianzzerlegung (Gl. 7.30) und berechne die Korrelationsverhältnisse und den Korrelationskoeffizienten r_{DM} für das Beispiel 7.2!

Lösung 7.12:

Es soll der Einfachheit halber nur von Dur (D) und Moll (M) gesprochen werden.

Daten zur Regressionslinie zur Schätzung von Dur:

wenn M	n_j	bed.Mittelw.	bedingte Varianz von Dur (D)
10	1	30	0 da $n_{.1} = 1$
20	2	55	½ $[(40-55)^2 + (70-55)^2] = 225$
30	4	35	¼ $[(20-35)^2 + (30-35)^2 +$ $(40-35)^2 + (50-35)^2] = 125$
40	2	35	½ $[(10-35)^2 + (60-35)^2] = 625$
50	1	50	0 da $n_{.5} = 1$

Man erkennt an der Folge der bedingten arithmetischen Mittelwerte, dass kein *linearer* Zusammenhang besteht (es ist auch $r_{DM}=0$).

Berechnung der Summe der Quadrate der Abweichungen SAQ für Dur:
da das Gesamtmittel $\bar{D} = 40$ ist

$SAQ_{tot} = (10-40)^2 + (20-40)^2 + 2(30-40)^2 + 2(40-40)^2 + 2(50-40)^2 + (60-40)^2 + (70-40)^2 = 3000$ (so dass wegen n = 10 für die Varianz gilt $s_D^2 = 300$).

$SAQ_{int} = 2\cdot225 + 4\cdot125 + 2\cdot625 = 2200;$

$SAQ_{ext} = (30-40)^2 + 2(55-40)^2 + 6(35-40)^2 + (50-40)^2 = 800.$

Man sieht, dass gilt $SAQ_{tot} = SAQ_{int} + SAQ_{ext} = 3000 = 2200 + 800$.

Für das Korrelationsverhältnis erhält man dann $\eta_{DM}^2 = SAQ_{ext}/SAQ_{tot} = 800/3000 = 4/15 = 0{,}267$ und folglich ist $\eta_{DM} = \sqrt{4/15} = 0{,}5164$.

Entsprechend erhält man mit der Regressionslinie zur Schätzung von Moll:

$SAQ_{tot} = 1200 \quad (s_M^2 = = 120)$

$SAQ_{int} = 2\cdot100 + 2\cdot25 + 2\cdot100 = 450$

$SAQ_{ext} = 100 + 200 + 50 + 200 + 100 + 100 = 750$

Es gilt wieder $SAQ_{tot} = SAQ_{int} + SAQ_{ext} = 1200 = 450 + 750$.

Korrelationsverhältnis ($\eta^2_{MD} = SAQ_{ext}/SAQ_{tot} = 750/1200 = 0{,}625$ und somit $\eta_{MD} = 0{,}79057$ (was ungleich $\eta_{DM} = 0{,}5164$ ist).

Beispiel 7.13:

(Die emanzipierte Fassung von Aufg. 7.2/7.12: Das Experiment einer Dreierbeziehung und ein neuerliches Beispiel für konkrete Lebenshilfe durch Statistik)

Nachdem Andrea (A) zwei Jahre mit Charlie (C) ging, haben sie sich `ne echt besitzhafte Identität aufgebaut, aus der sich A nun emanzipieren will. Sie ist jetzt mehr so auf Bernd (B) drauf, kann aber noch nicht total auf B einflippen. Und weil ihr bisheriger Typ C die Trennungsverarbeitung erst einmal konkret abgecheckt haben will und das, was zwischen A und B so läuft emotional noch nicht so auffangen kann, haben sie jetzt alle drei beschlossen, das Problem bis spätestens zum nächsten Jahr zu dritt ganz konkret aufzuarbeiten. In ihrer total fixierenden Art, mit der sie mit jeder Beziehungskiste umgeht hat A die folgenden Aufzeichnungen gemacht über die Tage, die sie im Monat mit B bzw. C verbracht hatte:

	Tage mit C			
Tage mit B	0-10	10-20	>20	Σ
0-10	0	2	4	6
10-20	1	2	0	3
>20	3	0	0	3
Σ	4	4	4	12

Man verifiziere die Varianzzerlegung (Gl. 7.30) und berechne die Korrelationsverhältnisse sowie den Korrelationskoeffizient r_{BC}! Als Klassenmittelwerte sind die Zahlen 5, 15 und 25 anzusetzen.

Lösung 7.13:

Während Beispiel 7.12 die Berechnung des Korrelationsverhältnisses bei gruppierten Daten demonstrieren soll, gilt es hier, Varianzzerlegung und Berechnung des Korrelatinsverhältnisses sowie des Korrelationskoeffizienten bei einer zweidimensionalen klassierten Verteilung zu zeigen.

Parameter der Randverteilungen

Bernd	Charlie
Mittel $\bar{B} = 12{,}5$	Mittel $\bar{C} = 15$
Varianz $s^2_B = 825/12$	Varianz $s^2_C = 800/12$,

so dass SAQ_{tot} 825 bzw. 800 ist.

Regressionslinie Bernd

wenn C	$n_{.j}$	bed.Mittelwert	bedingte Varianz der Regressionslinie v. Bernd
5	4	22,5	$\frac{1}{4}[(15-22,5)^2+ 3(25 - 22,5)^2] = 75/4$
15	4	10	$\frac{1}{4}[2(5-10)^2+2(5 -10)^2] = \frac{1}{4}100 = 25$
25	4	5	0 da alle 4 Beobachtungen gleich sind

Regressionslinie Charlie

wenn B	$n_{i.}$	bed. Mittelw.	bed. Varianz
5	6	130/6	133,33/6
15	3	35/3	22,222
25	3	5	0

$SAQ_{int} = 75 + 100 = 175$ $SAQ_{int} = 133,33 + 66,67 = 200$

$SAQ_{ext} = 4(22,5-12,5)^2+4(10-12,5)^2+4(5-12,5)^2 = 650$ $SAQ_{ext} = 600$ (SAQ_{tot}=800)

$SAQ_{tot} = 825 = SAQ_{ext} + SAQ_{int} = 650 + 175.$

Berechnung der Korrelationsverhältnisse:

Bernd: $\eta_{BC} = \sqrt{SAQ_{ext}/SAQ_{tot}} = \sqrt{650/825} = 0,8876$

Charlie: $\eta_{CB} = \sqrt{600/800} = \sqrt{3/4} = 0,8660$

Die beiden Korrelationsverhältnisse sind nicht gleich. Man kann an ihnen auch nicht erkennen dass eigentlich eine negative Korrelation besteht.

Berechnung des Korrelationskoeffizienten:

Kovarianz s_{BC}: $(5\cdot5\cdot0 + 5\cdot15\cdot2 + 5\cdot25\cdot4 + 15\cdot5\cdot1 + 15\cdot15\cdot2 + 15\cdot25\cdot0 + 25\cdot5\cdot3 + 25\cdot25\cdot0 + 25\cdot25\cdot0)/12 - 12,5\cdot15 = -700/12$ und da die Varianzen $s_B,^2 = 825/12$ und $s_C,^2 = 800/12$ betragen erhält man $r_{BC} = s_{BC}/s_B s_C = - 0,86164$.

4. Zusammenhang bei nicht metrisch skalierten Variablen

a) Maße des Zusammenhangs und Skalenniveaus (Übersicht)

In Abhängigkeit von dem Skalenniveau der Variablen X und Y gibt es zahlreiche Maße für den Grad des Zusammenhangs zwischen X und Y (vgl. Übers. 7.2). Man kann in allen Fällen auch von Korrelationen im weiteren Sinne sprechen und die bisher behandelte Korrelation als "**Maßkorrelation**" (Korrelation im engeren Sinne) bezeichnen. Im englischen Sprachgebrauch ist auch "association" ein entsprechender Oberbegriff, während Assoziation im engeren Sinne nur den Zusammenhang zweier dichotomer Merkmale bezeichnet. Auf einige der in Übersicht 7.2 genannten Maße des Zusammenhangs wird in den folgenden Abschnitten eingegangen.

Übersicht 7.2:
Maße des Zusammenhangs zwischen den Merkmalen X und Y

a) Fallunterscheidung nach dem Skalentyp der beiden Variablen

	Variable X		
Var. Y	nominal	ordinal	metrisch
nominal	3*)	5	4
ordinal	5	2*)(2a/2b)	
metrisch	4		1

*) weitere Fallunterscheidung unter b)

b) Fälle im einzelnen und Maßzahlen

1. Beide Variablen sind metrisch skaliert (*Fall 1*): **Maßkorrelation**
   - Produkt-Moment-Korrelationskoeffizient (Def. 7.8)
   - Korrelationsverhältnis (Def. 7.11)
   - Gelegentlich wird auch auf den älteren, kaum noch gebräuchlichen Korrelationskoeffizient r_F von Fechner in diesem Zusammenhang verwiesen (vgl. Bem. 3 zu Def. 7.7). Da es aber bei ihm nur auf die Vorzeichen der Abweichungen vom Schwerpunkt $(\overline{x},\overline{y})$ bzw. vom Medianpunkt $(\tilde{x}_{0,5}, \tilde{y}_{0,5})$ ankommt, paßt r_F eher zur Situation 2b.

2. Die ordinale Abstufung (*Fall 2*) kann
   2a) durch Rangplätze beschrieben werden (**Rangkorrelation**).
   2b) ohne Rangplätze bestehen:
   bei beiden Merkmalen X und Y werden die Ausprägungen allein ordinal unterschieden, etwa $x_1 < x_2 < x_3$ usw.; sie werden nicht mit [wie immer gefundenen] Zahlenwerten (z.B. Rangplätzen) codiert (**Rangassoziation** oder ordinale Assoziation [order association]).

   Bem.: Im Falle 2a) sind zwar nicht die Meßwerte x_1, x_2 usw. (Merkmalsausprägungen) metrisch skaliert, wohl aber die hierfür verwendeten Rangplätze.

3. Weitere Fallunterscheidung (*Fälle 3 und 4*)

   Nominalskala mit
   p > 2 Ausprägungen: **Polytomie (P)**,
   p = 2 Ausprägungen: **Dichotomie (D)**.
   Speziell im Falle einer Dichotomie kann man unterscheiden:

   D1: Echte Dichotomie (qualitative Unterscheidung),
   D2: Dichotomie mit einer zugrundeliegenden Rangordnung (z.B. Unterscheidung von gut/schlecht, positiv/negativ usw.),

D3: Dichotomie bei an sich zugrundeliegender Normalverteilung.

Y	X-Skala				
Skala	D1/D2	D3	P	M*)	O*)
D1/D2	A	*	K	PB	RB
D3	*	T	*	B	
P	K	*	K		

*) M = metrisch-skaliert O = ordinal-skaliert

A = Assoziationsmaße (Vierfelderkorrelation)
T = Tetrachorische Korrelation (tetrachoric correlation)
K = Kontingenzmaße
B = Biserielle Korrelation (biserial correlation, analog z.B. triseriell wenn Y als Trichotomie vorliegt)
PB = Punkt-biserielle-Korrelation (point biserial correlation)
RB = Rang-biserieller Korrelationskoeffizient.

b) Assoziation und Kontingenz

1. Allgemeines

Konstruktionsprinzipien für Zusammenhangsmaße

Man kann Assoziations- und Kontingenzmaße (Übersicht 7.2) konstruieren aufgrund folgender Überlegungen:

a) *Auf χ^2 basierende Maße*: Vergleich der Häufigkeiten n_{ij} für die Kombination (x_i,y_j) der Merkmale X und Y mit den zu erwartenden Häufigkeiten bei Unabhängigkeit [Def. 7.5] (sind sie gleich, so liegt kein Zusammenhang vor).
b) *Prädikationsmaße*: Liegt ein Zusammenhang zwischen X und Y vor (also keine Unabhängigkeit), so ist bei Kenntnis der Verteilung von X die Verteilung von Y (bzw. umgekehrt) "besser" vorauszusagen als ohne Kenntnis.
c) Man kann die *Häufigkeit konkodanter und diskonkodanter Merkmalskombinationen* vergleichen, was bei Nominalskalen ohne jede dahinterstehende Rangordnung nicht sinnvoll ist, wohl aber z.B. bei Dichotomien im Sinne von D2/D3 der Übers. 7.2 (also in der Assoziationsanalyse).
d) Speziell in der Assoziationsanalyse: Berechnung von r_{xy}, der Produkt-Moment-Korrelation (Def. 7.8) für mit 0 und 1 codierte Variablen X und Y ("Vierfelderkorrelation" [vgl. Def. 7.16]).

Normierungsprobleme

Während die Untergrenze aller Maße des Zusammenhangs eindeutig ist und durch Unabhängigkeit (Def. 7.5) gegeben ist, macht es Schwierigkeiten, einen "maximalen" Zusammenhang zu definieren und so Kontingenz- und Assoziationsmaße auf den Wertebereich von 0 bis 1 oder -1 bis +1 zu normieren.
Abhängigkeit (Unabhängigkeit) besteht in der Unterschiedlichkeit (Gleichheit) der bedingten Verteilungen. Aber:

- eine maximale Unterschiedlichkeit ist nicht eindeutig definiert;
- während Unabhängigkeit eine symmetrische Eigenschaft ist, muss dies für die Abhängigkeit nicht gelten.

Axiomatik

Für ein Assoziations- oder Kontingenzmaß (AK) sollte gelten

A1/K1 Das Maß AK sollte dann und nur dann den Wert AK = 0 annehmen, wenn die beiden Variablen unabhängig sind.

A2/K2 Bei einer genau definierten "maximalen Abhängigkeit" sollte AK die Obergrenze AK = +1 annehmen. Diese Obergrenze ist im Falle der Assoziation, nicht aber in dem der Kontingenz eindeutig definiert.

A3/K3 Das Maß AK sollte nicht invariant sein gegenüber einer Ver-k-fachung von einzelnen Zeilen oder Spalten.

A4/K4 Das Maß AK sollte von der Gesamtzahl n der Beobachtungen unabhängig sein.

Durch eine Ver-k-fachung der Häufigkeiten einer einzelnen Zeile (oder Spalte) ändert sich die entsprechende bedingte Verteilung nicht, wohl aber die Abhängigkeit zwischen den Merkmalen, was im Fall der Assoziation durch Hinweis auf die Regressionsanalyse gezeigt werden wird. Es gibt Maße, die invariant sind gegenüber einer solchen Veränderung der Häufigkeiten. Im Unterschied hierzu soll eine Ver-k-fachung **aller** Häufigkeiten nach A4/K4 das Maß AK nicht verändern.

Beispiel/Lösung 7.14:

Man könnte von "vollständiger Abhängigkeit" der Variablen Y von der Variablen X (oder von einem "maximalen Zusammenhang" zwischen den Variablen X und Y) sprechen, wenn aus der Verteilung von X eindeutig die Verteilung von Y hervorgeht und umgekehrt, wie dies bei der folgenden Tafel 1 der Fall ist. Es ist klar, dass dieses Konzept eines "maximalen Zusammenhangs" nur Sinn macht bei quadratischen Kontingenztafeln. Da für X und Y nur Nominalskalen vorausgesetzt werden, müßte der Grad der Abhängigkeit gleich bleiben wenn Zeilen und Spalten permutiert oder auch

zusammengefasst werden. Es müßten also die folgenden vier Tafeln jeweils die gleiche Kontingenz (als Grad der Abhängigkeit) aufweisen:

Tafel 1

	y_1	y_2	y_3	Σ
x_1	20	0	0	20
x_2	0	30	0	30
x_3	0	0	50	50
Σ	20	30	50	100

Tafel 2

	y_2	y_1	y_3	Σ
x_3	0	0	50	50
x_2	30	0	0	30
x_1	0	20	0	20
Σ	30	20	50	100

Kontingenzmaße betrachten i.d.R. die Tafeln 1 und 2 als gleichwertig und sie nehmen jeweils ihren maximalen Wert 1 an, nicht dagegen in den Fälle der Tafel 3 und 4.

Tafel 3

	y_1+y_3	y_2	Σ
x_3	50	0	50
x_1	20	0	20
x_2	0	30	30
Σ	70	30	100

Tafel 4

	y_1+y_3	y_2	Σ
x_2+x_3	50	30	80
x_1	20	0	20
Σ	70	30	100

3. Kontingenzmaße

a) auf den Vergleich mit der Unabhängigkeit basierende Maße

Sind die Merkmale X und Y Polytomien, also Nominalskalen mit jeweils zwei und mehr Ausprägungen, so kann man die Größe Chi-Quadrat (χ^2) berechnen. Sie beruht auf einem Vergleich der beobachteten Häufigkeiten n_{ij} mit den (bei Unabhängigkeit) zu erwartenden Häufigkeiten f_{ij}.

Def. 7.12: Chi-Quadrat

Mit den beobachteten Häufigkeiten n_{ij} ($n = \Sigma\Sigma n_{ij}$) und den bei Unabhängigkeit [Def. 7.5] zu erwartenden Häufigkeiten f_{ij} ist die Größe Chi-Quadrat (χ^2) definiert als

(7.33) $\chi^2 = \Sigma\Sigma(n_{ij} - f_{ij})^2/f_{ij}$ mit $f_{ij} = n_{i.}n_{.j}/n$ $\quad (i = 1,2,...,r \text{ und } j = 1,2,...,c)$

Bemerkungen zu Def. 7.12:

1. Die Größe χ^2 selber ist nicht als Kontingenzmaß geeignet, weil sie direkt von der Anzahl der Beobachtungen abhängt. Eine Ver-k-fachung aller Häufigkeiten führt auch zu einem k-fachen Wert von χ^2.

2. Einfluß hat auch die Anzahl r der Zeilen (rows) und c der Spalten (columns) der zugrundeliegenden Kontingenztafel (i = 1,2,...,r und j = 1,2,...,c).

3. Die Punkte 1 und 2 haben die Konstruktion einiger auf χ^2 basierender Kontingenzmaße angeregt (Def. 7.13), die wegen χ^2 stichprobentheoretisch gewisse Vorteile haben, andererseits aber kaum anschaulich interpretierbar sind. Es gibt deshalb auch Kontingenzmaße, die einem anderen Konzept folgen (Def. 7.14).

4. Aus der Definition von χ^2 (Gl. 7.33) folgt, dass χ^2 bei Unabhängigkeit den Wert Null annimmt und in allen anderen Fällen größer als Null ist.

5. Die Quadrierung in der Größe χ^2 ist damit zu motivieren, dass die Zeilen- und Spaltensumme der einfachen Abweichungen jeweils verschwindet, denn:

   (7.33a) $$\sum_i (n_{ij} - f_{ij}) = \sum_j (n_{ij} - f_{ij}) = \sum_i \sum_j (n_{ij} - f_{ij}) = 0$$

   Der Zusammenhang wird im Beispiel 7.15 demonstriert.

Def. 7.13: auf Chi-Quadrat beruhende Kontingenzmaße

(7.34) $\phi = \sqrt{c^2/n}$ Phi-Koeffizient

$\phi^2 = \chi^2/n$ heisst auch mittlere quadratische Kontingenz

(7.35) $C = \sqrt{c^2/(c^2 + n)}$ Kontingenzmaß von Pearson

so dass $C^2 = \phi^2/(\phi^2+1)$

(das ist das unnormierte Kontingenzmaß von Pearson, das maximal den Wert $C_{max} = \sqrt{(m-1)/m}$ mit m = min(r,c) annimmt; der auf den Wertebereich [0,1] normierte Kontingenzkoeffizient von Pearson lautet:

(7.35a) $$C^* = \frac{C}{C_{max}} = \sqrt{\frac{mc^2}{(m-1)(c^2 + n)}} \qquad 0 \le C^* \le 1$$

(7.36) $T^2 = \chi^2/[n\cdot\sqrt{(r-1)(c-1)}]$ T: Kontingenzmaß von Tschuprow

(7.37) $V^2 = \chi^2/[n \cdot \min(r-1,c-1)]$ V: Kontingenzmaß von Cramer

Folgerungen:

- bei einer quadratischen Tabelle (r=c) ist T = V
- bei einer Vierfeldertafel [Def. 7.15], d.h. bei r = c = 2 (Assoziation) ist T = V = ϕ und $C^* = \sqrt{2f^2/(f^2+1)}$ mit $0 \le C^* \le 1$.

Beispiel 7.15:

Man bestimme die Kontingenzmaße der Def. 7.13 für die vier Tafeln von Bsp. 7.14!

Lösung 7.15:

Die Bestimmung der Tafel bei Unabhängigkeit gem. Def. 7.5 (auch "Indifferenztabelle" genannt) und damit der Größe χ^2 wird für Tafel 1 ausführlich gezeigt:

empirische (beobachtete) Häufigkeiten n_{ij}

	y_1	y_2	y_3	Σ
x_1	20	0	0	20
x_2	0	30	0	30
x_3	0	0	50	50
Σ	20	30	50	100

erwartete (bei Unabhängigkeit) Häufigkeiten f_{ij}

	y_1	y_2	y_3	Σ
x_1	4	6	10	20
x_2	6	9	15	30
x_3	10	15	25	50
Σ	20	30	50	100

Damit erhält man die Abweichungen $n_{ij} - f_{ij}$

	y_1	y_2	y_3	Σ
x_1	16	-6	-10	0
x_2	-6	21	-15	0
x_3	-10	-15	25	0
Σ	0	0	0	0

(womit auch Gl. 7.33a verifiziert ist) χ^2 ergibt sich nun durch Summierung der Größen $(n_{ij} - f_{ij})^2/f_{ij}$ über alle Zeilen und Spalten.

Für Tafel 1 ist also $\chi^2 = 200$. Mit n = 100 und r=c=3 erhält man außerdem $\phi = \sqrt{2}$, T = V = 1 ferner $C = C_{max} = \sqrt{2/3}$.
Für Tafel 2 erhält man die gleichen Ergebnisse wie für Tafel 1 und für Tafel 3 ergibt sich $\chi^2 = 54{,}857$, $\phi = V = 0{,}74065$, $T = 0{,}62282$, $C = 0{,}59518$.
Für Tafel 4 ist $\chi^2 = 10{,}7143$ und $\phi = T = V = 0{,}32733$ und $C = 0{,}31109$.

b) Prädikationsmaße der Kontingenz (Konzept der Fehlerreduktion)

Eine Alternative zu der wenig anschaulichen Größe χ^2 ist eine Klasse von Maßzahlen, die auf dem Konzept der Fehlerreduktion beruhen und von Leo A. Goodman und William H. Kruskal (1954) in die Diskussion gebracht wurden. Danach sind X und Y dann "korreliert", wenn es gelingt, bei Kenntnis des Merkmalswertes x_v der v-ten Einheit, deren Wert y_v besser (mit geringerem Fehler, mit größerer Treffsicherheit) vorherzusagen als ohne Kenntnis von x_v. Aus dieser Definition des "Zusammenhangs" zwischen X und Y, bzw. der Abhängigkeit der Variablen Y von X

> Y hängt von X in dem Maße ab, in dem Kenntnis über X die Unsicherheit über Y reduziert

folgt auch, dass Zusammenhang nicht notwendig eine symmetrische Relation ist. Die Maße von Goodman und Kruskal nach dem Konzept der pro-

portionalen (=relativen) Fehlerreduktion (proportional reduction in error PRE) sind **asymmetrische Kontingenzmaße**. Sie setzen voraus:

1. ein Konzept (ein Maß) für den **Fehler** der Vorhersage unter Berücksichtigung des Skalenniveaus;
2. eine (i.d.R. für beide Fälle a und b identische) **Vorhersageregel** für die Vorhersage von Y durch X
   a) *ohne* Kenntnis von X (aus der Randverteilung von Y)
   b) aufgrund der (durch X) bedingten Verteilungen von Y (also *mit* Kenntnis von X) und
3. ein Maß für das Konzept der "**Fehlerreduktion**".

Die Konstruktion von Kontingenzmaßen aufgrund bestimmter Festlegungen zu diesen drei Punkten soll anhand des folgenden Beispiels demonstriert werden. Dargestellt werden die Koeffizienten

- *lambda (**λ**)* der *optimalen* (modalen) Vorhersage und
- *tau (**τ**)* der *proportionalen* Vorhersage

von Goodman und Kruskal, wobei sich λ und τ nur durch die Vorhersageregel (Punkt 2a und 2b) unterscheiden.

Beispiel 7.16:

Entwicklung und Demonstration der Maße von Def. 7.14
Anhand der folgenden 3x4 Kontingenztafel (mit absoluten Häufigkeiten n_{ij}) sollen die Maße λ und τ hergeleitet werden.

	y_1	y_2	y_3	y_4	Σ
x_1	9	1	2	13	25
x_2	6	19	6	15	46
x_3	6	5	10	8	29
Σ	21	25	18	36	100

Die Zahlen sind so gewählt (n=100), dass eine Umrechnung in relative Häufigkeiten sehr einfach ist.

Zu 1 und 2a:

Vorhersage von Y **ohne** Kenntnis von X (Vorhersageregel/-fehler)

a) optimale (modale) Vorhersage [λ]
   100 Einheiten, deren y-Wert unbekannt ist werden vollständig der häufigsten (modalen) Ausprägung y_4 zugeordnet: damit werden 36 Einheiten richtig und 64 Einheiten falsch zugeordnet. Der Vorhersagefehler ist somit:
   $E_1 = 1 - p_{.4} = 1 - \max_j p_{.j} = 0{,}64$.

b) proportionale Vorhersage [τ]

100 Einheiten, deren y-Wert unbekannt ist werden proportional zu den Häufigkeiten der Randverteilung zugeordnet, also 21 in y_1, 25 in y_2 usw. Der Vorhersagefehler ist somit:
$0{,}21 \cdot 0{,}79 + 0{,}25 \cdot 0{,}75 + 0{,}18 \cdot 0{,}82 + 0{,}36 \cdot 0{,}64 = 0{,}7314$
$F_1 = \Sigma p_{.j}(1 - p_{.j}) = 1 - \Sigma(p_{.j})^2 = 0{,}7314.$

Zu 2b:

Vorhersage von Y **mit** Kenntnis von X, also aufgrund der bedingten Verteilungen von Y (Vorhersageregel/-fehler)

a) optimale (modale) Vorhersage [λ]
Die Einheiten werden jeweils vollständig der bei gegebenem Wert von X häufigsten Ausprägung von Y zugeordnet. Der Vorhersagefehler ist somit:
bei $X = x_1$: 0,25 - 0,13 = 0,12,
bei $X = x_2$: 0,46 - 0,19 = 0,27 usw. insgesamt also
$E_2 = \Sigma_i [p_{i.} - \max_j (p_{ij})] = 1 - \Sigma_i \max_j (p_{ij}) = 0{,}58.$

b) proportionale Vorhersage [τ]
Die Einheiten werden proportional zu den Häufigkeiten der bedingten Verteilungen zugeordnet. Der Vorhersagefehler ist dann: bei $X = x_1$:
$9/25 \cdot 16/25 + 1/25 \cdot 24/25 + 2/25 \cdot 23/25 + 13/25 \cdot 12/25 = 0{,}592$
oder allgemein:
$f_1 = \sum_j (p_{1j}/p_{1.})(1 - p_{1j}/p_{1.}) = 1 - \sum_j (p_{1j}/p_{1.})^2,$

entsprechend bei $X = x_2$: $f_2 = 1 - \sum_j (p_{2j}/p_{2.})^2 = 0{,}689$ usw.

($f_3 = 0{,}7325$)

Der gesamte Fehler ist dann:

$F_2 = \Sigma_i p_{i.} - f_i = \Sigma_i p_{i.}[1 - \Sigma_j (p_{ij}/p_{i.})^2] = 1 - \Sigma_j (p_{ij})^2/p_i)$.

(im Beispiel $F_2 = 0{,}67737$).

Zu 3:

Konzept der proportionalen (relativen) Fehlerreduktion
a) optimale Vorhersage [λ] b) proportionale Vorhersage [τ]

Ergebnis:

λ_{xy}	= (E1 - E2)/E1	τ_{xy}	= (F1 - F2)/F1
	= (0,64 - 0,58)/0,64		= (0,7314 - 0,67737)/0,7314
	= 0,09375		= 0,07387

Def. 7.14: Kontingenzmaße von Goodman-Kruskal

Nach Umformungen erhält man für die asymmetrischen Kontingenzmaße λ und τ mit den relativen Häufigkeiten $p_{ij} = n_{ij}/n$ und mit den absoluten Häufigkeiten

a) Koeffizient der optimalen Vorhersage λ

aa) Vorhersage von Y durch X

$$(7.38) \qquad \lambda_{xy} = \frac{\Sigma_i \max_j(p_{ij}) - \max_j(p_{\cdot j})}{1 - \max_j(p_{\cdot j})} = \frac{\Sigma_i \max_j(n_{ij}) - \max_j(n_{\cdot j})}{n - \max_j(n_{\cdot j})}$$

ab) Vorhersage von X durch Y

$$(7.38a) \qquad \lambda_{yx} = \frac{\Sigma_j \max_i(p_{ij}) - \max_i(p_{i\cdot})}{1 - \max_i(p_{i\cdot})} = \frac{\Sigma_j \max_i(n_{ij}) - \max_i(n_{i\cdot})}{n - \max_i(n_{i\cdot})}$$

b) Koeffizient der proportionalen Vorhersage τ

ba) Vorhersage von Y durch X

$$(7.39) \tau_{xy} = \frac{\Sigma\Sigma p_{ij}^2 / p_{i\cdot} - \Sigma p_{\cdot j}^2}{1 - \Sigma p_{\cdot j}^2} = \frac{n\Sigma\Sigma n_{ij}^2 / n_{i\cdot} - \Sigma n_{\cdot j}^2}{n^2 - \Sigma n_{\cdot j}^2}$$

bb) Vorhersage von X durch Y

$$(7.39a) \tau_{yx} = \frac{\Sigma\Sigma p_{ij}^2 / p_{\cdot j} - \Sigma p_{i\cdot}^2}{1 - \Sigma p_{i\cdot}^2} = \frac{n\Sigma\Sigma n_{ij}^2 / n_{\cdot j} - \Sigma n_{i\cdot}^2}{n^2 - \Sigma n_{i\cdot}^2}$$

Bemerkungen zu Def. 7.14:

1. Es ist offensichtlich, dass λ und τ asymmetrisch sind, also die Abhängigkeit nicht vertauscht werden darf.

   Im Beispiel 7.16 erhält man
   $\lambda_{xy} = 0{,}09375$ und $\tau_{xy} = 0{,}07387$ aber
   $\lambda_{yx} = 0{,}26563$ und $\tau_{yx} = 0{,}11601$
   (Zwischenergebnisse bei λ_{yx}: E1 = 0,64 und E2 = 0,47

bzw. bei τ_{yx} F1 = 0,6418 und F2 = 0,56734)

2. Aus Gl. 7.39 folgt unmittelbar, dass τ bei Unabhängigkeit (wenn $p_{ij} = p_{i.}p_{.j}$ für alle Werte von i und j) den Wert Null annimmt. Man kann zeigen, dass dies auch für λ der Fall ist. Unabhängigkeit ist aber nur eine notwendige, nicht eine hinreichende Bedingung für das Verschwinden von λ.

   Die folgende Kontingenztafel

	y_1	y_2	y_3	Σ
x_1	20	40	10	70
x_2	10	15	5	30
Σ	30	55	15	100

   führt zu $\lambda_{xy} = \lambda_{yx} = 0$ obgleich keine Unabhängigkeit vorliegt (übrigens ist τ_{xy} = 0,002849 und τ_{yx} = 0,00433 also nicht Null).

   Das Beispiel ist so konstruiert, dass alle Zeilenmaxima in Spalte 2 und alle Spaltenmaxima in Zeile 1 sind.

3. Es ist leicht zu sehen, dass λ und τ dann den Wert 1 annehmen, wenn jede Zeile und jede Spalte einer quadratischen Kontingenztafel mit nur einer Häufigkeit besetzt ist, wie dies in Tafel 1 von Bsp. 7.15 der Fall ist [was oben als "vollständige Abhängigkeit" bezeichnet wurde].

4. Zu weiteren Eigenschaften von τ und λ (im Spezialfall der Assoziationsanalyse) vgl. Bem. 8 zu Def. 7.16.

5. Als ein Maß der proportionalen (relativen) Fehlerreduktion könnte man auch interpretieren:

   - das **Quadrat des Korrelationsverhältnisses**, wobei als Fehler E1 (oder F1) die Summe der Abweichungsquadrate SAQ_{tot} und als Fehler E2 (oder F2) SAQ_{int} auftritt;
   - das **Bestimmtheitsmaß** (im Falle linearer Regression) $B_{xy} = r_{xy}^2$.

   Das legt den Gedanken nahe, dass nicht λ (τ) sondern die Wurzel von λ (bzw. τ) eigentlich das Kontingenzmaß sein sollte.

3. Assoziation Vierfelderkorrelation

a) Dichotome Merkmale

Von Assoziation spricht man, wenn X und Y dichotome (binäre) Merkmale sind. Man kann dann wie folgt codieren:

$$x = \begin{cases} 0 & \text{wenn eine bestimmte Eigenschaft nicht vorhanden ist} \\ 1 & \text{wenn diese Eigenschaft (das Attribut) vorhanden ist} \end{cases}$$

und entsprechend ist Y als 0-1-Variable gegeben (d.h. es gibt nur die beiden Ausprägungen y=0 und y=1).

Man kann auch codieren: ja (1), nein (0) oder "+" (für 1) und "-" (für 0), um anzuzeigen, dass lediglich danach unterschieden wird, ob eine Eigenschaft (ein "Attribut") gegeben ist oder nicht gegeben ist.

Def. 7.15: Vierfeldertafel

Die zeilenweise (spaltenweise) Anordnung des dichotomen Merkmals X (Y) mit den Häufigkeiten a,b,c,d (statt $n_{11},n_{12},n_{21},n_{22}$) in der folgenden Art

	Variable Y		
Variable. X	y=1	y=0	Σ
x=1	a	b	a+b
x=0	c	d	c+d
Σ	a+c	b+d	n

heißt Vierfeldertafel (oder Assoziationstafel; vgl. Bsp. 7.1 für Tafeln dieser Art).

b) Bedingte Mittelwerte, Unabhängigkeit, Regressionsgeraden

1. Mit der Codierung als 0-1-Variablen sind Mittelwerte und bedingte Mittelwerte als relative Häufigkeiten (bzw. bedingte relative Häufigkeiten) zu interpretieren:

$$(7.40) \quad \bar{x} = \frac{1\cdot(a+b) + 0\cdot(c+d)}{n} = \frac{a+b}{n}$$

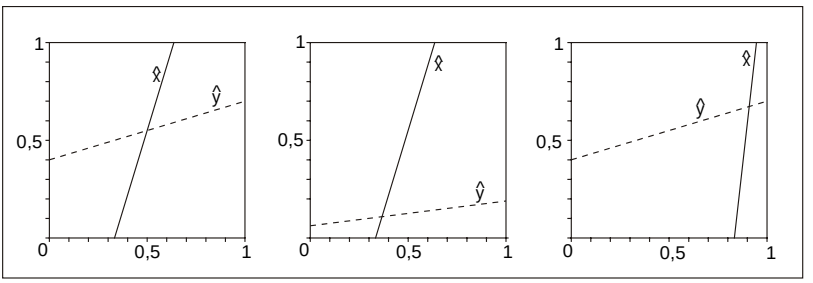
$$(7.41) \quad \bar{x}(y_1) = a/(a+c) \text{ und } \bar{x}(y_2) = b/(b+d)$$

   Entsprechendes gilt für den unbedingten Mittelwert $\bar{y}$ und für die bedingten Mittelwerte $\bar{y}(x_i)$.

2. Die Punkte $Px_1[a/(a+c),1]$ und $Px_2[b/(b+d),0]$ stellen im x,y-Koordinatensystem die bedingten Mittelwerte $\bar{x}(y_j)$ (mit j=1,2) dar. Ihre lineare Verbindung stellt die Regressionslinie zur Schätzung von x aufgrund

von y dar. Weil es sich nur um zwei Punkte handelt, ist es eine **Regressionsgerade** und es lässt sich zeigen, dass diese mit der im Kap. 8 eingeführten Regressionsgeraden $\hat{x}$ identisch ist. Entsprechend ist die Regressionsgerade $\hat{y}$ durch die Punkte $Py_1[1,a/(a+b)]$ und $Py_2[0,c/(c+d)]$ gegeben. In Abb. 7.5 sind die Regressionsgeraden für ein Zahlenbeispiel dargestellt.

Abb. 7.5: Regressionsgeraden für das Beispiel 7.17



3. Die Unterschiedlichkeit der bedingten Mittelwerte ist Ausdruck der Abhängigkeit der Merkmale untereinander. Bei Unabhängigkeit (Def. 7.5) gilt:

$$\bar{x}(y_1) = \bar{x}(y_2) = \bar{x} \text{ und } \bar{y}(x_1) = \bar{y}(x_2) = \bar{y}.$$

Hieraus folgt als Bedingung für die Unabhängigkeit

(7.42) $ad = bc.$

Weil die Merkmale X und Y hier nur zwei Ausprägungen haben, kann die Abhängigkeit nur linear sein, weshalb hier zwischen Unabhängigkeit und Unkorreliertheit (keine Assoziation) nicht unterschieden werden kann.

c) Normierung eines Assoziationsmaßes

Ein Assoziationsmaß A soll bei Unabhängigkeit den Wert Null annehmen. Die Obergrenze des Betrags von A ist dagegen nicht eindeutig. Besteht zwischen den Ausprägungen eine Ordnungsrelation etwa dergestalt, dass $x_1 > x_2$ und $y_1 > y_2$ (Dichotomien mit dahinterstehender Rangordnung), so kann auch sinnvoll zwischen positiver und negativer Assoziation unterschieden werden. Es sollte dann gelten:

(7.43) $-1 \le A \le +1.$

Man kann dann unterscheiden (nach M.G.Kendall):

total association	$b = c = 0$
total disassociation	$a = d = 0$
complete association	$b = 0$ oder $c = 0$

complete disassociation $a = 0$ oder $d = 0$

Es soll im folgenden von totaler bzw. vollständiger Assoziation bzw. Disassoziation gesprochen werden.

- Bei totaler Assoziation bzw. Disassoziation nehmen jeweils **beide** bedingten Mittelwerte (bedingte relative Häufigkeiten) einer Regressionsgeraden die Werte Null und Eins an.
- Bei vollständiger Assoziation bzw. bei vollständiger Disassoziation nimmt jeweils nur einer der beiden Endpunkte einer Regressionsgeraden die extremen Werte Null und Eins an.

Die Funktionen der beiden "Regressionsgeraden" lauten:

(7.44) $$\hat{y} = \frac{c}{c+d} + \left[\frac{a}{a+b} - \frac{c}{c+d}\right] x$$

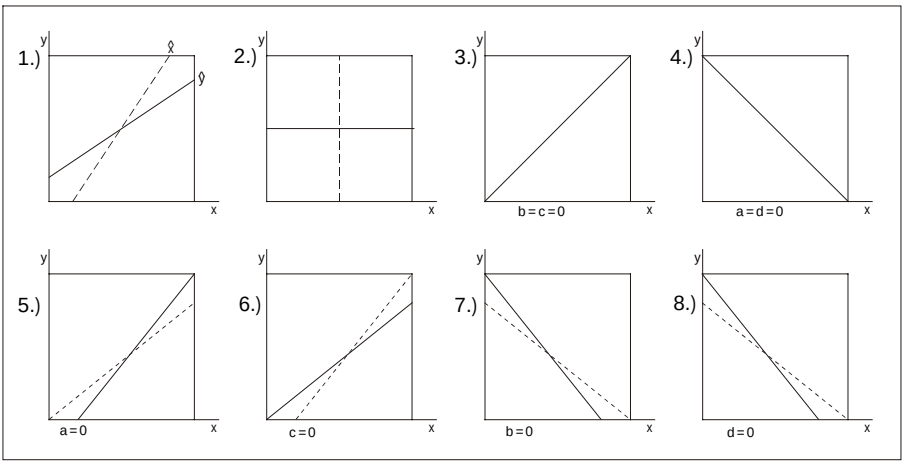
und

(7.45) $$\hat{x} = \frac{b}{b+d} + \left[\frac{a}{a+c} - \frac{b}{b+d}\right] y.$$

Zur Gestalt der Regressionsgeraden in den beschriebenen Fällen vergleiche man Abb. 7.6. Dabei zeigt sich, dass es durchaus sinnvoll ist, totale und vollständige Assoziation/Disassoziation als unterschiedlich anzusehen.

Abb.7.6: Dichotome "Regression"

1: allgemeiner Fall; 2: Unabhängigkeit; 3: totale Assoziation 4: totale Disassoziation, 5/6: vollständige Assoziation; 7/8: vollständige Disassoziation



d) Axiomatik für Assoziations- und Kontingenzmaße

Die unter c) getroffenen Aussagen über den Wertebereich eines Assoziationsmaßes führen zu dem Gedanken, auch hier eine Axiomatik anzugeben (vgl. Seite 222f). Für das Assoziationsmaß A sollte speziell die Forderung A2 noch wie folgt spezifiziert werden:

A2: Ist totale Assoziation gegeben, so sollte A = +1 sein. Bei geordneten Ausprägungen $x_1 > x_2$ und $y_1 > y_2$ sollte auch Assoziation und Disassoziation unterscheidbar sein und bei totaler Disassoziation den Wert A = -1 annehmen.

Die Axiome A1, A3 und A4 (vgl. Seite 222f) sind auch bei Kontingenzmaßen sinnvoll zu fordern. Ein Assoziationsmaß kann sehr wohl invariant sein gegenüber einer Ver-k-fachung aller Häufigkeiten, gleichwohl aber A3 nicht erfüllen (Das gilt z.B. für das Maß Q, vgl. Gl. 7.50).

Def. 7.16: Einige Assoziationsmaße

a) Die in Def. 7.13 zusammengestellten Kontingenzmaße haben im speziellen Fall einer Vierfeldertafel die folgende Gestalt:

$$(7.46)\quad \phi = \frac{ab - cd}{\sqrt{(a+b)(a+c)(b+d)(c+d)}} = \sqrt{\frac{c^2}{n}}$$

heisst Phi-Koeffizient oder Vierfelderkorrelation.
Es gilt $\phi = T = V$. Das Kontingenzmaß von Pearson ist speziell bei zwei dichotomen Merkmalen mit $C = \sqrt{\frac{f^2}{f^2+1}}$ kein sinnvoller Ausdruck.

b) Vorgeschlagen wurde auch

(7.47)	$\lvert xy\rvert = (ad - bc)/n^2$	Kreuzprodukt von Lazarsfeld
(7.48)	$cpr = ad/bc$	Kreuzproduktverhältnis
(7.49)	$\delta_x = a/(a+c) - b/(b+d)$	und
(7.49a)	$\delta_y = a/(a+b) - c/(c+d)$	(die sog. Anteilsdifferenzen)

Offenbar ist $\phi = \sqrt{d_x d_y}$ und die Anteilsdifferenzen δ_x und δ_y sind die Steigungen der Regressionsgeraden $\hat{x}$ und $\hat{y}$.

c) Auf dem Vergleich konkodanter und diskonkodanter Merkmalskombinationen (Paare) beruht das Assoziationsmaß Q von Yule

$$(7.50)\quad Q = \frac{ad - bc}{ad - bc} = \frac{cpr - 1}{cpr + 1} \quad .$$

sowie der Verbundenheitskoeffizient von Yule (coefficient of colligation):

$$(7.50a)\quad Y = \frac{\sqrt{ad} - \sqrt{bc}}{\sqrt{ad} + \sqrt{bc}} \quad ,$$

der nur von geringer Bedeutung ist, da er nur eine monotone Transformation von Q darstellt.

d) Ohne Bedeutung für die Assoziationsmessung sind die in Def. 7.14 definierten Kontingenzmaße von Goodman und Kruskal.

Bemerkungen zu Def. 7.16:

1. Bei einer Codierung von X und Y mit den Variablenwerten 0 und 1 ist das Kreuzprodukt |xy| die Kovarianz zwischen X und Y. Für Varianzen gilt $s_x^2 = (a+b)(c+d)/n^2$ und $s_y^2 = (a+c)(b+d)/n^2$, so dass ϕ nichts anderes ist als der Produkt-Moment-Korrelationskoeffizient r_{xy} für den Fall einer 0-1-Codierung von X und Y.

2. Bedeutet x = 1 (y = 1) das Auftreten des Ereignisses A (bzw. B) so ist |xy| die Abweichung von der stochastischen Unabhängigkeit |xy| = |AB| = P(AB) - P(A)P(B). Als Kovarianz hat das Kreuzprodukt den Wertebereich $-1/4 \leq |xy| \leq +1/4$. Die Schreibweise |xy| oder |AB| beruht darauf, dass diese Größe die Determinante der Vierfeldertafel (mit relativen statt absoluten Häufigkeiten, bzw. mit Wahrscheinlichkeiten) darstellt. Das Kreuzprodukt spielt in der "Latent Structure Analysis" eine große Rolle.

3. Yules Assoziationsmaß Q erfüllt die Axiome A1 und A4. Es nimmt aber die extremen Werte +1 bzw. -1 nicht nur im Falle totaler, sondern auch vollständiger Assoziation bzw. Disassoziation an. Q erfüllt also A2 nicht (vollständig) und auch A3 nicht.
   Die drei Vierfeldertafeln von Beispiel 7.17 führen alle zum gleichen Wert von Q = 5/9 und auch zum gleichen Kreuzproduktverhältnis cpr = 3,5, obgleich die Phi-Koeffizienten unterschiedlich sind.

4. Aus Gl. 7.46 folgt, dass bei Ver-k-fachung aller Häufigkeiten χ^2 auch k-mal so groß ist wie bisher, so dass χ^2 das Axiom A4 nicht erfüllt. χ^2 hat jedoch sehr günstige Aggregationseigenschaften, weil es in beliebige Teilsummen aufzuspalten ist. Außerdem sind χ^2 und alle auf χ^2 basierenden Maße invariant gegenüber Vertauschungen von Zeilen und Spalten.

5. Die mit den Axiomen A1 und A2 geforderte Normierung des Wertebereichs wird vom Kreuzprodukt nicht erfüllt: Die Grenzen des Intervalls $-1/4 \leq |xy| \leq +1/4$ werden nur erreicht bei totaler Assoziation **und** wenn die Varianzen von X und Y jeweils 1/4 (also maximal) sind, also a=d bzw. b=c beträgt.

6. Zum Kreuzproduktverhältnis: Es ist nichtnegativ und es gilt $0 \leq cpr \leq 1$ bei Disassoziation, cpr = 1 bei Unabhängigkeit und $1 < cpr < \infty$ bei Assoziation. Schon bei vollständiger Disassoziation ist cpr = 0.

7. Wegen der Abhängigkeit von χ^2 gilt für C die Einschrankung: $0 \le C \le \sqrt{1/2}$, mit der Obergrenze $\sqrt{1/2}$ bei totaler Assoziation oder Disassoziation.

8. Überträgt man die Kontingenzmaße von Goodman und Kruskal (Def. 7.14) auf den speziellen Fall der Assoziationsanalyse (r=c=2) so ergeben sich komplizierte Ausdrücke. Die Maße haben auch erhebliche Nachteile: λ kann auch ohne Unabhängigkeit den Wert Null annehmen, andererseits sind λ und τ zwar bei totaler, nicht aber bei vollständiger Assoziation 1, wie die folgenden zwei Tabellen zeigen:

	y_1=1	y_2=0	Σ
x_1=1	4	2	6
x_2=0	3	1	4
Σ	7	3	10

$\lambda_{xy} = \lambda_{yx} = 0$
dagegen $\phi = 0{,}0891$ und
$\tau_{xy} = 0{,}00794 = 1/126 = \tau_{yx}$

	y_1=1	y_2=0	Σ
x_1=1	7	1	8
x_2=0	0	2	2
Σ	7	3	10

$\lambda_{xy} = 2/3$, $\lambda_{yx} = 1/2$
(vollständige Assoziation)
$\tau_{xy} = 0{,}58333 = 7/12 = \tau_{yx}$

Man kann zeigen, dass τ bei einer Vierfeldertafel stets symmetrisch ist.

Beispiel 7.17:

Man berechne die Assoziationsmaße der Def. 7.16 (a-c) sowie die Regressionsgeraden für die folgenden drei Vierfeldertafeln:

Tafel 1

	y=1	y=0	Σ
x=1	70	30	100
x=0	40	60	100
Σ	110	90	200

Tafel 2

	y=1	y=0	Σ
x=1	7	30	37
x=0	4	60	64
Σ	11	90	101

Tafel 3

	y=1	y=0	Σ
x=1	70	30	100
x=0	4	6	10
Σ	74	36	110

Welche Besonderheiten gelten für das Assoziationsmaß Q sowie das Kreuzproduktverhältnis (cpr)? Wie reagiert χ^2 bei einer Verdoppelung aller Häufigkeiten der Tafel 1?

Lösung 7.17:

1) Als Muster jeweils die Berechnung für Tafel 1
Q = (70·60-30·40)/(70·60+30·40) = (4200-1200)/(4200+1200) = 5/9
cpr = 4200/1200 = 3,5
Phi-Koeffizient $\phi = (70\cdot60-30\cdot40)/\sqrt{100\cdot100\cdot110\cdot90} = 0{,}30151$
$\chi^2 = n\phi^2 = 200/11 = 18{,}182$ und $C = \sqrt{1/12} = 0{,}288675$
Regressionsfunktionen $\hat{y} = 0{,}4 + 0{,}3x$ und $\hat{x} = 1/3 + (10/33)y$
daraus folgt: Anteilswertdifferenzen $\delta_x = 10/33$ und $\delta_y = 0{,}3$ sowie für den Phi-Koeffizienten $\phi = \sqrt{d_x\, d_y} = \sqrt{30/330} = 0{,}30151$.

2) Berechnungen für Tafel 2 und Tafel 3:
$Q_2 = Q_3 = 5/9$ (= Q_1)

$cpr_2 = cpr_3 = 3{,}5$ (= cpr_1).

Es fällt also auf, dass Q und cpr invariant sind gegenüber proportionalen Transformationen einzelner Zeilen und Spalten, was für die übrigen Assoziationsmaße nicht gilt:

$\phi_2 = 0{,}195935$ ($\neq \phi_1 = 0{,}3015$) $\phi_3 = 0{,}183804$

$\chi^2_2 = 3{,}87746$, $\chi^2_3 = 3{,}71622$ und $C_2 = 0{,}192289$, $C_3 = 0{,}18078$

Regressionsgeraden und Anteilsdifferenzen:

Tafel 2	Tafel 3
$\hat{y} = 0{,}0625 + 0{,}1267x$	$\hat{y} = 0{,}4 + 0{,}3x$
$\hat{x} = 1/3 + (10/33)y$	$\hat{x} = 5/6 + 0{,}1126y$
$\delta_{y2} = 0{,}1267$ $\delta_{x2} = 10/33$	$\delta_{y3} = 0{,}3$ $\delta_{x3} = 0{,}1126$

3) Veränderung von χ^2 bei Verdoppelung der Häufigkeiten:

Tafel 1

	y=1	y=0	Σ
x=1	70	30	100
x=0	40	60	100
Σ	110	90	200

$\chi^2 = 200/11 = 18{,}1818$

$\phi^2 = 1/11$

Tafel 1a

	y=1	y=0	Σ
x=1	140	60	200
x=0	80	120	200
Σ	220	180	400

$\chi^2 = 400/11 = 36{,}3636$

$\phi^2 = 1/11$

Verdoppelung der Häufigkeiten führt auch zu einer Verdoppelung von χ^2, lässt aber den Phi-Koeffizienten unberührt.

Beispiel 7.18:

Man berechne ϕ sowie Goodman Kruskals λ für die folgenden Vierfeldertafeln!

	$y_1=1$	$y_2=0$	Σ
$x_1=1$	4	6	10
$x_2=0$	7	5	12
Σ	11	11	22

	$y_1=1$	$y_2=0$	Σ
$x_1=1$	4	7	11
$x_2=0$	6	5	11
Σ	10	12	22

Lösung 7.18:

Man erkennt, dass die rechte Tafel nur die transponierte linke Tafel ist. Für λ erhält man für die linke Tafel $\lambda_{xy} = 1/30 = 0{,}0333$ und $\lambda_{yx} = 1/30$. λ_{xy} der linken Tafel ist gleich λ_{yx} der rechten Tafel. Man erhält also stets den gleichen Wert 1/30 für λ. Das gleiche gilt für ϕ, das ohnehin symmetrisch und deshalb für die linke und rechte Tafel gleich ist, nämlich $\phi = 0{,}18257$.

c) Rangkorrelation, Zusammenhang bei ordinalskalierten Variablen

1. Rangkorrelation

Verzichtet man auf eine Metrik und nutzt nur die Ranginformation in den Tupeln (x_v, y_v) oder liegen nur ordinalskalierte Variablen vor, die in eine Rangskala transformiert werden können, so lassen sich Rangkorrelationskoeffizienten berechnen. Man kann dann nicht mehr von einem linearen Zusammenhang sprechen, sondern von einem Zusammenhang, der (bei Rangkorrelation) linear in den Rängen $R(x_v)$, $R(y_v)$ ist. Bei der Rangassoziation werden keine Ränge vergeben (Def. 7.17), der Zusammenhang ist dann monoton.

Def. 7.17: Rangtransformation

Der zweidimensionalen Variable (X,Y) mit den der Größe nach geordneten Ausprägungen x_v, y_v werden Rangzahlen mit $R(x_v) = v$ und $R(y_v) = v$ zugeordnet.

Bemerkungen zu Def. 7.17:

1. Die Definition der Rangzahlen setzt voraus, dass X und Y mindestens ordinalskaliert sind und alle Ausprägungen verschieden sind, so dass $x_{(v-1)} < x_{(v)} < x_{(v+1)}$ (und die Ordnung der y-Werte entsprechend definiert ist). Die Rangzahlen (Ränge) sind natürliche Zahlen.

2. Sind die Ausprägungen nicht jeweils alle verschieden, d.h. treten **Bindungen** (ties) auf, so wird jeweils k gleichen Werten das arithmetische Mittel der auf sie entfallenden Rangzahlen zugeordnet.

   Beispiel: Bei $x_{(1)} < x_{(2)} < x_{(3)} = x_{(4)} = x_{(5)} < x_{(6)} = x_{(7)}$ erhalten die 3te bis 5te Ausprägung alle den Rangplatz 4 und die 6te und 7te Ausprägung den Rang 6,5. Die Reihe der Rangzahlen lautet also 1, 2, 4, 4, 4, 6½, 6½. Diese Art Bindungen zu behandeln verändert die Rangsumme $\Sigma R(x_v)$ (und $\Sigma R(y_v)$ entsprechend) nicht, wohl aber die Summe der quadrierten Ränge $\Sigma[R(x_v)]^2$. Sie ist bei Auftreten von Bindungen kleiner als bei Ausprägungen, die alle unterschiedlich sind. Denn in $\Sigma[R(x_v)]^2$ tritt im Falle n gleicher Werte n-mal die Größe $\bar{x}^2$ auf und bei n unterschiedlichen Werten dagegen die Größe Σx_j^2 (j = 1,2,...,n). Es gilt als Konsequenz des Verschiebungssatzes der Varianz $\Sigma x_j^2 > n\bar{x}^2$, denn $n^{-1}\Sigma x_j^2 - \bar{x}^2 = s_x^2 > 0$.

Def. 7.18: Rangkorrelationskoeffizient nach Spearman

Der Korrelationskoeffizient nach Bravais-Pearson für Rangzahlen

$$(7.51)\quad R_{xy}^{S} = 1 - \frac{6\Sigma d_v^2}{n(n^2-1)} \qquad \text{mit } d_v = R(x_v) - R(y_v)$$

heisst Rangkorrelationskoeffizient nach Spearman (v = 1,2,...,n).

Bemerkungen zu Def. 7.18:

1. *Zur Größe d:* Hat z.B. die Einheit v (die v-te Beobachtung) bezüglich des Merkmals X den 5ten Platz und bei Y den 3ten Rang, so ist $R(x_v) = 5$ und $R(y_v) = 3$, so dass $d_v = 5 - 3 = 2$ ist.

2. *Zusammenhang mit Produkt-Moment-Korrelation:* Da die Ränge $R(x_1),...,R(x_n)$ natürliche Zahlen sind, gilt $\Sigma R(x_v) = 1 + 2 + ... + n = n(n+1)/2$ (v = 1,2,...,n). Den gleichen Ausdruck erhält man für $\Sigma R(y_v)$. Die arithmetischen Mittel sind also $\overline{R(x)} = \overline{R(y)} = (n+1)/2$. Das gilt auch, wenn Bindungen auftreten. Ferner gilt: $\Sigma[R(x_v)]^2 = \Sigma[R(y_v)]^2 = n(n+1)(2n+1)/6$, so dass man für die Varianzen $s^2_{R(x)}$ und $s^2_{R(y)}$ der Rang-reihen R(x), R(y) den Ausdruck $(n^2-1)/12$ erhält.

   Treten Bindungen auf, so werden die Varianzen geringer sein (vgl. Bem. 2 zu Def. 7.17).

   Nach Satz 7.4 ist die Kovarianz zwischen den Rängen $(n^2-1)/12 - (\Sigma d^2{}_v)/2n$, woraus sich mit Def. 7.8 (Gl. 7.20) für die Korrelation r_{xy} zwischen den Rängen Gl. 7.51 er-gibt. Der Rangkorrelationskoeffizient von Spearman ist also die Produkt-Moment-Korrelation zwischen den Rangzahlen $R(x_v)$, $R(y_v)$.

3. *Skalen*: Mit den Rangplätzen (Rängen) wird gerechnet wie mit metrisch skalierten Variablen, was eigentlich voraussetzt, dass die Abstände zwischen den Rängen gleich groß sind. Es wird also angenommen, dass zwar nicht die Ausprägungen der zu korrelierenden Merkmale X und Y selber, wohl aber die hierfür vergebenen Rangplätze metrisch skaliert sind.

4. Im Falle von Bindungen werden im Nenner $n(n^2-1)$ von Gl. 7.51 Korrekturen angebracht. Der Rangkorrelationskoeffizient ist dann gleichwohl meist kleiner als im Falle ohne Bindungen.

Axiomatik

Es ist von einem Rangkorrelationskoeffizienten R zu fordern:

R1: $-1 \le R \le 1$

R2: Bei vollständiger Übereinstimmung der Rangreihen, also $R(x_v) = R(y_v)$ für alle v = 1,2,...,n, soll R = +1 sein.

R3: Bei inverser Rangordnung $R(y_v) = (n+1) - R(x_v)$ soll R = -1 sein.

R4: Bei Unabhängigkeit von X und Y soll R = 0 gelten.

R5: Rangkorrelationskoeffizienten sollten invariant sein gegenüber monoton steigenden Transformationen der Rangzahlen.

Weitere Bemerkungen zu Spearmans Rangkorrelation

Man sieht leicht, dass R^s die Axiome R1 bis R3 erfüllt. Bei Übereinstimmung ist $d_v = 0$ für alle v = 1,2,...,n, so dass $\Sigma d_v^2 = 0$ und $R^s = 1$. Bei inverser Rangordnung ist $d_v = (n+1) - 2R(x_v)$, so dass $\Sigma d_v^2 = -n(n+1)^2 + 2n(n+1)(2n+1)/3 = n(n^2-1)/3$. Dies eingesetzt in Gl. 7.51 liefert $R^s = -1$. Ist $R^s = 0$, so ist der Mittelwert der Rangdifferenzen gleich der Summe der Varianzen, d.h. es gilt $(\Sigma d_v^2)/n = (n^2 -1)/6$ und die Kovarianz zwischen den Rängen ist Null. Wie man auch sieht, ist $(n^2-1)/6$ der halbe Betrag dessen, was sich für $(\Sigma d_v^2)/n$ bei $R^s = -1$ ergibt. Während die Situationen $R^s = +1$ und $R^s = -1$ eindeutig sind, ist $R^s = 0$ mit verschiedenen Konstellationen verträglich (vgl. Beispiel 7.19). Spearmans Rangkorrelationskoeffizient erfüllt auch nicht das Axiom R5. Eine Verdoppelung aller Rangzahlen führt zu einer Vervierfachung der Σd_v^2, wodurch sich R^s verändert (wobei R^s dann nicht mehr zwischen -1 und +1 liegen muss; die Ränge sind dann ja auch nicht mehr, wie in Def. 7.18 vorausgesetzt, die Folge natürlicher Zahlen).

Beispiel 7.19:

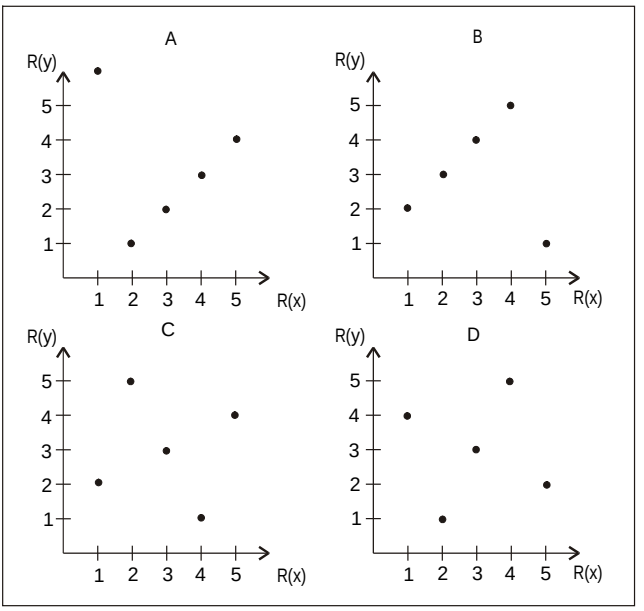
Man zeige, dass bei gegebener Rangfolge der X-Werte die folgenden vier Rangordnungen der Y-Werte jeweils zu $R^s = 0$ führen.

R(x)	Rangordnungen von y			
	A	B	C	D
1	5	2	2	4
2	1	3	5	1
3	2	4	3	3
4	3	5	1	5
5	4	1	4	2

Lösung 7.19:

In allen vier Fällen (A bis D)- vgl. Abb.7.7 - ist die Summe der quadrierten Rangdifferenzen 20, so dass bei n=5 gilt: $R^s = 1 - \frac{6 \cdot 20}{5(25-1)} = 1 - 1 = 0$.

Abb. 7.7: Fälle mit einer Rangkorrelation von R^S = 0 (Bsp 7.19)



2. Paarvergleiche, Bindungen, Rangassoziation

Die im Teil 1 betrachteten Daten waren vom folgenden Typ:

- Einzelbeobachtungen (n Einheiten),
- den Merkmalswerten werden Ränge zugeordnet (Rangtransformation Def. 7.17) und sie
- sind i.d.R. alle unterschiedlich, so dass die rangtransformierten Merkmalswerte $R(x_v)$, $R(y_v)$ natürliche Zahlen von 1 bis n sind,
- nur als Ausnahme gibt es auch "Bindungen" (Bem. 2 zu Def. 7.17).

Die Merkmale X und Y sind ordinalskaliert, aber durch die **Rangtransformation** entsteht mehr als eine Ordinalskala: Ränge sind äquidistant, die Merkmalsausprägungen, für die diese Ränge vergeben werden, sind es aber nicht. Es ist zweifelhaft, ob eine solche Anhebung des Skalenniveaus vertretbar ist. Wünschenswert ist demgegenüber

1. eine Datenanalyse ordinaler Merkmale, die nur die Information einer Ordinalskala ausnutzt, d.h. nur dass bei $x_i > x_j$ die Ausprägung x_i des Merkmals X "größer", "höher" o.ä. ist als x_j,
2. die Berücksichtigung häufigen (nicht nur als Ausnahme) Auftretens gleicher Merkmalsausprägungen (also von Bindungen).

Soll ohne Rangtransformationen "Zusammenhang" bei ordinalen Merkmalen definiert werden (Rangassoziation statt Rangkorrelation), so ist vom Paarvergleich auszugehen.

Def. 7.19: Paarvergleiche

a) Die folgende Tabelle soll ***Kontingenztabelle*** genannt werden, obgleich dieser Begriff speziell häufig für nominalskalierte Merkmale benutzt wird. Die Anzahl der Zeilen und Spalten ist nur beispielhaft. Die Zusammenhänge gelten allgemein:

	y_1	y_2	y_3	Σ
x_1	n_{11}	n_{12}	n_{13}	$n_{1.}$
x_2	n_{21}	n_{22}	n_{23}	$n_{2.}$
Σ	$n_{.1}$	$n_{.2}$	$n_{.3}$	n

Es gilt $y_1 > y_2 > y_3$ und $x_1 > x_2$. Die n Einheiten lassen n(n-1)/2 Paarvergleiche zu. Alle n_{ij} Einheiten sind jeweils bezüglich X **und** Y gleich (Bindungen).

b) ***Bindungen (Verknüpfungen, ties)***: Einheiten sind verknüpft, wenn sie bezüglich eines Merkmals oder beider Merkmale gleiche Ausprägungen haben.

c) Jeder der n(n-1)/2 Paarvergleiche ist ein Vergleich im Sinne eines der folgenden fünf Typen:

1. Konkordante Vergleiche:
n_{ij} Einheiten mit $X = x_i$ und $Y = y_j$ werden verglichen mit allen Einheiten, für die gilt: $X < x_i$ und $Y < y_j$. Die Anzahl der konkordanten Paarvergleiche ist $N_c = n_{11}(n_{22} + n_{23}) + n_{12}n_{23}$.

2. Diskonkordante Vergleiche:
N_d ist die Anzahl der Vergleiche von jeweils n_{ij} Einheiten mit solchen Einheiten, bei denen $X > x_i$ und $Y < y_j$. Für die Tabelle gilt
$N_d = n_{13}(n_{21} + n_{22}) + n_{12}n_{21}$.

3. Es gibt T_x in X verknüpften Vergleichen
$T_x = n_{11}(n_{12} + n_{13}) + n_{12}n_{13} + n_{21}(n_{22} + n_{23}) + n_{22}n_{23}$.

4. Bei T_y-Vergleichen liegen Bindungen bezüglich Y vor:
(Vergleiche mit den restlichen Einheiten einer Spalte):
$T_y = n_{11}\,n_{21} + n_{12}\,n_{22} + n_{13}n_{23}$.

5. Verknüpft in X und Y
sind T_{xy}-Vergleiche:
$T_{xy} = 1/2\ \Sigma\Sigma n_{ij}(n_{ij}-1) = 1/2[n_{11}(n_{11}-1) + n_{12}(n_{12}-1 + ... + n_{23}(n_{23}-1)]$.

d) Es gilt

$$n(n-1)/2 = N_c + N_d + T_x + T_y + T_{xy}. \quad (7.52)$$

Bemerkungen zu Def. 7.19:

1. Man kann N_c auch bestimmen durch Vergleiche mit Einheiten, für die gilt $X > x_i$ und $Y > y_j$ statt mit Einheiten, für die gilt $X < x_i$ und $Y < y_i$: Die Summen $n_{11}(n_{22} + n_{23}) + n_{12}n_{23}$ und $n_{23}(n_{12} + n_{11}) + n_{22}n_{11}$ sind gleich. Bei einer Vierfeldertafel ist $N_c = ad$.

2. Auch hier könnte man alternativ definieren: Vergleiche mit Einheiten, bei denen $X < x_i$ und $Y > y_j$. Man erhält N_d auch mit $n_{21}(n_{12} + n_{13}) + n_{22}n_{13} = n_{13}(n_{21} + n_{22}) + n_{12}n_{21}$. Bei einer Vierfeldertafel ist $N_d = bc$.

3. Bei einer Vierfeldertafel ist $T_x = ab + cd$ und $T_y = ac + bd$.

Def. 7.20: Maße der Rangassoziation

Auf der Anzahl konkordanter und diskonkordanter Paarvergleiche beruhen die folgenden Maße der Rangassoziation.

(7.53) $\gamma = (N_c - N_d)/(N_c + N_d)$ index of order association von Goodman/Kruskal.
[Es wird auch ω als Symbol verwendet; bei Vierfeldertafeln sind γ und Yules Q identisch].

(7.54a) $\tau_a = \frac{2(N_c - N_d)}{n(n-1)} = \frac{N_c - N_d}{\binom{n}{2}}$ (Kendalls τ) oder in einer anderen Version

(7.54b) $\tau_b = \frac{N_c - N_d}{\sqrt{(N_c+N_d+T_x)(N_c+N_d+T_y)}}$

[bei Vierfeldertafeln ist dies gleich dem Phi-Koeffizient, vgl. Def. 7.16, Gl. 7.46].

Bemerkungen zu Definition 7.20:

1. Sämtliche Werte in einem Tabellenfeld stellen untereinander verglichen Bindungen in X **und** Y dar. Es gilt bei $n = \Sigma\Sigma n_{ii}$, also bei einer symmetrischen Tabelle, die nur in der Hauptdiagonalen besetzt ist: $T_{xy} = \Sigma\Sigma[n_{ii}(n_{ii}-1)]/2$, so dass $n(n-1)/2 = N_c + T_{xy}$. Die Konsequenz ist, dass selbst in diesem Falle vollständiger Rangassoziation τ_a nicht den Maximalwert 1 annimmt, denn $N_c < \binom{n}{2}$ und $N_d = 0$, es sei denn, für alle i gilt $n_{ii} = 1$.

2. Der Koeffizient τ_b enthält entsprechende Korrekturen. Er ist betragsmäßig nie kleiner als τ_a. Ob er seinen Maximalwert erreicht, hängt auch davon ab, ob die Tabelle quadratisch und die Randverteilungen Gleichverteilungen sind.

3. Kendalls τ wird aus Einzelwerten (n Tupel für die n Einheiten) wie folgt berechnet
(7.55) $\tau_a = \Sigma\Sigma d_{ij}/[n(n-1)/2]$,

wobei d_{ij} den Wert +1 oder –1 annimmt, je nachdem, ob zwei verglichene Objekte (Einheiten) rangmäßig in X und Y gleich oder ungleich abgestuft sind und alle Einheiten mit jeder anderen Einheit verglichen werden.

4. Die Vergabe der Werte +1 für konkordante und und -1 für diskonkordante Paarvergleiche legt auch die folgende Schreibweise nahe:

   (7.56) $$\tau_a = \frac{\Sigma\Sigma a_{ij} b_{ij}}{\sqrt{\Sigma\Sigma a_{ij}^2 \Sigma\Sigma b_{ij}^2}}$$

   wobei $a_{ij} = +1$ wenn $R(x_i) > R(x_j)$ und $a_{ij} = -1$ wenn $R(x_i) < R(x_j)$ und b_{ij} entspre-chend für R(y) definiert ist. Konkordante Paarvergleiche ergeben $a_{ij}b_{ij} = +1$ und ent-sprechend ist $a_{ij}b_{ij}$ bei diskonkordanten Vergleichen -1. Gl. 7.56 erinnert an den Pro-dukt-Moment-Korrelationskoeffizienten.

5. Im Unterschied zu den Koeffizienten von Goodman und Kruskal sind die Maße der Def. 7.20 symmetrische Zusammenhangsmaße.

6. Der Koeffizient γ (Gl. 7.53) ist sehr beliebt und kann anders als Q für beliebig dimensionierte Kontingenztafeln berechnet werden. Außerdem kann γ im Sinne der proportionalen Fehlerreduktion interpretiert werden.

Beispiel 7.20:

Gegeben seien folgende Rangdaten (ohne Bindungen) von X und Y mit n = 4 Objekten (Einheiten) für die τ_a nach Gl. 7.55 und γ zu berechnen sind:

v	R(x)	R(y)
1	1	3
2	3	2
3	4	1
4	2	4

Lösung 7.20:

Die Objektpaare, die von X jeweils höher bewertet werden (niedrigerer Rang) als von Y sind 1 und 4: denn $R(x_1) < R(y_1)$ und $R(x_4) < R(y_4)$. Alle übrigen Objektpaare werden von X und Y unterschiedlich bewertet. Man erhält also die folgende Tabelle der Werte d_{ij} von Paarvergleichen, die nur oberhalb der Hauptdiagonalen ausgefüllt wird.

Objekte	1	2	3	4
1	-	-1	-1	+1
2		-	-1	-1
3			-	-1
4				-

Dann ist $\Sigma\Sigma d_{ij} = -4$, also bei n = 4 und $n(n-1)/2 = 6$, $\tau_a = -4/6 = -2/3$ oder in der Schreibweise von Gl. 7.53: $N_c = 1$, $N_d = 5$. Für γ erhält man (1-5)/(1+5) = -4/6, was mit τ_a identisch ist, weil keine Bindungen auftreten.

3. Spezialfall: ohne Bindungen

Die im letzten Abschnitt behandelten Konzepte lassen sich auch anwenden im Fall fehlender Bindungen (wie im Abschn. 1 angenommen). Man kann zeigen, dass dann gilt für Kendalls τ:

(7.57) $\tau_a = \frac{4\Sigma u_i}{n(n-1)} - 1,$

wobei $R(x_i) = i$ die Referenzrangreihe ist und u_i die Anzahl der in der nachfolgenden Reihe der Werte R(y) höheren Ränge als $R(y_i)$ ist. Ferner ist im Fall ohne Bindungen $\Sigma u_i = N_c$ und entsprechend die Summe niedrigerer Ränge $\Sigma v_i = N_d$ und $N_c + N_d = n(n-1)/2$. In diesem Fall ist Kendalls τ (τ_a) mit Goodman Kruskals γ identisch.

4. Konkordanzkoeffizient

Als Verallgemeinerung von Kendalls τ gilt der Konkordanzkoeffizient W von Kendall für den Vergleich von mehr als zwei Rangordnungen. Werden n Objekte durch m Personen rangmäßig beurteilt, so ergibt sich eine Matrix von Rangzahlen mit n Spalten und m Zeilen. Die zu erwartende Summe der Rangzahlen (Summe einer Spalte) eines jeden Objekts beträgt dann $M = n(n+1)/2$ und die tatsächlich erreichte Rangsumme des j-ten Objekts ($j = 1,2,..., n$) sei R_j. Gäbe es keinen Zusammenhang zwischen den m Rangordnungen, so wären die Rangsummen aller n Objekte jeweils gleich und somit $R_j = M$.

Def. 7.21: Konkordanzkoeffizient

Der Konkordanzkoeffizient mißt die Abweichung von der Gleichheit der Rangvergabe, d.h. er mißt, in welchem Maße sich die Rangordnungen unterscheiden. Er lautet:

(7.58) $W = 12\,\Sigma\, D_j^2/m^2n(n^2-1)$ für $D_j = R_j - M$ und $j=1,2,...,m$

d) Weitere Maße des Zusammenhangs

Auf einige weitere in Übers. 7.2 genannte Korrelationsmaße kann aus Platzgründen hier nicht eingegangen werden. Es soll jedoch kurz die punktbiserielle Korrelation hergeleitet werden.

Ist X dichotom (mit den Ausprägungen $x_o = 0$ und $x_1 = 1$ und den absoluten Häufigkeiten n_0 und n_1) und Y metrisch skaliert und sind $\bar{y}_0$ bzw. $\bar{y}_1$ die bedingten Mittelwerte von Y wenn $X = x_o$ bzw. $X = x_1$ ist, so gilt für die (zweipunktverteilte) Variable X:

Mittelwert: $\bar{x} = n_1/n = p_{1x} = 1 - p_{ox}$

Varianz: $s_x^2 = n_o n_1/n^2 = p_{ox}p_{1x}$

und für die Kovarianz zwischen X und Y

$s_{xy} = p_{1x}\bar{y}_1 - p_{1x}\bar{y} = p_{1x}(\bar{y}_1 - \bar{y})$, da $(1/n)\Sigma xy = p_{1x}\,\bar{y}_1$ ist.

Daraus ergibt sich für den Produkt-Moment-Korrelationskoeffizienten r_{xy} in diesem speziellen Fall einer dichotomen Variablen X:

$$(7.59)\quad r_{xy} = \frac{p_{1x}(\bar{y}_1-\bar{y})}{s_y\sqrt{p_{ox}p_{1x}}} = \frac{\bar{y}_1-\bar{y}}{s_y}\cdot\sqrt{\frac{p_{1x}}{p_{ox}}} \quad .$$

Das ist der punkt-biserielle Korrelationskoeffizient R_{pb}, der wegen $\bar{y} = p_{ox}\bar{y}_0 + p_{1x}\bar{y}_1$ auch mit unterschiedlichen Umformungen bekannt ist.

Def. 7.22: point biserial correlation

Der punkt-biserielle Korrelationskoeffizient R_{pb} ist definiert als

$$(7.60)\quad R_{pb} = (\bar{y}_1 - \bar{y}_0)\cdot\frac{s_x}{s_y}$$

mit $s_x = \sqrt{p_{0x}p_{1x}}$, wenn X zweipunktverteilt (dichotom) ist, d.h. die relativen Häufigkeiten betragen p_{0x} wenn $X = x_0$ und p_{1x} wenn $X = x_1$ und Y metrisch skaliert ist. In der umgekehrten Situation (Y dichotom) gilt entsprechend

$$(7.60a)\quad R_{pb} = (\bar{x}_1 - \bar{x}_0)\cdot(s_y/s_x) \quad \text{mit } s_y = \sqrt{p_{0y}p_{1y}}.$$

Bemerkungen zur Def. 7.22:

1. Wie aus der Herleitung von Gl. 7.59/60 hervorgeht, ist R_{pb} die Produkt-Moment-Korrelation für den speziellen Fall, dass eine der beiden Variablen dichotom ist. Gl. 7.60 hat eine gewisse Ähnlichkeit mit der Prüfgröße beim t-Test für zwei unabhängige Stichproben über den Unterschied zweier Mittelwerte.

2. Verabredet man s_o^2 für die bedingte Varianz von Y, wenn $X = x_o$ und s_1^2 entspre-chend, wenn $X = x_1$, so ist in Analogie zu Gl. 5.11 (Satz 5.5) $s_y^2 = (p_{0x}\, s_o^2 + p_{1x}\, s_1^2) + [p_{0x}\,(\bar{y}_0 - \bar{y})^2 + p_{1x}\bar{y}_1 - \bar{y})^2]$, worin der erste Ausdruck die interne und der zweite die externe Varianz ist, die auch zu $(\bar{y}_1 - \bar{y}_0)^2\ (p_{0x}\, p_{1x})^2 = (\bar{y}_1 - \bar{y}_0)^2\, s_x^2$ umgeformt werden kann. R_{pb} ist also auch ein analog dem Korrelationsverhältnis η konstruiertes Maß.

Hinweise auf weitere Korrelationskoeffizienten:

1. Der punktbiserielle Korrelationskoeffizient R_{pb} ist nicht zu verwechseln mit dem **biseriellen Korrelationskoeffizienten** R_b (vgl. Übers. 7.2), bei dem man voraus-setzt, dass das dichotome Merkmal X an sich normalverteilt ist und so dichotomisiert ist, dass für $-\infty < X \le x^*$ der Wert $X = x_0$ vergeben wurde und für $X > x^*$ der Wert $X = x_1$. Die Dichtefunktion der Normalverteilung hat an der Stelle x^* den Wert $f(x^*)$. Dann ist $R_b = [(\bar{y}_1-\bar{y})p_{1x}]/[s_y f(x^*)]$

2. Erwähnt sei auch der **tetrachorische Korrelationskoeffizient** mit zwei dichotomisierten, aber an sich normalverteilten Variablen. Er ist ein Assoziationsmaß und in exakter Form schwer zu berechnen. Meist wird er aufgrund des Kreuzproduktverhältnisses cpr aus Tabellen bestimmt.

3. Es gibt auch weitere, nicht im Abschnitt 4b behandelte Versuche, Kontingenzmaße auf der Basis des Konzepts der Fehlerreduktion zu konstruieren, wobei als "Fehler" auch Streuungsmaße benutzt werden, wie z.B. die Entropie.

5. Korrelation und Kausalität

Die Suche nach kausalen Erklärungen oder "Gesetzen" entspringt dem Bedürfnis, Regelmäßigkeiten zu finden. Praktisches Handeln ist unmöglich, ohne das Vertrauen darauf, dass unter ähnlichen Bedingungen auch ähnliches geschehen wird. Es gibt mithin pragmatische, aber auch theoretische Motive einer kausalen Betrachtung.

Gerade in den Wirtschaftswissenschaften laufen viele Kontroversen auf die Frage hinaus, ob eine Variable X (etwa Lohn- oder Geldmengensteigerung) "nur" eine passive Begleiterscheinung oder aber eine aktive (auslösende) Ursache für Y (etwa Inflation) ist. Es ist deshalb eine wichtige Frage, ob und wie empirisch (statistisch) festgestellt werden kann, ob Y die Wirkung von X oder nur eine passive Begleiterscheinung ist.

Im folgenden soll versucht werden, den Kausalbegriff zu definieren und auf zwei Mißverständnisse über das Verhältnis zwischen Kausalität und Korrelation einzugehen, nämlich die Aussagen:

- man könne Kausalität positiv beweisen und (das andere Extrem)
- Korrelation und Kausalität habe nichts miteinander zu tun.

Es ist vergeblich, nach einer Methode zu suchen, mit der man allein auf statistische Daten gestützt, "beweisen" kann, dass X die Ursache von Y ist und nicht nur eine Begleiterscheinung, weil eine Kausalaussage nicht verifiziert, sondern nur falsifiziert werden kann.

Es war das Ziel von Hume und Mill, axiomatisch oder konstruktiv zu einer Festlegung darüber zu gelangen, wie ein empirischer Befund beschaffen sein muss, um auf Kausalität induktiv schließen zu können. Dabei wurde von der Induktion dieselbe Sicherheit verlangt wie von der Deduktion. Ein solches Programm ist zum Scheitern verurteilt. Wie gezeigt wurde (einleitend zu Def. 7.9) beweist ein hoher Betrag der Korrelation r_{xy} nicht, dass X die Ursache für Y ist (oder umgekehrt Y für X), weil dies auch Ergebnis einer Scheinkorrelation sein kann. Hierauf wird auch in Kapitel 8 eingegangen (Gl. 8.37).

Kausalität "beweisen", hieße ausschließen zu können, dass irgendein anderer als der vermutete Kausalzusammenhang für das Zustandekommen der Beobachtungen verantwortlich ist. Das ist in einer positiven, direkten Art nicht möglich, wohl aber kann man indirekt vorgehen. Wie jede andere Hypothese kann die Kausalhypothese dadurch und **nur** dadurch (indirekt) geprüft werden, dass man feststellt, ob der empirische Befund

nicht evtl. im Widerspruch steht zu den bei Geltung der Hypothese zu erwartenden Beobachtungen. Dabei stellen sich aber zwei Fragen:

1. Kann man ein für das praktische Handeln ausreichend sicheres Urteil über eine Kausalhypothese erreichen?
2. Inwiefern können statistische Betrachtungen hierzu beitragen?

zu 1.:
Kausalität kann zwar als Allsatz ("Immer dann, wenn") streng genommen nie verifiziert sondern nur falsifiziert werden. Eine positive Aussage ist trotzdem aus praktischen Gründen oft notwendig. Sie kann aber stets nur vorläufig und unsicher sein. Das damit verbundene Problem kann auch ethischer Natur sein, nicht nur methodischer: Wieviele Menschen müssen z.B. durch Rauchen gestorben sein, bis die Hypothese der Schädlichkeit des Rauchens annehmbar ist?

zu 2.:
Mit der Korrelation r_{xy} kann die Existenz einer (kausalen) Beziehung nicht bewiesen werden, da r_{xy} stets die Summe direkter und indirekter (oder auch nur indirekter, wie bei der Scheinkorrelation!) Einflüsse zwischen X und Y ist.

Im Vorgriff auf Kapitel 8 sei folgendes Modell der multiplen Regression angenommen.

$$Y = b_{yx}X + b_{yz}Z + u_y$$

Hierbei sind alle Variablen standardisiert (also auch $s_x^2 = s_z^2 = 1$), die Koeffizienten b_{yx} und b_{yz} sind sind somit auch die standardisierten Regressionskoeffizienten und die Störgröße u_y ist mit den Regressoren X und Z nicht korrelliert. Man erhält dann

$$(7.61) \quad r_{xy} = b_{yx} + b_{yz}\, r_{xz} \ .$$

Hier misst b_{yx} den direkten und b_{yz} den indirekten (d.h. den über Z vermittelten) Einfluss von X auf Y.

Die in den drei Pfeilschemen der Abb. 7.4 dargestellten Kausalmodelle stellen sich dann wie folgt dar; z.B. beim linken Bild:

$$X = b_{xz}\, Z + u_x$$
$$Y = b_{yz}\, Z + u_y$$

Daraus folgt, wenn u_x und u_y mit Z und auch untereinander nicht korreliert sind,

$$(7.62) \quad r_{xy} = b_{xz}\, b_{yz} = r_{xz}\, r_{yz}.$$

Es gibt jetzt keinen (sich in b_{yx} ausdrückenden) direkten Einfluß von X auf Y, gleichwohl ist aber $|r_{xy}| > 0$.
Entsprechend erhält man bei einer Kausalstruktur, wie man sie im mittleren Teil der Abb. 7.4 dargestellt ist

$$Z = b_{zx}\, X + u_z \qquad \text{so dass } r_{xz} = b_{zx}$$

$Y = b_{yz} Z + u_y$ so dass $r_{yz} = b_{yz}$

woraus folgt, wenn u_y und u_z nicht miteinander korreliert sind

$r_{xy} = b_{yz} r_{xz} = r_{yz} r_{xz}$

wie Gl. 7.62, d.h. beide Pfeilschemen (und auch das dritte) der Abb. 7.4 sind hinsichtlich ihrer beobachtbaren Konsequenz, nämlich Gl. 7.62 nicht unterscheidbar. Sie sind aber unterscheidbar vom Regressionsmodell $x \rightarrow y \leftarrow z$, das zu gl. 7.61 führt. Betrachtungen dieser Art sind Gegenstand der Pfadanalyse, auf die hier nicht weiter eingegangen werden kann.

Aus dem gleichen Grunde ist es auch falsch anzunehmen, man könne über die Richtung der Kausalität (X als Ursache von Y statt Y als Ursache von X) damit entscheiden, ob z.B. die verzögerte Variable X_{t-1} mit Y_t stärker korreliert ist als Y_{t-1} mit X_t (dann $X \rightarrow Y$) oder umgekehrt (dann $Y \rightarrow X$). Hinzu kommt: Weder ist bei allen kausalen Vorgängen diese Asymmetrie von Ursache und Wirkung zu fordern, noch ist diese eindeutig aus der zeitlichen Reihenfolge oder aus den Ergebnissen der - diese Asymmetrie nicht voraussetzenden - Korrelationsanalyse zu folgern.

Andererseits gilt aber: wenn eine bestimmte Kausalstruktur angenommen wird (vgl. Abb. 7.4), dann müßten bestimmte Beziehungen für die Korrelation folgen, etwa (in den drei Fällen von Abb. 7.4) $r_{xy} = r_{xz} r_{yz}$. Trifft dies bei den empirischen Beobachtungen nicht zu, so könnte diese Kausalhypothese verworfen werden.

Das Konzept der Kausalität weist über die bloße Beobachtung hinaus, denn es muss Bezug nehmen auf eine theoretische Fundierung und zwar aus den folgenden drei Gründen:

1. Erklärung:

Oft wird unterschieden zwischen Gesetzmäßigkeit und (evtl. zufälliger) Regelmäßigkeit, je nachdem, ob die beobachteten Zusammenhänge deduktiv in Verbindung mit Sätzen einer Theorie gebracht werden können oder ob dies (noch) nicht der Fall ist. Vom Standpunkt der Beobachtung kann meist zwischen den beiden Arten der Regelmäßigkeit nicht unterschieden werden.

2. Sprachgebundenheit:

In jedem Fall muss sich die Erklärung der Sprache bedienen, d.h. ihre Gültigkeit ist nicht unabhängig davon, wie die Ereignisse oder Variablen bezeichnet und operationalisiert werden, die kausal verknüpft werden. Je enger z.B. das (ursächliche) Ereignis definiert wird, desto kleiner ist die Beobachtungsbasis und Geltungsdauer für den vermuteten Kausalzusammenhang: Im Extremfall mag man die historische Einmaligkeit des Ereignisses behaupten, weil so viele Aspekte einer Erscheinung als "wesentlich" erscheinen, dass die Vergleichbarkeit ausgeschlossen ist. Dann gibt es natürlich keine Möglichkeit, auf Regelmäßigkeiten zu schließen.

3. Modellbildung:

Das Kausalkonzept weist stets ins Unendliche: Jede Ursache hat unendlich viele Wirkungen und selbst wieder unendlich viele "tiefere" oder "letzte" Ursachen. Es ist eine triviale und nutzlose "Erkenntnis", dass alles irgendwie mit allem zusammenhängt. Eine Theorie hat deshalb die Aufgabe, zu einem überprüfbaren Modell zu gelangen durch Ausschluß bestimmter denkbarer Kausalbeziehungen einerseits und durch Festlegung der Art und der Richtungen der Verursachung zwischen ausgewählten Variablen andererseits (causal ordering). Das geschieht z.B. durch die Annahme linearer stochastischer Beziehungen nach Art der oben betrachteten Pfeilschemen von Abb. 7.4.

Die Kennzeichen der Kausalität lassen sich somit in folgender Definition zusammenfassen:

Def. 7.23: Kausalität

Eine Kausalbeziehung zwischen Ereignissen bzw. Variablen X und Y dergestalt, dass X die Ursache von Y ist, bedeutet:

1. Eine Änderung von X "bewirkt" systematisch auch eine Veränderung von Y (Produktionsaspekt).
2. Die Beziehung ist i.d.R. asymmetrisch, d.h. ist X die Ursache von Y, so ist nicht gleichzeitig Y die Ursache von X.
3. Für den Fall, dass ein Experiment nicht durchführbar ist, kann keine Methodik gefunden werden, die es erlaubt, ohne die Interpretationshilfe einer Theorie von bestimmten Daten auf einen Kausalmechanismus zu schließen, d.h. diesen zu beweisen. Man kann aber falsifizierend vorgehen und denkbare Kausalhypothesen ausschließen und zwar auch bei Beobachtungsdaten. Dabei ist auch der Vergleich von beobachteten und erwarteten Korrelationen bedeutsam.

Ob ein statistisch gemessener Zusammenhang kausal interpretiert werden darf oder nicht, ist somit i.d.R. nicht allein aus den Daten zu erkennen.

Bemerkungen zu Def. 7.22:

1. Es ist sinnvoll, zwischen Kausalität in bezug auf Ereignisse (0-1-Variablen; vgl. Kap. 9) und Kausalität in bezug auf Variablen zu unterscheiden.

2. Es wird nicht gefordert, dass ein eindeutiger (funktionaler) Zusammenhang besteht: aus X=x folgt Y=y, wohl aber sollte der mit dem Bestimmtheitsmaß gemessene Anteil der systematischen (mit X erklärten) Variation an der Gesamtvariation von Y (in einem nicht näher bestimmten Maße) beträchtlich sein.

3. Mit der Asymmetrie ist die kausal nicht interpretierbare Interdependenz (feed back) ausgeschlossen. Die zeitliche Folge von Ursache und Wirkung ist nur insofern bedeutsam, als sie eine Möglichkeit bietet, sich empirisch der Asymmetrie zu

vergewissern. Sie ist für sich genommen genauso wenig ein Beweis für Kausalität wie eine gelungene Prognose (post hoc ergo propter hoc - Fehlschluß). Die zeitliche Abfolge ist oft nur der einzige Anhaltspunkt für die Existenz einer Kausalkette. Sie empirisch nachzuweisen ist als solches bereits ein schwieriges methodisches Problem. Eine Methode zur Überprüfung von Kausalhypothesen kann, muss aber nicht notwendig, eine explizite Zeitvariable vorsehen.

4. Eine Methode, sich empirisch des Produktionsaspektes und der Asymmetrie zu vergewissern, ist das Experiment, d.h. die alleinige Variation von X bei Konstanz aller übrigen Einflüsse. Es ist offensichtlich, dass die Ursächlichkeit von X für irgendeine Wirkung von Y nur demonstriert werden kann, wenn X variiert. Ist X konstant, so ist immer Unkorreliertheit gegeben. Eine Konstante X scheidet als erkennbare Ursache (aber auch als Wirkung) aus: eine Konstante ist der Erklärung weder fähig noch bedürftig.

5. Eine Theorie hat eine Erklärung und ein falsifizierbares Modell (causal ordering) zu liefern. Auf die statistischen Methoden zur Behandlung solcher Modelle (z.B. Pfadanalyse) kann hier nicht eingegangen werden.