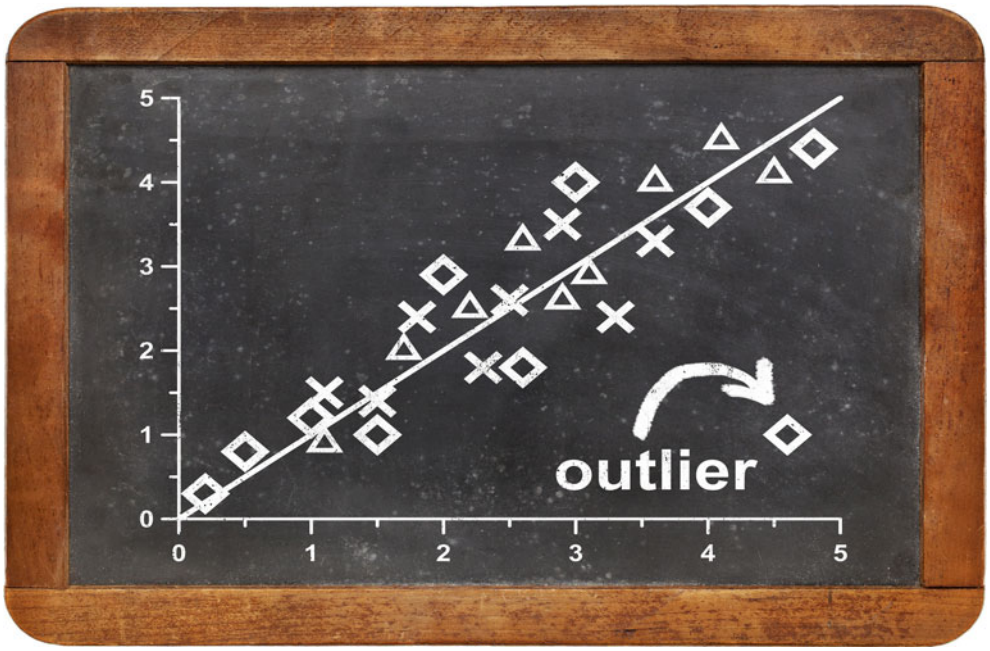
16 Zweidimensionale deskriptive Statistik – den Zusammenhang zwischen zwei Merkmalen beschreiben



Wie überprüft man den Zusammenhang zwischen zwei Merkmalen?

Was versteht man unter Regression?

Welche Rolle spielt die Korrelation?

Teil IV

L. Göllmann et al., *Mathematik für Ingenieure: Verstehen – Rechnen – Anwenden*, DOI 10.1007/978-3-662-53867-8_16

Solange wir nur ein Merkmal betrachten, haben wir jetzt eine Vorstellung, wie man die Daten darstellen und analysieren kann. Interessant wird es aber, wenn man zwei (oder sogar noch mehr) Merkmale gleichzeitig betrachtet und sich für Wechselwirkungen zwischen den Merkmalen, z. B. mögliche Einflüsse, interessiert.

Auch bei diesem Aufgabengebiet kommt es immer wieder darauf an, mit welcher Art Merkmale man es zu tun hat. Daher lernen wir Methoden für nominale, ordinale und quantitative Merkmale kennen.

16.1 Zusammenhangsuntersuchung bei nominalen Merkmalen

In diesem Abschnitt interessieren wir uns für zwei nominale Merkmale, also die Merkmale, mit denen man gar nicht rechnen kann, und wir möchten wissen, ob es einen Zusammenhang zwischen den beiden Merkmalen gibt und, wenn ja, wie stark er ist.

Kontingenztabellen liefern einen Überblick über zwei qualitative Merkmale

Das Einzige, was man mit nominalen Merkmalen tun kann, ist, die Häufigkeiten ihrer Ausprägungen zu betrachten. Daher erstellen wir zuerst eine **Kontingenztabelle**, in der die Häufigkeiten der einzelnen Ausprägungen beider Merkmale eingetragen werden. Dies geschieht nicht isoliert, sondern zweidimensional, d. h., man trägt die Häufigkeiten der betrachteten Merkmalsträger ein, die gleichzeitig bei Merkmal A die Ausprägung A_i und bei Merkmal B die Ausprägung B_j besitzen.

Definition: Kontingenztabelle

Eine Kontingenztabelle betrachtet zwei Merkmale gleichzeitig und stellt die **gemeinsamen Häufigkeiten** (die je zwei Merkmalsausprägungen betreffen) und die **Randhäufigkeiten**, die nur ein Merkmal erfassen, übersichtlich dar. Kontingenztabellen werden sowohl mit absoluten als auch mit relativen Häufigkeiten erstellt.

In einer Kontingentabelle mit absoluten Häufigkeiten bezeichnet n_{ij} die absolute gemeinsame Häufigkeit, mit der bei Merkmal A die Ausprägung A_i und gleichzeitig bei Merkmal B die Ausprägung B_j auftrat. $n_{i\bullet}$ bzw. $n_{\bullet j}$ bezeichnet die absolute Randhäufigkeit, mit der bei Merkmal A die Ausprägung A_i bzw. bei Merkmal B die Ausprägung B_j auftrat, ohne dass man sich für das jeweils andere Merkmal interessiert. Man berechnet diese Randhäufigkeit, indem man die zugehörigen gemeinsamen Häufigkeiten addiert. n bezeichnet den Stichprobenumfang, den man erhält, indem man die Randhäufigkeiten eines Merkmals addiert.

	B_1	B_2	...	B_l	Randhäufigkeit Merkmal A
A_1	n_{11}	n_{12}	...	n_{1l}	$n_{1\bullet}$
A_2	n_{21}	n_{22}	...	n_{2l}	$n_{2\bullet}$
...	...	...	...	...	...
A_k	n_{k1}	n_{k2}	...	n_{kl}	$n_{k\bullet}$
Randhfk. B	$n_{\bullet 1}$	$n_{\bullet 2}$	...	$n_{\bullet l}$	n

In einer Kontingenztabelle mit relativen Häufigkeiten bezeichnet h_{ij} die relative gemeinsame Häufigkeit, mit der bei Merkmal A die Ausprägung A_i und gleichzeitig bei Merkmal B die Ausprägung B_j auftrat. $h_{i\bullet}$ bzw. $h_{\bullet j}$ bezeichnet die relative Randhäufigkeit, mit der bei Merkmal A die Ausprägung A_i bzw. bei Merkmal B die Ausprägung B_j auftrat, ohne dass man sich für das jeweils andere Merkmal interessiert. Man berechnet diese Randhäufigkeit, indem man die zugehörigen gemeinsamen Häufigkeiten addiert. Wenn man die Randhäufigkeiten eines Merkmals addiert, muss das Ergebnis 1 sein.

	B_1	B_2	...	B_l	Randhäufigkeit Merkmal A
A_1	h_{11}	h_{12}	...	h_{1l}	$h_{1\bullet}$
A_2	h_{21}	h_{22}	...	h_{2l}	$h_{2\bullet}$
...	...	...	...	...	...
A_k	h_{k1}	h_{k2}	...	h_{kl}	$h_{k\bullet}$
Randhfk. B	$h_{\bullet 1}$	$h_{\bullet 2}$	...	$h_{\bullet l}$	1

Schauen wir uns direkt ein Beispiel zu den Kontingenztabellen an.

Beispiel

Prinzipiell kann man Kontingenztabellen für alle Arten von Merkmalen darstellen. Einzige Ausnahme sind quantitativ stetige bzw. diskrete unendliche Merkmale, denn dafür bräuchte man unendlich viele Zeilen und Spalten. Da es für ordinale und quantitative Merkmale aber bessere Zusammenhangsmaße gibt, schauen wir uns zwei nominale Merkmale an, da es für diese keine bessere Methode gibt. Die klassischen nominalen Beispielsmerkmale sind Haarfarbe und Augenfarbe. Also betrachten wir eine Stichprobe von 100 Studierenden, bei denen gleichzeitig Haarfarbe (Merkmal *HF*) und Augenfarbe (Merkmal *AF*) erhoben wurde.

Folgendermaßen sah das Ergebnis (mit absoluten Häufigkeiten) aus:

		Augenfarbe (AF) blau	braun	grün	Randhfk. HF
Haar-	braun	12	23	8	43
farbe	blond	19	4	15	38
(HF)	rot	2	1	5	8
	schwarz	1	8	2	11
Randhäufigkeit AF		34	36	30	100

Man kann an der Tabelle erkennen, dass insgesamt 100 Studierende befragt wurden. Vier von ihnen hatten blonde Haare und braune Augen; insgesamt waren 30 Studierende grünäugig.

Und so sieht die Kontingenztabelle mit relativen Häufigkeiten aus:

		Augenfarbe (AF) blau	braun	grün	Randhfk. HF
Haar-	braun	0.12	0.23	0.08	0.43
farbe	blond	0.19	0.04	0.15	0.38
(HF)	rot	0.02	0.01	0.05	0.08
	schwarz	0.01	0.08	0.02	0.11
Randhäufigkeit AF		0.34	0.36	0.30	1

◀

Bedingte Häufigkeiten berechnen die relativen Häufigkeiten des einen Merkmals bezogen auf eine Ausprägung des anderen Merkmals

Bei der gleichzeitigen Betrachtung zweier Merkmale gibt es ein interpretatorisches Problem. Schauen wir uns dazu noch einmal das obige Beispiel an, und zwar die Studenten mit grünen Augen und roten Haaren. Davon gab es fünf, d. h., der Anteil betrug 0.05 oder 5 %, was sicherlich nicht viel ist. Wenn man jetzt aber nur die Rothaarigen betrachtet, so muss man zugeben, dass viele rothaarige Studierende grüne Augen besitzen, nämlich 62.5 %. Das erscheint doch schon recht viel. Betrachtet man umgekehrt nur die Menschen mit grünen Augen, so sind von diesen 30 Menschen fünf rothaarig, was einem Anteil von 16.7 % entspricht.

Was das Beispiel zeigen soll, ist, dass man neben den gemeinsamen Häufigkeiten („Wie viele weisen Merkmalsausprägung A_i und Merkmalsausprägung B_j gleichzeitig auf?“) noch eine weitere Häufigkeitsart kennen muss – die **bedingten Häufigkeiten** („Wie viele von denjenigen, die bereits Merkmalsausprägung A_i aufweisen, weisen Merkmalsausprägung B_j auf?“).

Achtung Bedingte Häufigkeiten sind etwas völlig anderes als gemeinsame Häufigkeiten, und (was es noch schlimmer macht) Psychologen haben festgestellt, dass Menschen mit bedingten Häufigkeiten ein Problem haben. (Näheres findet man z. B. unter dem Stichwort „conditional probability fallacy“ bei Von Nitzsch, R. (2015): „Entscheidungslehre“, 8. Auflage (Aachen)). ◀

Viele Menschen können die beiden bedingten Häufigkeiten nicht auseinanderhalten. Der Anteil derjenigen Objekte, die Merkmalsausprägung B_j aufweisen, an denen, die Merkmalsausprägung A_i aufweisen, und der Anteil derjenigen Objekte, die Merkmalsausprägung A_i aufweisen, an denen, die Merkmalsausprägung B_j aufweisen, sind etwas völlig verschiedenes. Das obige Beispiel zeigt deutlich, dass es einen Unterschied gibt, denn 62.5 % und 16.7 % sind einfach unterschiedliche Zahlen!

Definition: Bedingte Häufigkeiten, bedingte (Häufigkeits-)Verteilung

Es sei $n_{\bullet j} > 0$.

Dann heißt $h_{A=A_i|B=B_j} = \frac{n_{ij}}{n_{\bullet j}} = \frac{h_{ij}}{h_{\bullet j}}$ die bedingte Häufigkeit von $A = A_i$ unter der Bedingung, dass $B = B_j$.

Das geht natürlich auch umgekehrt:

Es sei $n_{i\bullet} > 0$.

Dann heißt $h_{B=B_j|A=A_i} = \frac{n_{ij}}{n_{i\bullet}} = \frac{h_{ij}}{h_{i\bullet}}$ die bedingte Häufigkeit von $B = B_j$ unter der Bedingung, dass $A = A_i$.

Die Auflistung aller bedingten Häufigkeiten heißt **bedingte Häufigkeitsverteilung**.

Die bedingten Häufigkeitsverteilungen stellt man am besten auch tabellarisch dar. Das sehen wir uns mal an dem Beispiel von oben an:

Beispiel

Zuerst betrachten wir die bedingte Häufigkeitsverteilung, bei der Merkmal B, also die Augenfarbe feststeht.

		Augenfarbe (AF) blau	braun	grün
Haar-	braun	0.353	0.639	0.267
farbe	blond	0.559	0.111	0.5
(HF)	rot	0.059	0.028	0.167
	schwarz	0.029	0.222	0.067
Summe		1	1	1

Und nun die umgekehrte bedingte Häufigkeitsverteilung, bei der das erste Merkmal, also die Haarfarbe bereits fest-

steht:

		Augenfarbe (AF)			
		blau	braun	grün	Summe
Haar-	braun	0.279	0.535	0.186	1
farbe	blond	0.5	0.105	0.395	1
(HF)	rot	0.25	0.125	0.625	1
	schwarz	0.091	0.727	0.182	1

◀

Empirische Unabhängigkeit zweier Merkmale basiert auf bedingten Häufigkeiten

Bevor wir die Abhängigkeit messen können, müssen wir erstmal festlegen, was eigentlich Unabhängigkeit bei zwei (in der Regel nominalen) Merkmalen bedeuten soll. Wir brauchen die bedingten Häufigkeiten, um die Idee der Abhängigkeit zu erklären. Man muss sich übrigens unbedingt klarmachen, dass wir in der Statistik nur **empirische Abhängigkeiten** betrachten können, also Abhängigkeiten, die aufgrund der Häufigkeitsverteilungen feststellbar sind. Wir können in der Statistik niemals inhaltliche Abhängigkeiten berechnen, sondern nur gefundene empirische Abhängigkeiten näher analysieren und dann mit Fachexperten überlegen, welche inhaltlichen Gründe es für dieses empirische (beobachtbare) Phänomen gibt.

Aber zurück zu den bedingten Häufigkeiten und ihrer Bedeutung für die Unabhängigkeit. Man würde sich sicherlich davon überzeugen lassen, dass die Augenfarbe eines Menschen unabhängig von seiner Haarfarbe ist, wenn die Häufigkeitsverteilung der Augenfarben bei Blonden genauso aussieht wie bei den Braunhaarigen und wie bei den Schwarzhaarigen etc. Dann würde man schließen, dass Haarfarbe und Augenfarbe keinen (inhaltlichen und empirischen) Zusammenhang aufweisen.

Das formuliert man mathematisch wie folgt:

Definition: Empirische Unabhängigkeit

Zwei Merkmale A und B heißen **empirisch unabhängig**, wenn die bedingten Häufigkeiten $h_{A=A_i|B=B_j}$ gleich sind, egal welche Ausprägung das Merkmal B annimmt, d. h. $h_{A=A_i|B=B_{j1}} = h_{A=A_i|B=B_{j2}}$.

Außerdem muss gelten: Die bedingten Häufigkeiten $h_{B=B_j|A=A_i}$ sind gleich, egal welche Ausprägung das Merkmal A annimmt, d. h. $h_{B=B_j|A=A_{i1}} = h_{B=B_j|A=A_{i2}}$.

Es kann sehr mühsam sein, immer die bedingten Häufigkeiten auszurechnen, um beurteilen zu können, ob zwei Merkmale empirisch unabhängig sind. Das ist auch nicht nötig. Man kann die Unabhängigkeit auch in der Kontingenztabelle sehen.

Empirische Unabhängigkeit

Zwei Merkmale A und B sind empirisch unabhängig, wenn für die relativen Häufigkeiten gilt:

$$h_{ij} = h_{i\bullet} \cdot h_{\bullet j} \text{ für alle } i \in \{1, \ldots, k\} \text{ für alle } j \in \{1, \ldots, l\}$$

bzw. wenn für die absoluten Häufigkeiten gilt:

$$n_{ij} = \frac{n_{i\bullet} \cdot n_{\bullet j}}{n} \text{ für alle } i \in \{1, \ldots, k\} \text{ für alle } j \in \{1, \ldots, l\}.$$

Das bedeutet also, dass im Falle der empirischen Unabhängigkeit die gemeinsame Häufigkeit durch die beiden Randhäufigkeiten vollständig bestimmt wird. In diesem Fall sind übrigens die bedingten Häufigkeiten und die „unbedingten" Häufigkeiten identisch.

Achtung Falls in einer Kontingenztabelle bei den gemeinsamen Häufigkeiten (also den Zahlen in der Mitte) Nullen auftreten, kann man sich darauf verlassen, dass die Merkmale abhängig sind. Denn gemäß dem berühmten Satz vom Nullprodukt müsste für die obige Gleichung dann auch mindestens eine Randhäufigkeit 0 sein, was aber keinen Sinn macht, da dies bedeuten würde, dass es diese Merkmalsausprägung nicht gibt. Warum sollte sie dann in der Kontingenztabelle auftauchen? ◀

Den Grad empirischer Abhängigkeit kann man messen

Aber wie sieht es aus, wenn die beiden Merkmale eben nicht unabhängig sind? Dann kann man den Grad der Abhängigkeit tatsächlich messen. Basis für die Berechnung ist hierbei die **χ^2-Größe**. (Das spricht sich: Chi-Quadrat-Größe.)

Definition: χ^2-Größe

Bei positiven Randhäufigkeiten ist die χ^2-Größe wie folgt definiert:

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{\left(n_{ij} - \frac{n_{i\bullet} \cdot n_{\bullet j}}{n}\right)^2}{\frac{n_{i\bullet} \cdot n_{\bullet j}}{n}}$$

Wenn man unbedingt mit dieser Formel rechnen will, sollte man sich eine separate Tabelle für die $\frac{n_{i\bullet} \cdot n_{\bullet j}}{n}$-Werte anlegen. Aber es gibt eine einfachere Möglichkeit, die χ^2-Größe zu berechnen.

Satz (Berechnung der χ^2-Größe)
Es gilt:

$$\chi^2 = n \cdot \left(\sum_{i=1}^{k} \sum_{j=1}^{l} \frac{n_{ij}^2}{n_{i\bullet} \cdot n_{\bullet j}} \right) - n$$ ◀

Das kann man relativ einfach berechnen, und wir betrachten das direkt an unserem Beispiel:

Beispiel

Im obigen Beispiel berechnet sich die χ^2-Größe wie folgt:

$$\begin{aligned}\chi^2 = 100 \cdot \Bigg(&\frac{12^2}{34 \cdot 43} + \frac{23^2}{36 \cdot 43} + \frac{8^2}{30 \cdot 43} + \frac{19^2}{34 \cdot 38} \\ &+ \frac{4^2}{36 \cdot 38} + \frac{15^2}{30 \cdot 38} + \frac{2^2}{34 \cdot 8} + \frac{1^2}{36 \cdot 8} \\ &+ \frac{5^2}{30 \cdot 8} + \frac{1^2}{34 \cdot 11} + \frac{8^2}{36 \cdot 11} + \frac{2^2}{30 \cdot 11} \Bigg) - 100 \\ = 27.707092 \end{aligned}$$

◀

Das Problem besteht nun darin, diese Zahl zu interpretieren, und das ist dann schon schwieriger. Denn diese Größe kann prinzipiell jeden positiven Wert annehmen, und ob der Wert jetzt hoch oder niedrig ist, kann man nicht beurteilen. Sie ist also in dieser Gestalt denkbar ungeeignet, um damit etwas (insbesondere Abhängigkeit) zu messen. Außerdem hat man sich darauf geeinigt, dass man an Maßzahlen bestimmte Anforderungen stellt. Sie müssen positiv sein, eine Maßzahl muss zwischen 0 und 1 liegen, und sie muss umso größer sein, je stärker das, was man messen möchte, ausgeprägt ist. Außerdem soll die Maßzahl den Wert 1 annehmen, wenn das, was sie misst, maximal vorhanden ist, und sie soll den Wert 0 haben, wenn das, was sie misst, nicht vorhanden ist.

Aufgrund dieser Konvention müssen wir die χ^2-Größe noch umrechnen, bevor wir eine Maßzahl erhalten, die wir sinnvoll interpretieren können, nämlich den (korrigierten) Kontingenzkoeffizienten von Pearson.

Definition: Kontingenzkoeffizient von Pearson

Der **Kontingenzkoeffizient von Pearson** berechnet sich wie folgt:

$$K = \sqrt{\frac{\chi^2}{n + \chi^2}}$$

Dieser Kontingenzkoeffizient hat zwar schon fast alle gewünschten Eigenschaften, er wird aber maximal nicht genau 1, sondern es gilt:

$$0 \leq K \leq \sqrt{\frac{\min(k,l) - 1}{\min(k,l)}} < 1$$

Hierbei ist k die Anzahl der Ausprägungen des ersten Merkmals (Zeilen) und l die Anzahl der Ausprägungen des zweiten Merkmals (Spalten). ($\min(k,l)$ bedeutet hierbei, dass man von den beiden Zahlen die kleinere nehmen muss.)

K wird also niemals 1, sondern bleibt immer kleiner als 1. Da dies nicht erwünscht ist, müssen wir den Kontingenzkoeffizienten noch (geringfügig) korrigieren und erhalten endlich unsere Maßzahl:

Definition: Korrigierter Kontingenzkoeffizient von Pearson

Der **korrigierte Kontingenzkoeffizient von Pearson** K^* ist definiert durch:

$$K^* = K \cdot \sqrt{\frac{\min(k,l)}{\min(k,l) - 1}}$$

Und der erfüllt alle unsere Anforderungen:

Eigenschaften des korrigierten Kontingenzkoeffizienten

Der korrigierte Kontingenzkoeffizient von Pearson K^* besitzt folgende Eigenschaften:

- $0 \leq K^* \leq 1$
- $K^* = 0$, genau dann, wenn die beiden Merkmale empirisch unabhängig sind.
- $K^* = 1$, genau dann, wenn die Häufigkeiten in genau einem Feld konzentriert sind.
- Je größer K^* ist, desto stärker sind die beiden Merkmale voneinander abhängig.

Auch die beiden Kenngrößen K und K^* berechnen wir für das obige Beispiel:

Beispiel

Da die χ^2-Größe schon berechnet wurde, ist nicht mehr viel zu tun:

$$K = \sqrt{\frac{27.707092}{27.4329687 + 100}} = 0.4635788$$

und in der korrigierten Version:

$$K* = 0.465788 \cdot \sqrt{\frac{3}{2}} = 0.5704714$$

Damit können wir nun urteilen, dass in der vorgegebenen Stichprobe ein mittlerer Zusammenhang zwischen Augen- und Haarfarbe besteht. Diesen gilt es nun inhaltlich zu spezifizieren. Die Zahl alleine genügt auf keinen Fall. Sie sollten also erwähnen, dass unter den Blonden in erste Linie Blauäugige zu finden sind, dass die Braunhaarigen vorwiegend braune Augen haben u.ä.

Hieran kann man schön verstehen, dass zwischen empirischem Zusammenhang und inhaltlichem Zusammenhang ein großer Unterschied besteht. Denn die Statistik liefert nur einen Hinweis, dass hier ein Zusammenhang besteht, aber wie er begründet ist, das kann man nur herausfinden, wenn man sich mit Biologie beschäftigt. ◀

16.2 Zusammenhangsuntersuchung bei ordinalen Merkmalen

Mit nominalen Daten können wir nun umgehen. Was gibt es aber noch für Möglichkeiten, wenn die Daten ein höheres Skalenniveau besitzen, d. h., zumindest ordinal skaliert sind? Dann hat man ja die zusätzliche Information, dass die Ausprägungen sinnvoll geordnet werden können. Und diese Information sollte man auch benutzen. (Man kann es natürlich auch ignorieren und einfach eine Kontingenztabelle erstellen und den Kontingenzkoeffizienten berechnen, aber mehr Inputinformationen liefern tendenziell auch ein besseres Outputergebnis.)

Die Lösung ist der Rangkorrelationskoeffizient von Spearman.

Definition: Rangkorrelationskoeffizient von Spearman

Seien R_i die Ränge der Ausprägungen des ersten Merkmals und S_i die Ränge der Ausprägungen des zweiten Merkmals. $\overline{R}$ und $\overline{S}$ seien die Durchschnitte der jeweiligen Ränge.

Dann ist der **Rangkorrelationskoeffizient von Spearman** r_{Sp} wie folgt definiert:

$$r_{Sp} = \frac{\sum_{i=1}^{n} (R_i - \overline{R}) \cdot (S_i - \overline{S})}{\sqrt{\sum_{i=1}^{n} (R_i - \overline{R})^2} \cdot \sqrt{\sum_{i=1}^{n} (S_i - \overline{S})^2}}$$

Wichtig ist bei dieser Formel, dass man sich nie für die eigentlichen Messwerte interessiert, sondern nur für die Ränge, also die Platzierungen. Wenn gleiche Platzierungen auftreten, muss man daran denken, dass dann der nächste Platz unbesetzt ist. Also wenn es drei zweite Plätze gibt, dann gibt es keinen dritten und keinen vierten Platz und es geht weiter mit Platz 5.

Ein wenig einfacher kann man sich das Leben machen, wenn man zum Rechnen die folgende Formel benutzt:

Berechnungsformel für den Rangkorrelationskoeffizienten von Spearman

$$r_{Sp} = \frac{\overline{R \cdot S} - \overline{R} \cdot \overline{S}}{\sqrt{\overline{R^2} - (\overline{R})^2} \cdot \sqrt{\overline{S^2} - (\overline{S})^2}}$$

Diese Formel muss man erstmal lesen können. Das Meiste kennen wir schon: $\overline{R}$ und $\overline{S}$ sind die Mittelwerte der Ränge des ersten bzw. des zweiten Merkmals, $\overline{R^2}$ und $\overline{S^2}$ sind die Mittelwerte über die quadrierten Ränge der beiden Merkmale. Das einzig Neue ist $\overline{R \cdot S}$. Das bedeutet, dass wir für jeden Merkmalsträger die Ränge der beiden Merkmale multiplizieren müssen und den Mittelwert über diese Produkte bilden.

$$\overline{R \cdot S} = \frac{1}{n} \sum_{i=1}^{n} R_i \cdot S_i.$$

Auch diese Formel ist noch recht arbeitsaufwendig. Allerdings gibt es für einen speziellen Fall eine Vereinfachung, die aber nur gilt, wenn alle Messwerte eines Merkmals unterschiedlich sind, wenn also jede Platzierungsnummer (d. h. jeder Rang) pro Merkmal nur einmal auftritt.

Berechnungsformel für den Rangkorrelationskoeffizienten von Spearman bei unterschiedlichen Rängen

Wenn alle Merkmalsausprägungen der beiden betrachteten Merkmale unterschiedlich sind, lässt sich der Rangkorrelationskoeffizient von Spearman wie folgt berechnen:

$$r_{Sp} = 1 - \frac{6}{n(n^2 - 1)} \sum_{i=1}^{n} (R_i - S_i)^2$$

Beispiel

Wir betrachten zehn Maschinenbaustudenten, die im ersten Semester die Module Mathematik und Informatik absolviert haben und folgende Ergebnisse erzielt haben:

Student	1	2	3	4	5	6	7	8	9	10
Mathem.	2.3	4.0	3.3	2.7	1.3	1.7	5.0	2.0	3.0	1.0
Inform.	2.7	3.7	3.3	2.3	1.0	1.7	4.0	2.0	5.0	1.3

Die Tabelle der Ränge sieht dann wie folgt aus:

Student	1	2	3	4	5	6	7	8	9	10
Math. (R_i)	5	9	8	6	2	3	10	4	7	1
Info. (S_i)	6	8	7	5	1	3	9	4	10	2

Alle Ränge sind also unterschiedlich, und wir können die vereinfachte Formel anwenden:

$$\begin{aligned} r_{Sp} &= 1 - \frac{6}{10 \cdot 99} \left((-1)^2 + 1^2 + 1^2 + 1^2 + 1^2 \right. \\ &\qquad \left. + 0^2 + 1^2 + 0^2 + (-3)^2 + (-1)^2\right) \\ &= 1 - \frac{6}{990} \cdot 16 = 1 - 0.0969 = 0.9031 \end{aligned}$$

Verwenden wir die allgemeine Berechnungsformel, erhalten wir $\overline{R} = 5.5 = \overline{S}$ sowie $\overline{R \cdot S} = 37.7$, $\overline{R^2} = 38.5$ und $\overline{S^2} = 38.5$. Also berechnet sich der Rangkorrelationskoeffizient von Spearman wie folgt:

$$\begin{aligned} r_{Sp} &= \frac{37.7 - 5.5 \cdot 5.5}{\sqrt{38.5 - 5.5^2} \cdot \sqrt{38.5 - 5.5^2}} \\ &= \frac{7.45}{\sqrt{8.25} \cdot \sqrt{8.25}} \\ &= 0.9031 \end{aligned}$$

◀

Auch der Rangkorrelationskoeffizient von Spearman besitzt einige Eigenschaften, die man sich merken sollte.

Eigenschaften des Rangkorrelationskoeffizienten

- Der Rangkorrelationskoeffizient von Spearman liegt immer zwischen -1 und 1, d. h. $-1 \leq r_{Sp} \leq 1$.
- Wenn r_{Sp} positiv ist, dann liegt ein positiver Zusammenhang vor, d. h., je besser die Ausprägungen beim ersten Merkmal sind, desto besser sind sie auch beim zweiten Merkmal. Im Extremfall gilt $r_{Sp} = 1$, dann sind alle Ränge des ersten Merkmals identisch zu den Rängen des zweiten Merkmals, d. h. der erstplatzierte Wert beim ersten Merkmal war auch erster beim zweiten Merkmal usw.
- Wenn r_{Sp} negativ ist, dann liegt ein negativer Zusammenhang vor, d. h., je besser die Ausprägungen beim ersten Merkmal sind, desto schlechter sind sie beim zweiten Merkmal und umgekehrt. Im Extremfall gilt $r_{Sp} = -1$, dann sind die Ränge genau entgegengesetzt, d. h. der Erstplatzierte beim ersten Merkmal war Letzter beim zweiten Merkmal, der Zweitplatzierte beim ersten Merkmal der Vorletzte beim zweiten usw.
- Wenn $r_{Sp} = 0$, dann geht man davon aus, dass zwischen den beiden ordinalen Merkmalen kein Zusammenhang besteht, zumindest keiner der Form „je mehr ..., desto mehr ..." bzw. „je mehr ..., desto weniger ...".
- Der Rangkorrelationskoeffizient von Spearman r_{Sp} verändert sich nicht, wenn die Merkmale beide entweder monoton steigend oder monoton fallend transformiert werden.

16.3 Zusammenhangsuntersuchung bei quantitativen Merkmalen

Quantitative Merkmale können am genauesten untersucht werden, weil sie die meisten Informationen beinhalten. Bei diesen Merkmalen kann man nicht nur die Stärke des Zusammenhangs angeben, sondern man kann die Art des Zusammenhangs analysieren, hier gibt es nämlich Funktionen, die den Zusammenhang beschreiben.

Das Streuungsdiagramm liefert eine Vermutung zur Art des Zusammenhangs

Der erste Analyseschritt besteht meistens darin, die gegebenen Merkmalsausprägungen in ein xy-Diagramm einzuzeichnen, um eine Vermutung zu entwickeln, welcher Funktionstyp den Zusammenhang zwischen den Merkmalen am besten beschreibt. Dieses Diagramm nennt man **Streuungsdiagramm**. Mit der Entscheidung, welches Merkmal man x und welches man y nennt, hat man bereits festgelegt, welches Merkmal welches beeinflusst. Denn wir sagen ja immer (meistens), dass $y = f(x)$ ist, und das bedeutet, dass wir davon ausgehen, dass y von x beeinflusst wird und nicht umgekehrt. Diese Entscheidung nimmt einem kein statistisches Verfahren ab, das kann man nur aus Plausibilitätsgründen oder theoriegestützt festlegen.

Beispiel

Bei zehn Maschinenbaustudenten wurde zusammen mit der Körpergröße auch das Gewicht erhoben. Die folgende Tabelle zeigt das Ergebnis:

Student	Größe	Gewicht
1	1.56	58
2	1.67	63
3	1.68	74
4	1.72	74
5	1.86	95
6	1.92	78
7	1.56	49
8	1.67	62
9	1.72	80
10	1.53	50

Das zugehörige Streuungsdiagramm sieht dann wie in Abb. 16.1 aus.

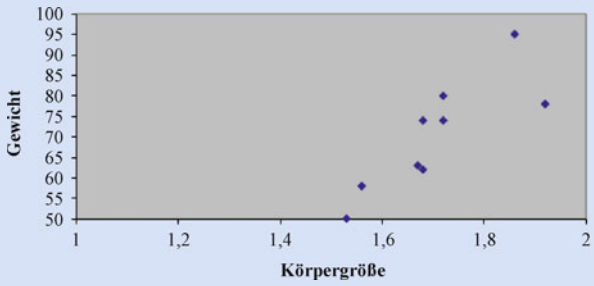


Abb. 16.1 Beispiel eines Streuungsdiagramms ◀

Die Methode der kleinsten Fehlerquadrate liefert die Idee zur Funktionsbestimmung

Wenn man keine Idee hat, welche Funktion den Zusammenhang am besten darstellt, ist das Problem schwer zu lösen. Man sollte daher zumindest eine Funktionsart im „Verdacht" haben, z. B. eine Gerade, eine Parabel oder eine e-Funktion. Dann erscheint es einigermaßen logisch, dass man etwaige zu bestimmende Parameter der Funktion so wählt, dass der Abstand zwischen den Messpunkten (x_i, y_i) und der Funktion $f(x)$ minimal wird. Genauer gesagt minimiert man die Summe der vertikalen Abstandsquadrate (Mal wieder aus dem Grund, dass die Abstände

teilweise positiv und teilweise negativ sind und die Summe also geringer ausfallen würde, als es gerechtfertigt wäre.), also

$$\sum_{i=1}^{n} (y_i - f(x_i))^2$$

Die Lösung dieses Minimierungsproblems erfordert Kenntnisse der Analysis mehrerer Variablen, die im Teil Analysis in Bd. 2 vermittelt werden. Dort findet man auch eine genauere Beschreibung der Methode der kleinsten Fehlerquadrate. Wir nehmen hier ein Resultat vorweg.

Die lineare Regression stellt den wichtigsten Spezialfall der Regression dar

Für den Fall, dass man begründet annimmt, dass zwischen den Messwerten ein linearer Zusammenhang herrscht, dass also $y_i \approx b \cdot x_i + a$ gilt, kann man nachweisen, dass die Methode der kleinsten Fehlerquadrate folgende Koeffizienten für die optimale Gerade ergibt.

Berechnung der Regressionsgeraden

Seien (x_i, y_i) mit $i = 1, \ldots, n$ gegebene Messwerte.

Dann ist $f(x) = \hat{b} \cdot x + \hat{a}$ die Gerade, die die Messwerte nach der Methode der kleinsten Fehlerquadrate am besten annähert, wenn

$$\hat{b} = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\overline{x^2} - (\overline{x})^2}$$

und

$$\hat{a} = \overline{y} - \hat{b} \cdot \overline{x}$$

Hierbei bedeutet $\overline{xy} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i \cdot y_i$. Wir müssen also für jede Messung das Produkt aus erstem und zweitem Messwert bilden, diese Produkte aufaddieren und zum Schluss durch die Anzahl der Messungen dividieren.

Logischerweise muss man zuerst $\hat{b}$ berechnen, bevor man $\hat{a}$ berechnen kann. (Übrigens heißen die beiden Parameter jetzt $\hat{a}$ und $\hat{b}$, weil man die Parameter geschätzt hat. Was das bedeutet, kommt in Abschn. 18.1 noch genauer.)

Im obigen Beispiel ergibt sich dann:

Beispiel

Zum Nachrechnen mit allen Zwischenergebnissen:

$$\overline{x} = 1.689, \quad \overline{y} = 68.3, \quad \overline{xy} = 116.783,$$
$$\overline{x^2} = 2.86711, \quad \overline{y^2} = 4853.9$$
$$\hat{b} = 98.98533602, \quad \hat{a} = -98.88623254$$

Also lautet die Regressionsgerade: $y = 98.98533602 \cdot x - 98.88623254$.

Grafisch ist dies in Abb. 16.2 dargestellt.

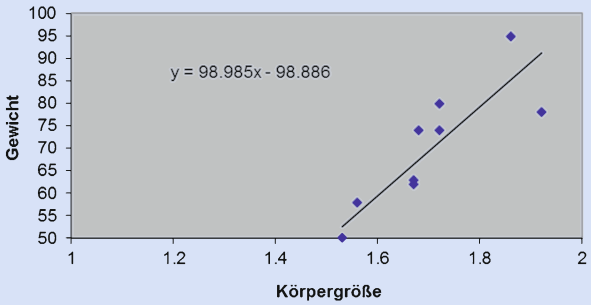


Abb. 16.2 Beispiel einer Regressionsgeraden ◀

Zusammenhangsmaße bei quantitativen Daten messen die Güte des Regressionsmodells

Die **Regressionsgerade** können wir nun zeichnen. Wir müssen uns aber fragen, wie wir beurteilen können, ob eine Gerade tatsächlich den realen Zusammenhang zwischen den Messwerten widerspiegelt. Manchmal drängt sich der Verdacht auf, dass in Wirtschaft und Wissenschaft oft mit **linearer Regression** gearbeitet und argumentiert wird, weil es das einzige Verfahren ist, das man kennt und vielleicht auch verstanden hat. Trotzdem sollte man immer genau verbal und anhand von theoretisch basierten Überlegungen argumentieren, dass ein linearer Zusammenhang für die vorliegende Problemstellung auch tatsächlich sinnvoll ist. Denn auch in der folgenden Situation können wir eine Regressionsgerade berechnen, aber wie sinnvoll ist das?

Beispiel

Wir betrachten die folgenden Messwerte:

Messwert	x	y
1	0	8
2	1	5
3	2	4
4	3	5
5	4	8

Das zugehörige Streuungsdiagramm ist in Abb. 16.3 dargestellt:

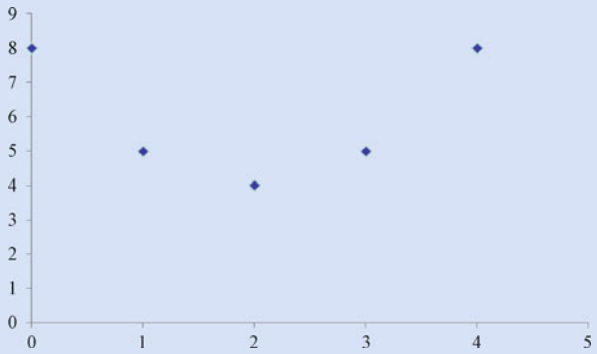


Abb. 16.3 Beispiel eines Streuungsdiagramms mit quadratischem Zusammenhang

Mit bloßen Augen erkennt man, dass es hier einen quadratischen Zusammenhang gibt, aber wenn man sich das Streuungsdiagramm nicht ansieht, ist die Gefahr groß, dass man ohne nachzudenken eine lineare Regression durchführt. Und statistische Verfahren sind hier gnadenlos. Sie funktionieren einwandfrei, auch wenn die Anwendung vollkommen idiotisch ist. In obigem Beispiel erhält man:

$$\overline{x} = 2, \quad \overline{y} = 6, \quad \overline{xy} = 12, \quad \overline{x^2} = 6, \quad \overline{y^2} = 38.8,$$
$$\hat{b} = 0, \quad \hat{a} = 6$$

Also lautet die Regressionsgerade: $y = 6$.

Grafisch sieht das Ganze wie in Abb. 16.4 aus:

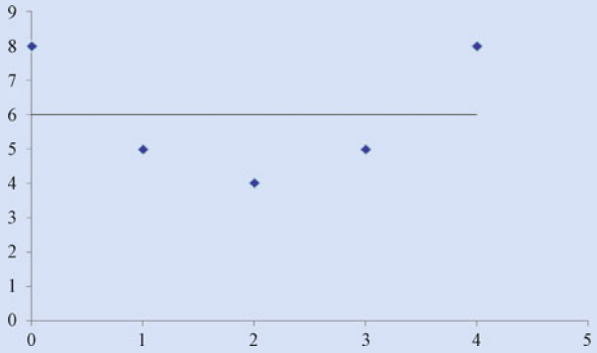


Abb. 16.4 Beispiel einer Regressionsgeraden bei quadratischem Zusammenhang

◀

Mathematisch kann man die Argumentation durch die Berechnung einer Maßzahl zur Messung des linearen Zusammenhangs unterstützen: durch den **Korrelationskoeffizienten von Bravais-Pearson**.

Definition: Korrelationskoeffizient von Bravais-Pearson

Der Korrelationskoeffizient von Bravais-Pearson r_{xy} wird wie folgt berechnet:

$$r_{xy} = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\sqrt{\overline{x^2} - (\overline{x})^2} \cdot \sqrt{\overline{y^2} - (\overline{y})^2}}$$

Für den Korrelationskoeffizienten von Bravais-Pearson gilt:

Eigenschaften des Korrelationskoeffizienten

- Der Korrelationskoeffizient liegt immer zwischen -1 und 1, d. h. $-1 \leq r_{xy} \leq 1$.
- $r_{xy} = 1$ genau dann, wenn alle Messwerte exakt auf einer Geraden mit positiver Steigung liegen.
- $r_{xy} = -1$ genau dann, wenn alle Messwerte exakt auf einer Geraden mit negativer Steigung liegen.

Üblicherweise erreichen die Messwerte nie eine Korrelation von ± 1. Um die Aussagekraft der Korrelation dann beurteilen zu können, gibt es einige Daumenregeln, die aber in Abhängigkeit von der empirischen Untersuchung variiert werden müssen. Bei Befragungen von Menschen zu heiklen Themen z. B. hat man oft so große Störeffekte, dass es selten zu starken Korrelationen kommt. Auch das Skalenniveau der Antwortmöglichkeiten verzerrt die Korrelationswerte.

Definition: Korreliertheit

- Die Merkmale X und Y heißen **positiv korreliert**, wenn $r_{xy} > 0$.
- Die Merkmale X und Y heißen **negativ korreliert**, wenn $r_{xy} < 0$.
- Die Merkmale X und Y heißen **unkorreliert**, wenn $r_{xy} = 0$. In diesem Fall liegt kein linearer Zusammenhang zwischen den Merkmalen vor.
- Die Merkmale X und Y heißen nicht/kaum korreliert, wenn $-0.3 \leq r_{xy} \leq 0.3$.
- Die Merkmale X und Y heißen schwach korreliert, wenn $-0.7 \leq r_{xy} \leq -0.3$ bzw. $0.3 \leq r_{xy} \leq 0.7$.
- Die Merkmale X und Y heißen stark korreliert, wenn $-1 \leq r_{xy} \leq -0.7$ bzw. $0.7 \leq r_{xy} \leq 1$.

Streng genommen handelt es sich beim Korrelationskoeffizienten nicht um eine Maßzahl, denn wie weiter oben gefordert, soll eine Maßzahl immer zwischen 0 und 1 liegen, und das tut r_{xy} offensichtlich nicht. Durch Quadrieren des Korrelationskoeffizienten erreicht man aber genau dies, erkauft sich das allerdings durch den Verlust der Orientierung der Korrelation (positiv oder negativ korreliert).

Definition: Bestimmtheitsmaß

Die entstehende Maßzahl heißt **Bestimmtheitsmaß** $B_{xy} = r_{xy}^2$.

Eigenschaften des Bestimmtheitsmaßes

- Es gilt: $0 \leq B_{xy} \leq 1$.
- $B_{xy} = 1$ genau dann, wenn alle Messwerte auf einer Geraden liegen.
- Die Merkmale X und Y heißen nicht/kaum korreliert, wenn $0 \leq B_{xy} \leq 0.1$.
- Die Merkmale X und Y heißen schwach korreliert, wenn $0.1 \leq B_{xy} \leq 0.5$.
- Die Merkmale X und Y heißen stark korreliert, wenn $0.5 \leq B_{xy} \leq 1$.

Die Korrelation und das Bestimmtheitsmaß berechnen und interpretieren wir für unser Beispiel der Körpergrößen und Gewichte von Studierenden.

Beispiel

Im obigen Beispiel ergibt sich

$$r_{xy} = \frac{116.783 - 1.689 \cdot 68.3}{\sqrt{2.86711 - 1.689^2} \cdot \sqrt{4853.9 - 68.3^2}} = 0.863661931$$

sowie

$$B_{xy} = 0.863661931^2 = 0.745911931$$

Aus beiden Kenngrößen geht hervor, dass zwischen Körpergröße und Gewicht der Studierenden ein starker linearer Zusammenhang besteht; die Messgrößen sind stark korreliert. Auch sachlogisch macht diese Aussage Sinn. ◀

Beispiel

In dem Beispiel mit dem eindeutigen quadratischen Zusammenhang ergibt sich übrigens, dass sowohl

$$r_{xy} = \frac{12 - 2 \cdot 6}{\sqrt{6 - 2^2} \cdot \sqrt{38.8 - 6^2}} = 0$$

als auch $B_{xy} = 0$. Auch das legt nahe, dass man in der Situation keinen linearen Zusammenhang unterstellen sollte. ◀

Aufgaben

16.1 25 Unternehmer, von denen jeder einer der drei Branchen Automobil (A), Bau (B) oder Chemie (C) angehört, wurden zur wirtschaftlichen Entwicklung im kommenden Jahr befragt. Dabei ergab sich das folgende Ergebnis (b = besser, g = gleichbleibend, s = schlechter):

$(A,b), (C,b), (C,g), (A,b), (B,s), (A,g), (C,g), (B,g), (A,b),$
$(C,s), (A,b), (C,g), (A,s), (B,g), (C,b), (B,s), (A,g), (B,s),$
$(C,b), (A,b), (A,b), (C,b), (A,g), (B,g), (C,b)$

Der erste Eintrag beschreibt jeweils die Branche, während der zweite die Einschätzung für das kommende Jahr angibt.

1. Erstellen Sie eine Kontingenztabelle.
2. Berechnen Sie die bedingten Verteilungen.
3. Berechnen Sie die χ^2-Größe, den Kontingenzkoeffizienten und den korrigierten Kontingenzkoeffizienten.
4. Sind die Merkmale unabhängig?

16.2 Acht Läufer haben einen 100-m- und einen 400-m-Lauf absolviert. In der nachfolgenden Tabelle sind ihre Zeiten festgehalten.

Läufer	Zeit 100 m in s	Zeit 400 m in s
1	10.2	47.2
2	11.4	46
3	10.4	48.5
4	10.1	48
5	10.3	46.2
6	12	46.8
7	11.2	45.4
8	11	47

Berechnen Sie den Rangkorrelationskoeffizienten von Spearman.

16.3 Ein Unternehmen hat für die Jahre von 2005 bis 2014 die inflationsbereinigten Umsätze und die inflationsbereinigten Werbeausgaben notiert.

Jahr	Werbeausgaben in Mio. €	Umsatz in Mio. €
2005	5	120
2006	8	135
2007	8	145
2008	7	134
2009	9	165
2010	10	158
2011	12	170
2012	8	145
2013	10	175
2014	14	195

1. Zeichnen Sie das Streuungsdiagramm zur Einflussanalyse der Werbeausgaben auf den Umsatz.
2. Berechnen Sie die Regressionsgerade, die prognostiziert, welchen Einfluss die Werbeausgaben auf den Umsatz haben.
3. Berechnen Sie den Korrelationskoeffizienten von Bravais-Pearson und das Bestimmtheitsmaß.
4. Wie hoch ist Ihre Prognose für den Umsatz bei einem Werbebudget von 11 Mio. €? Wie beurteilen Sie die Situation bei einem Werbebudget von 100 Mio. €?