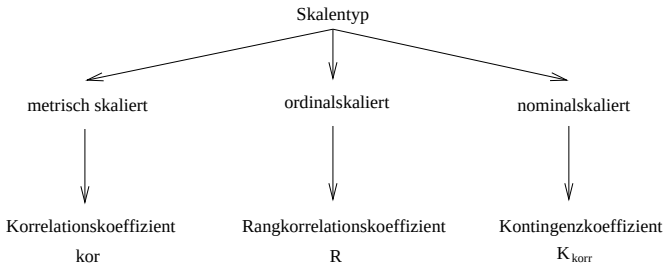
Statistik
Dr. Thomas Zehrt Zweidimensionale Häufigkeitsverteilung

Motivation

In der Praxis werden bei Datenerhebungen meist nicht nur ein Merkmal untersucht und wir werden uns in diesem Abschnitt auf zwei Merkmale einschränken. Bei der statistisch-mathematischen Auswertung dieser beiden Datenreihen kann es vorkommen, dass man den Eindruck gewinnt, dass sich beide Datenreihen ähnlich entwickeln (z.B. könnten bei jedem Merkmalsträger grosse Werte des einen Merkmals auch zu grossen Werten des anderen Merkmals gehören). Diese Stärke des Gleichlaufs beider Datenreihen wollen wir durch statistische Kennzahlen messen, wobei wir im Hinblick auf die Skalentypen unserer Merkmale, auch unterschiedliche Kennzahlen konstruieren werden.



Natürlich kann man aus statistischen Korrelationen von Datenreihen **nicht** direkt einen kausalen Zusammmenhang der beiden Merkmale ableiten, was aber oft genug passiert. Hier einige (mit Vorsicht zu geniessende) Beispiele aus den Medien:

- Grössere Leute verdienen mehr (Baz, 25.10.2003)
- Je mehr Lärm im Haus, desto dümmer die Kinder (Baz, 04.09.2004)
- Kaffee bewahrt vor Alzheimer (Stern, 06.06.2009)
- Senkung der Arbeitslosigkeit erfordert Wirtschaftswachstum (Okunsches Gesetz)

Benötigtes Schulwissen

- Bruchrechnung, Prozentrechnung und Summenzeichen
- ein wenig lineare Algebra für das Verständnis der Beweise (Vektoren, Länge von Vektoren, Skalarprodukt)
- die Minimum-Funktion $\min(l, k)$, die die kleinere (oder besser nicht grössere) der beiden Zahlen l und k auswählt, z.B. $\min(2, 4) = 2$, $\min(5, 4) = 4$ oder $\min(2, -4) = -4$.

1 Grundlagen

In diesem Abschnitt betrachten wir bei einer statistischen Masse **zwei** Merkmale X und Y und die damit verbundenen Besonderheiten. Beispielsweise könnten wir bei Menschen die Merkmale ,,Alter" und ,,Grösse" untersuchen.

Die Urliste gibt zu jedem statistischen Element i die beiden Merkmalsausprägungen x_i von X und y_i von Y nach folgendem Schema an:

Element Nr.	1	2	$\cdots$	i	$\cdots$	n
Ausprägung von X	x_1	x_2	$\cdots$	x_i	$\cdots$	x_n
Ausprägung von Y	y_1	y_2	$\cdots$	y_i	$\cdots$	y_n

Bei grösseren Datenmengen ist es wieder sinnvoll, alle Elemente zusammenzufassen, die in **beiden** Ausprägungen übereinstimmen.

Wir erhalten so eine zweidimensionale Häufigkeitstabelle, auch Kontingenztafel genannt. Dabei bezeichnen $a_1, \ldots, a_l$ die verschiedenen Ausprägungen von X und $b_1, \ldots, b_m$ die verschiedenen Ausprägungen von Y:

		Y						Vert.
		b_1	b_2	$\ldots$	b_k	$\ldots$	b_m	von X
	a_1	h_{11}	h_{12}	$\ldots$	h_{1k}	$\ldots$	h_{1m}	$h_{1\bullet}$
	a_2	h_{21}	h_{22}	$\ldots$	h_{2k}	$\ldots$	h_{2m}	$h_{2\bullet}$
	$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	
X	a_j	h_{j1}	h_{j2}	$\ldots$	h_{jk}	$\ldots$	h_{jm}	$h_{j\bullet}$
	$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	
	a_l	h_{l1}	h_{l2}	$\ldots$	h_{lk}	$\ldots$	h_{lm}	$h_{l\bullet}$
Vert. von Y		$h_{\bullet 1}$	$h_{\bullet 2}$	$\ldots$	$h_{\bullet k}$	$\ldots$	$h_{\bullet m}$	n

Es gilt (die in der Matrizenrechnung übliche) Konvention für Doppelindizes: Der erste Index bezeichnet die Zeile, der zweite die Spalte. h_{27} steht also in der 2-ten Zeile und 7-ten Spalte. Weiterhin gilt:

- h_{jk} = Anzahl Elemente mit $(X = a_j)$ **und** $(Y = b_k)$
- $h_{j\bullet}$ = Anzahl Elemente mit $(X = a_j) = \sum_{k=1}^{m} h_{jk}$ (j-te Zeilensumme)
  (wir ignorieren hier das Merkmal Y)
- $h_{\bullet k}$ = Anzahl Elemente mit $(Y = b_k) = \sum_{j=1}^{l} h_{jk}$ (k-te Spaltensumme)
  (wir ignorieren hier das Merkmal X)

In der Kontingenztafel sind die absoluten Häufigkeiten angegeben. Alternativ können wir auch die relativen Häufigkeiten $f_{jk} = \frac{h_{jk}}{n}$ verwenden, die dann die gemeinsame Verteilung der Merkmale X und Y darstellen.

Randverteilungen Natürlich kann man auch bei zwei Merkmalen das Augenmerk auf das eine oder andere Merkmal richten und die Zusammenhänge unbeachtet lassen. Bildlich gesprochen bedeutet das, dass wir nur einen Rand der Kontingenztafel anschauen und das Innere nicht beachten.

Die Häufigkeiten $h_{j\bullet}$ (bzw. $h_{\bullet k}$) sind die Häufigkeiten der Verteilung von X (bzw. von Y). Die relativen Häufigkeiten $f_{j\bullet} = \frac{h_{j\bullet}}{n}$ (bzw. $f_{\bullet k} = \frac{h_{\bullet k}}{n}$) heissen die Randverteilung von X (bzw. von Y).

Aufgabe 1.1 *Es wurden $n = 30$ Studenten nach ihren gerundeten Noten in Mathematik (Merkmal X) und in Statistik (Merkmal Y) befragt. Die Ergebnisse sind in folgender Urliste festgehalten:*

$$(6,6),(2,2),(4,4),(4,5),(5,4),(3,4),(4,4),(4,4),(4,4),(4,5),$$
$$(3,4),(4,4),(3,3),(4,5),(5,4),(3,4),(5,5),(4,4),(4,4),(4,5),$$
$$(3,4),(4,4),(2,3),(5,4),(5,4),(3,2),(5,5),(4,5),(4,4),(6,5)$$

Erstellen Sie die zweidimensionale Häufigkeitstabelle und identifizieren Sie die Randverteilungen.

Lösung:

X Mathematiknote

Y Statistiknote

		Y						*Vert. von X*
		$b_1 = 1$	$b_2 = 2$	$b_3 = 3$	$b_4 = 4$	$b_5 = 5$	$b_6 = 6$	
	$a_1 = 1$	0	0	0	0	0	0	0
	$a_2 = 2$	0	1	1	0	0	0	2
	$a_3 = 3$	0	1	1	4	0	0	6
X	$a_4 = 4$	0	0	0	9	5	0	14
	$a_5 = 5$	0	0	0	4	2	0	6
	$a_6 = 6$	0	0	0	0	1	1	2
Vert. von Y		0	2	2	17	8	1	30

Das rechteckige Zahlenschema (aus dem Inneren der Tabelle), das alle wesentlichen Daten enthält,

$$\begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 4 & 0 & 0 \\ 0 & 0 & 0 & 9 & 5 & 0 \\ 0 & 0 & 0 & 4 & 2 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \end{pmatrix}$$

wird als Matrix bzw. Datenmatrix bezeichnet. In der Vorlesung Mathematik 2 werden wir uns ausführlicher mit Matrizen und Matrizenrechnung beschäftigen. Dort werde ich versuchen Sie davon zu überzeugen, dass Matrizen sehr nützliche mathematische Objekte sind. Mit ihrer Hilfe lassen sich viele Probleme elegant formulieren und effizient lösen.

2 Zusammenhang zw. metrischskalierten Merkmalen

Sind die beiden Merkmale X und Y **metrisch skaliert**, so sind die Abstände zwischen den Merkmalsausprägungen interpretierbar und können bei der Konstruktion eines Zusammenhangsmasses berücksichtigt werden.

Für jedes der beiden Merkmale X und Y kann natürlich wieder das arithmetische Mittel und die Varianz bestimmt werden.

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \qquad var(X) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \qquad var(Y) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})^2$$

Beispiel 2.1 *Bei zwölf Kraftfahrzeugen eines Fuhrparkes werden die Merkmale X ,,Alter" (in Jahren) und Y ,,gefahrene Kilometer" (in 1000 km) untersucht. Es ergibt sich die folgende Urliste:*

Nr.	1	2	3	4	5	6	7	8	9	10	11	12
X	*1.5*	*5.2*	*4.5*	*0.5*	*2.4*	*2.6*	*1.8*	*4.2*	*6.2*	*3.6*	*2.5*	*5.1*
Y	*30*	*68*	*90*	*12*	*100*	*62*	*21*	*112*	*230*	*120*	*56*	*109*

Es ist ein deutlicher (und auch plausibeler) Zusammenhang zu erkennen: Ältere Autos haben mehr gefahrene Kilometer. Es scheint jedoch auch plausible zu sein, dass man **keine** *allgemeingültige Formel angeben kann, mit deren Hilfe sich für jedes beliebige Auto die gefahrenen Kilometer aus dem Alter errechnen lassen.*

Aufgabe 2.1 *Berechnen Sie $\bar{x}$, $\bar{y}$, $var(X)$ und $var(Y)$.*

Lösung: *($\overline{x} = 3.34$, $\overline{y} = 84.1$, $var(X) = 2.73$ und $var(Y) = 3158.8$)*

2.1 Die Kovarianz

Eine wirklich neue Masszahl, an der beide Messreihen gleichzeitig beteiligt sind, ist die Kovarianz.

Definition 2.1 *Die (empirische) Kovarianz* $cov(X,Y)$ *der Wertepaare* $(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)$ *ist definiert durch*

$$cov(X,Y) := \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}).$$

Beispiel 2.2 *Wir betrachten wieder unser Beispiel 2.1 mit der Urliste*

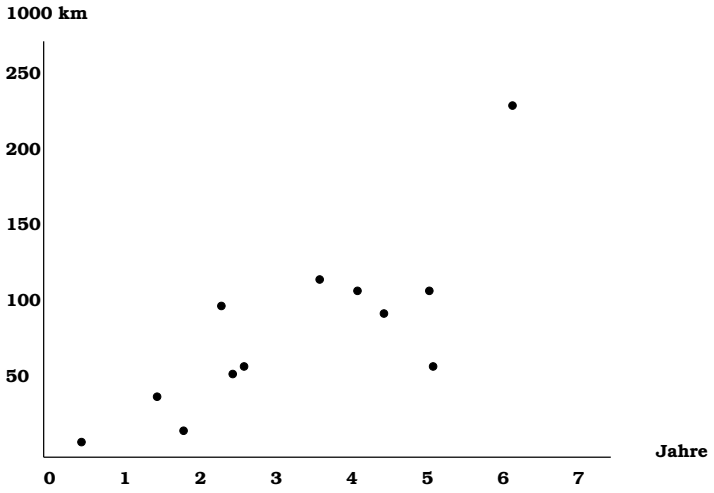
Nr.	1	2	3	4	5	6	7	8	9	10	11	12
X	*1.5*	*5.2*	*4.5*	*0.5*	*2.4*	*2.6*	*1.8*	*4.2*	*6.2*	*3.6*	*2.5*	*5.1*
Y	*30*	*68*	*90*	*12*	*100*	*62*	*21*	*112*	*230*	*120*	*56*	*109*

Aufgabe 2.2 *Berechnen Sie die Grösse* $cov(X,Y)$.

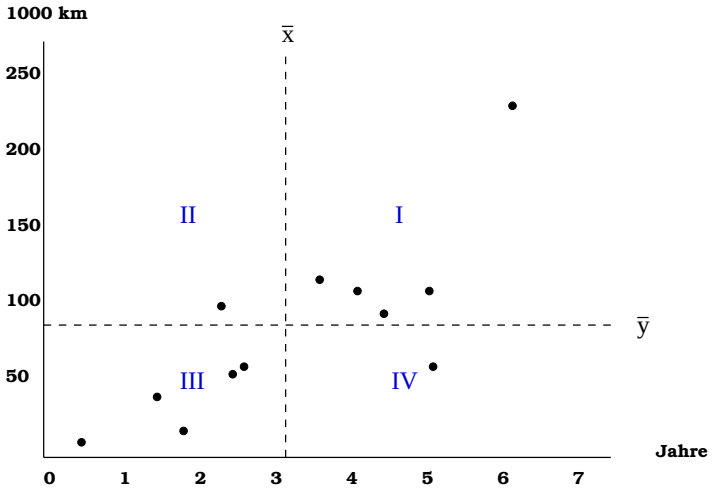
Lösung:

$$\begin{aligned} cov(X,Y) &= \frac{1}{12}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) \\ &= \frac{1}{12}\Big((1.5-3.34)(30-84.1)+(5.2-3.34)(68-84.1)+\ldots \\ &\qquad +(5.1-3.34)(109-84.1)\Big) \\ &= 74.81 \end{aligned}$$

Wir wollen an diesem Beispiel versuchen, eine geometrische Deutung des Vorzeichens von $cov(X,Y)$ zu finden. Dazu zeichnen wir ein Koordinatensystem mit der Jahresskala auf der x-Achse und der km-Skala auf der y-Achse. Dann tragen wir alle Merkmalsausprägungen als Punkte ein.



Nun zerlegen wir das Koordinatensystem durch das Einzeichnen der arithmetischen Mittel beider Merkmale in vier Sektoren und schauen uns jeden der entstandenen Sektoren genauer an.



Sektor I: $x_i > \bar{x}$ *und* $y_i > \bar{y}$ *also* $(x_i - \bar{x}) \cdot (y_i - \bar{y}) > 0$
Sektor II: $x_i < \bar{x}$ *und* $y_i > \bar{y}$ *also* $(x_i - \bar{x}) \cdot (y_i - \bar{y}) < 0$
Sektor III: $x_i < \bar{x}$ *und* $y_i < \bar{y}$ *also* $(x_i - \bar{x}) \cdot (y_i - \bar{y}) > 0$
Sektor IV: $x_i > \bar{x}$ *und* $y_i < \bar{y}$ *also* $(x_i - \bar{x}) \cdot (y_i - \bar{y}) < 0$

Also folgt:

$$cov(X,Y) = \frac{1}{n}\left(\sum_{(x_i,y_i)\in I} \underbrace{(x_i-\bar{x})(y_i-\bar{y})}_{positiv} + \sum_{(x_i,y_i)\in II} \underbrace{(x_i-\bar{x})(y_i-\bar{y})}_{negativ} + \sum_{(x_i,y_i)\in III} \underbrace{(x_i-\bar{x})(y_i-\bar{y})}_{positiv} + \sum_{(x_i,y_i)\in IV} \underbrace{(x_i-\bar{x})(y_i-\bar{y})}_{negativ} \right)$$

Ist also die Kovarianz grösser als Null, so überwiegen meist die Punkte in den Sektoren I und III. Die Punktwolke geht von links unten nach rechts oben. Ist die Kovarianz dagegen kleiner als Null, dann üeberwiegen meist die Punkte in den Sektoren II und IV. Die Punktwolke erstreckt sich von links oben nach rechts unten.

Die Kovarianz ist ein Mass für die Richtung des Zusammenhanges der Merkmale X und Y. Sie kann jeden reellen Wert annehmen.

Aufgabe 2.3 *Ergänzen Sie die fehlenden Rechenschritte:*

$$cov(X,Y) := \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$$

$$= \frac{1}{n}\left(\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}\right)$$

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

Satz 1 *Seien die n Wertepaare $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ gegeben. Dann gilt*

$$\boxed{|cov(X,Y)| \;\leq\; \sqrt{var(X)} \cdot \sqrt{var(Y)}}$$

Die Gleichheit gilt genau dann, wenn zwischen den x_i und den y_i ein linearer Zusammenhang besteht, d.h. es existieren reelle Zahlen a und b, so dass

$$y_i \;=\; ax_i + b$$

für alle $i = 1, 2, \ldots, n$.

Beweis: Wir formulieren die Behauptung des Satzes in der Sprache der linearen Algebra. Dazu definieren wir die folgenden beiden Vektoren:

$$\mathbf{v} = \begin{pmatrix} x_1 - \bar{x} \\ \cdots \\ x_n - \bar{x} \end{pmatrix} \quad \text{und} \quad \mathbf{w} = \begin{pmatrix} y_1 - \bar{y} \\ \cdots \\ y_n - \bar{y} \end{pmatrix}.$$

Dann gilt

$$\begin{aligned} var(X) &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \;=\; \frac{1}{n}\,||\mathbf{v}||^2 \\ var(Y) &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 \;=\; \frac{1}{n}\,||\mathbf{w}||^2 \\ cov(X,Y) &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \;=\; \frac{1}{n}\,\langle \mathbf{v}, \mathbf{w}\rangle \end{aligned}$$

Dabei bezeichnet $||\,.\,||$ die Euklidische Norm, d.h. die Länge eines Vektors und $\langle\ ,\ \rangle$ das gebräuchliche Skalarprodukt. Nutzen wir diese linear-algebraische Schreibweise, so geht die obige Gleichung in die folgende Gleichung über:

$$|\langle \mathbf{v}, \mathbf{w}\rangle| \quad\leq\quad ||\mathbf{v}|| \cdot ||\mathbf{w}||.$$

Diese Ungleichung ist die sogenannte Cauchy-Schwarzsche Ungleichung und für alle Vektoren richtig. Genauer gesagt, gilt sogar das folgende:

$$|\langle \mathbf{v}, \mathbf{w}\rangle| \quad=\quad ||\mathbf{v}|| \cdot ||\mathbf{w}|| \cdot \underbrace{|\cos(\angle(\mathbf{v}, \mathbf{w}))|}_{\leq 1} \quad \leq ||\mathbf{v}|| \cdot ||\mathbf{w}||,$$

wobei $\angle(\mathbf{v}, \mathbf{w})$ den Winkel zwischen den Vektoren $\mathbf{v}$ und $\mathbf{w}$ bezeichnet. Die Gleichheit gilt hier genau dann, wenn $|\cos(\angle(\mathbf{v}, \mathbf{w}))| = 1$ ist, d.h. wenn eine der folgenden Möglichkeiten eintritt:

1. $\cos(\angle(\mathbf{v}, \mathbf{w})) = 1$ oder $\angle(\mathbf{v}, \mathbf{w}) = 0^o$
2. $\cos(\angle(\mathbf{v}, \mathbf{w})) = -1$ oder $\angle(\mathbf{v}, \mathbf{w}) = 180^o$.

In beiden Fällen sind die beiden Vektoren linear abhängig, d.h. es gibt eine reelle Zahl a, so dass $\mathbf{w} = a\mathbf{v}$ oder

$$\begin{pmatrix} y_1 - \bar{y} \\ \cdots \\ y_n - \bar{y} \end{pmatrix} \quad=\quad a \begin{pmatrix} x_1 - \bar{x} \\ \cdots \\ x_n - \bar{x} \end{pmatrix}$$

gilt. Damit ergibt sich aber für jeden Index $i = 1, \ldots, n$

$$y_i \quad=\quad ax_i + \underbrace{\bar{y} - a\bar{x}}_{=:b}.$$

□

Beispiel 2.3

	1	2	3
x_i	1	2	3
y_i	2	4	5

Skizzieren Sie die Punktwolke!

Es gilt: $\bar{x} = 2$, $\bar{y} = 11/3$, $var(X) = 2/3$, $var(Y) = 14/9$ *und* $cov(X, Y) = 1$.

$$\underbrace{|cov(X,Y)|}_{=1} \quad < \quad \underbrace{\sqrt{var(X)}}_{=\sqrt{2/3}} \cdot \underbrace{\sqrt{var(Y)}}_{=\sqrt{14/9}}$$

Beispiel 2.4

	1	2	3
x_i	1	2	3
y_i	2	4	6

Skizzieren Sie die Punktwolke!

Es gilt: $\bar{x} = 2$, $\bar{y} = 4$, $var(X) = 2/3$, $var(Y) = 8/3$ *und* $cov(X, Y) = 4/3$.

$$\underbrace{|cov(X,Y)|}_{=4/3} \quad = \quad \underbrace{\sqrt{var(X)}}_{=\sqrt{2/3}} \cdot \underbrace{\sqrt{var(Y)}}_{=\sqrt{8/3}}$$

Hier gilt Gleichheit und tatsächlich liegt die Punktwolke auf einer Geraden.

2.2 Der Korrelationskoeffizient

Die Kovarianz $cov(X,Y)$ ist zwar ein Mass für die Richtung des Zusammenhanges von X und Y, jedoch als Mass für die Stärke des Zusammenhanges offenbar völlig ungeeignet, da sie von den physikalischen Einheiten der Merkmale abhängt. Durch geeignete Wahl der physikalischen Einheiten könnten wir sogar jede beliebige positive reelle Zahl als Kovarianz immer der selben Datenmenge erzwingen. Dies wird in Ordnung gebracht, indem man die Kovarianz durch die Standartabweichungen dividiert.

Definition 2.2 *Seien die n Wertepaare $(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)$ gegeben, wobei nicht alle x_i gleich sind und nicht alle y_i gleich sind. Der (empirische) Korrelationskoeffizient $kor(X,Y)$ ist definiert durch*

$$kor(X,Y) := \frac{cov(X,Y)}{\sqrt{var(X)}\cdot\sqrt{var(Y)}} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}}.$$

$kor(X,Y)$ ist nicht nur ein Mass für die Richtung des Zusammenhanges zwischen X und Y, sondern gleichzeitig ein Mass für die **Stärke des (statistischen) linearen Zusammenhanges** zwischen X und Y.

Satz 2 *Der Korrelationskoeffizient $kor(X,Y)$ kann nur Werte zwischen -1 und $+1$ annehmen. Weiterhin gilt:*

$$kor(X,Y)=+1 \iff y_i = ax_i+b \quad mit\ a>0$$
$$kor(X,Y)=-1 \iff y_i = ax_i+b \quad mit\ a<0.$$

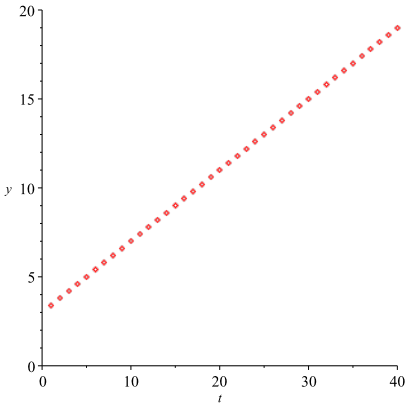
Beweis: Nutzen wir die Vektorschreibweise aus dem Beweis des Satzes , so gilt

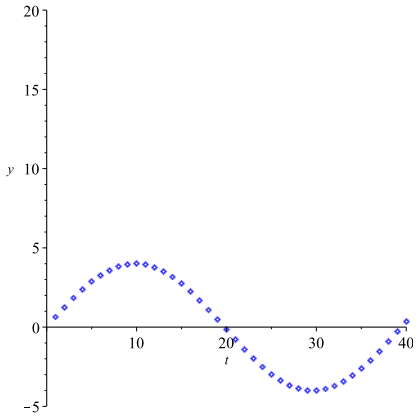
$$kor(X,Y) = \frac{\langle \mathbf{v},\mathbf{w}\rangle}{||\mathbf{v}||\,||\mathbf{w}||} = \cos(\angle(\mathbf{v},\mathbf{w})) \in [-1,1].$$

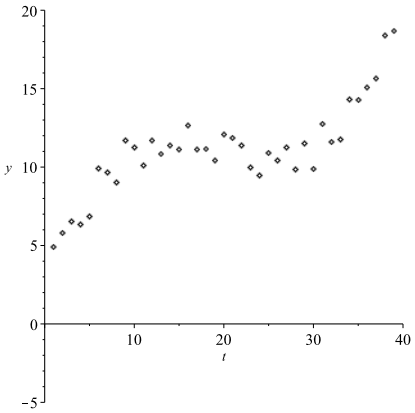
Der Korrelationskoeffizient nimmt also tatsächlich nur Werte zwischen -1 und 1 an. Ist nun $kor(X,Y) = +1 = \cos(\angle(\mathbf{v},\mathbf{w}))$ so folgt $\angle(\mathbf{v},\mathbf{w}) = 0^o$ und beide Vektoren liegen auf einer Geraden und zeigen in die gleiche Richtung. Somit existiert eine positive reelle Zahl a, so dass $\mathbf{w} = a\mathbf{v}$ gilt. □

Im Falle von $kor(X,Y)=0$ heissen die Merkmale X und Y unkorreliert. Im Falle von $kor(X,Y)=+1$ (bzw. $kor(X,Y)=-1$) heissen die Merkmale X und Y stark positiv (bzw. stark negativ) korreliert.

Aufgabe 2.4 *Schätzen Sie die Korrelationskoeffizienten der Daten, die in den folgenden Punktwolken dargestellt sind. Begründen Sie ihre Schätzung.*







Können Sie beliebig viele Punktwolken skizzieren, deren zugehörige Merkmale stark positiv bzw. stark negativ korelliert sind?

Achtung:

- Man darf aber hier **nicht** den Fehler machen, aus dem Korrelationskoeffizienten abzuleiten zu wollen, dass es zwischen den Merkmalen eine kausalen Zusammenhang gibt (falls $kor(X, Y) = +1$) bzw. nicht gibt ($kor(X, Y) = 0$)! Der Korrelationskoeffizient kann bestenfalls als Indiz angesehen werden.
- Der Korrelationskoeffizient sagt etwas über einen **funktionalen linearen** Zusammenhang der Datenreihen aus. Falls $kor(X, Y) \approx 0$ ist, kann es zwischen den Datenreihen aber immernoch einen anderen funktionalen Zusammenhang geben (z.B. könnten die Punkte auf einer quadratischen Parabel liegen).
- Nehmen wir an, dass zwei Merkmale X und Y miteinander korrelieren ($kor(X, Y) \approx 1$). Dann gibt es vier Möglichkeiten, weshalb sie das tun:
  1. X ist (kausale) Ursache von Y oder Y ist (kausale) Ursache von X.
  2. X und Y haben eine gemeinsame Ursache (Hintergrundvariable).
  3. Die Korrelation beruht auf einem systematischen Fehler.
  4. Die Korrelation ist zufällig, trotz statistischer Signifikanz.

Big Data

Wissenschaftler wiederholen gerne, dass Korrelation noch keine Kausalität bedeuten muss und dass man keine Ursache-Wirkungsbeziehung nur auf Basis der Korrelation (in eine kleinen Stichprobe) beweisen kann. Man braucht ein Modell und ein tiefes Verständnis der Prozesse. Erst dann kann man versuchen Zusammenhänge herzustellen und event. statistisch beweisen.

Aber im Big-Data-Zeitalter ist dieser Zugang eventuell überholt, meinen einige Leute. Gigantische Datensätze könnten uns erlauben zu sagen, dass **Korrelation reicht**!

Konzerne wie Google erobern die Werbung, ohne jegliche Fachkenntniss. Gute und schnelle mathematische Analyseverfahren genügen, um das Verhalten des Kunden erschreckend genau vorherzusagen, ohne das Warum zu verstehen.

Einige Fragen, die man mit Big-Data-Analysen zu beantworten versucht:

- Was kaufen Sie als nächstes ein?
- Wird am Sonntag bei Ihnen eingebrochen? (Predictive Policing)
- Wann werden Sie das nächste Mal im Spital sein und wie lange?

Seien Sie vorsichtig! Ihr Mobiltelefon sammelt permanent Daten über Sie (GPS, Anrufe, SMS, Bluetooth, WLAN, interne Sensoren, ...)! Aus diesen Daten kann man schon jetzt recht gut vorhersagen, wo Sie sich in der nächsten Woche aufhalten werden (bis auf 3 Meter genau) und mit wem Sie sich dann treffen!

3 Zusammenhang zw. ordinalskalierten Merkmalen

Seien X und Y ordinalskalierte Merkmale. Den Wertepaaren $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ werden zunächst Rangzahlen R_i bezüglich $(x_1, x_2, \ldots, x_n)$ und R'_i bezüglich $(y_1, y_2, \ldots, y_n)$ folgendermassen zugeordnet:

- Dem grössten Wert von $(x_1, x_2, \ldots, x_n)$ wird die Rangzahl 1 zugeordnet, dem zweitgrössten Wert die Rangzahl 2 usw. ... dem kleinsten Wert die Rangzahl n.
- Bezüglich $(y_1, y_2, \ldots, y_n)$ verfahren wir analog.

  Die geordnete Menge der Rangzahlen sei mit $R(X)$ bzw. $R(Y)$ bezeichnet.
- Haben zwei oder mehr Beobachtungen die gleiche Ausprägung von X (oder von Y), so liegt eine so genannte Bindung vor. Als Rang der einzelnen (identischen) Beobachtungen wird dann das arithmetische Mittel der zu vergebenen Ränge gewählt.

Definition 3.1 *Der Rangkorrelationskoeffizient der Wertepaare* $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ *ist definiert als der (empirische) Korrelationskoeffizient der Rangzahlen:*

$$R = R(X, Y) := \frac{cov(R(X), R(Y))}{\sqrt{var(R(X))}\sqrt{var(R(Y))}}.$$

Die Interpretation der erhaltenen Werte zwischen -1 und 1 erfolgt analog zum Korrelationskoeffizienten.

Satz 3 *Falls in den Daten* **keine Bindungen** *auftreten gilt:*

$$R = 1 - \frac{6\sum_{i=1}^{n}(R_i - R'_i)^2}{n(n^2-1)}.$$

Beispiel 3.1 *Die folgende Tabelle zeigt die Prüfungsresultate von zehn Schülern in den Fächern Mathematik (Merkmal M = Punktzahl in der Mathematikprüfung) und Physik (Merkmal P = Punktzahl in der Physikprüfung). Die maximal erreichbare Punktzahl war jeweils* 15*.*

Obwohl *es Bindungen in den Daten gibt, wollen wir die Formel aus dem Satz anwenden! Rechnet man den Korrelationskoeffizienten der Rangzahlen wie in der Definition aus, erhält man das exakte Resultat* 0.8580900871*.*

Schüler i	M	R_i	P	R'_i	$(R_i - R'_i)^2$
1	13	4	15	1	9
2	14	2.5	8	4	2.25
3	8	9	1	10	1
4	10	7	7	6.5	0.25
5	15	1	9	2	1
6	1	10	5	9	1
7	14	2.5	8	4	2.25
8	12	5	7	6.5	2.25
9	9	8	6	8	0
10	11	6	8	4	4

Also gilt

$$R \;=\; 1 - \frac{6 \cdot 23}{10(100-1)} \;=\; 0.8606$$

und wir erkennen eine (wie zu erwarten war) positive (Rang)korrelation.

4 Zusammenhang zw. nominalskalierten Merkmalen

Seien X und Y nominalskalierte Merkmale mit den in der folgenden (relativen) Kontingenztafel zusammengefassten Ausprägungen:

		Y						Vert.
		b_1	b_2	$\dots$	b_k	$\dots$	b_m	von X
	a_1	f_{11}	f_{12}	$\dots$	f_{1k}	$\dots$	f_{1m}	$f_{1\bullet}$
	a_2	f_{21}	f_{22}	$\dots$	f_{2k}	$\dots$	f_{2m}	$f_{2\bullet}$
	$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	
X	a_j	f_{j1}	f_{j2}	$\dots$	f_{jk}	$\dots$	f_{jm}	$f_{j\bullet}$
	$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	
	a_l	f_{l1}	f_{l2}	$\dots$	f_{lk}	$\dots$	f_{lm}	$f_{l\bullet}$
Vert. von Y		$f_{\bullet 1}$	$f_{\bullet 2}$	$\dots$	$f_{\bullet k}$	$\dots$	$f_{\bullet m}$	n

Wir bezeichnen die beiden Merkmale X und Y (bezüglich der obigen Messung) als unabhängig, **wenn** sich ihre gemeinsame Verteilung (hier reden wir von den **relativen** Häufigkeiten) als Produkt der beiden Randverteilungen ergäbe, d.h. für alle $j = 1, \dots, l$ und alle $k = 1, \dots, m$

$$f_{jk} \;=\; f_{j\bullet} \cdot f_{\bullet k}$$

gilt. In den absoluten Häufigkeiten ausgedrückt heisst das

$$h_{jk} \;=\; n \cdot f_{jk} \;=\; n \cdot f_{j\bullet} \cdot f_{\bullet k} \;=\; n \cdot \frac{h_{j\bullet}}{n} \cdot \frac{h_{\bullet k}}{n} \;=\; \frac{h_{j\bullet} \cdot h_{\bullet k}}{n}$$

Bei einem genaueren Blick auf diese Formel wird hoffentlich klar, warum man in diesem Fall über Unabhängigkeit redet.

$$\underbrace{h_{jk}}_{\substack{\text{Anzahl Elemente mit} \\ X = a_j \text{ und } Y = b_k}} \;=\; \underbrace{\frac{h_{j\bullet}}{n}}_{\substack{\text{Anteil der Elemente mit} \\ X = a_j}} \cdot \underbrace{h_{\bullet k}}_{\substack{\text{Anzahl Elemente mit} \\ Y = b_k}}$$

Definition 4.1 *Wir bezeichnen die beiden Merkmale X und Y (bezüglich der obigen Messung) als unabhängig, wenn für die Ausprägungen $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ und deren Häufigkeiten* **immer**

$$f_{jk} \;=\; f_{j\bullet} \cdot f_{\bullet k} \qquad \textit{oder} \qquad h_{jk} \;=\; \frac{h_{j\bullet} \cdot h_{\bullet k}}{n}$$

gilt.

Wir werden nun ein Zusammenhangsmass definieren, das in gewisser Weise die Abweichung der (tatsächlichen) Verteilung von der zugehörigen unabhängigen Verteilung misst.

Definition 4.2 *Der Chi-Quadrat-Koeffizient für die Wertepaare $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ (Ausprägungen eines nominalskalierten Merkmals) ist*

$$\chi^2 \;:=\; \sum_{j=1}^{l} \sum_{k=1}^{m} \frac{\left(h_{jk} - \frac{h_{j\bullet} \cdot h_{\bullet k}}{n}\right)^2}{\frac{h_{j\bullet} \cdot h_{\bullet k}}{n}}.$$

Der Kontingenzkoeffizient ist

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

Der Chi-Quadrat-Koeffizient wäre im Falle vollkommener Unabhängigkeit genau Null, in allen anderen Fällen strikt positiv.

Der Kontingenzkoeffizient ist ebenfalls Null, wenn der Chi-Quadrat-Koeffizient Null ist. Für grössere Werte von χ^2 wird auch K grösser, erreicht aber den Wert Eins nicht ganz:

$$0 \;\leq\; K \;\leq\; K_{max} \;=\; \sqrt{\frac{\min(l,m)-1}{\min(l,m)}} \;<\; 1.$$

Hier sei nochmals daran erinnert, dass l die Anzahl der Ausprägungen von X und m die Anzahl der Ausprägungen von Y. Somit ist $\min(l,m)$ einfach die Kleinere (oder nicht Grössere) der beiden Zahlen. Damit kommen wir letztendlich zu einem Mass für die Stärke des Zusammenhangs zweier nominalskalierter Merkmale.

Definition 4.3 *Der normierte Kontingenzkoeffizient ist definiert als*

$$\boxed{K_{korr} := \frac{K}{K_{max}}}$$

und nimmt Werte zwischen 0 *und* 1 *an.*

Beispiel 4.1 *Wir wollen den Zusammenhang zwischen den beiden Merkmalen $X =$,,Teilnahme an einer Grippeschutzimpfung,, und $Y =$,,Grippeerkrankung,, untersuchen. Untersuchungen an 1000 Personen ergab die folgende Häufigkeitstabelle:*

	Grippe	*keine Grippe*	*Vert. von X*
Impfung	40	458	498
keine Impfung	259	243	502
Vert. von Y	299	701	1000

Selbstverständlich ist $l = m = 2$ (jedes Merkmal hat zwei mögliche Ausprägungen), also ist $\min(2,2) = 2$. *Dann gilt nacheinander:*

$$\chi^2 \;=\; \frac{\left(40-\frac{498\cdot 299}{1000}\right)^2}{\frac{498\cdot 299}{1000}} + \frac{\left(458-\frac{498\cdot 701}{1000}\right)^2}{\frac{498\cdot 701}{1000}} + \frac{\left(259-\frac{502\cdot 299}{1000}\right)^2}{\frac{502\cdot 299}{1000}} + \frac{\left(243-\frac{502\cdot 701}{1000}\right)^2}{\frac{502\cdot 701}{1000}} \;=\; 226.334$$

$$K \;=\; \sqrt{\frac{226.334}{1000+226.334}} \;=\; 0.43$$

$$K_{korr} \;=\; 0.43 \Big/ \sqrt{\frac{2-1}{2}} \;=\; 0.43 \cdot \sqrt{\frac{2}{2-1}} \;=\; 0.61$$

Es besteht also ein Zusammenhang zwischen den beiden Merkmalen.

5 Übungsaufgaben

1. Es sei x_t die Anzahl der in Bayern registrierten Gästeübernachtungen im Monat t in Mio., y_t die Arbeitlosenzahl in Bayern am Ende des Monats t in 1000.

	Jan.	Feb.	März	Apr.	Mai	Juni	Juli	Aug.	Sept.	Okt.	Nov.	Dez.
x_t	4.55	4.36	4.67	5.01	5.55	6.87	6.99	7.16	7.76	6.33	5.02	4.16
y_t	431	427	428	403	423	347	337	321	349	357	366	377

   (a) Bestimmen Sie die Varianzen beider Messreihen, die Kovarianz und den Korrelationskoeffizienten.

   (b) Können Sie einen kausalen Zusammenhang der beiden Merkmale ableiten? Wie erklären Sie sich die berechnete Korrelation?

2. Ein Schuldirektor möchte prüfen, ob bei seinen Schülern ein Zusammenhang zwischen naturwissenschaftlicher und sprachlicher Begabung besteht. Welche Antwort erhält er bei der Auswertung der Punktzahlen x_i einer Biologieklausur und y_i einer Englischklausur von je 12 Schülern? Verwenden Sie dazu den Rangkorrelationskoeffizient.

Schüler	1	2	3	4	5	6	7	8	9	10	11	12
x_i	65	76	29	54	35	78	60	32	67	80	55	61
y_i	76	34	27	54	37	44	78	22	51	79	32	71

3. Ein Gesundheitsmagazin will im Fernsehen eine Sendung über das Gesundheitsbewusstsein 40-jähriger Männer und Frauen bringen. Dazu werden 206 Männer (47 von ihnen geben an, für die Gesundheit regelmässig etwas Sport zu treiben) und 294 Frauen (101 von ihnen geben an, für die Gesundheit regelmässig etwas Sport zu treiben) befragt. Besteht anhand dieser Daten ein Zusammenhang zwischen Geschlecht und Gesundheitsbewusstsein?

4. Gegeben seien die x-Werte $x_1 = -2$, $x_2 = -1$, $x_3 = 0$, $x_4 = 1$ und $x_5 = 2$.

   (a) Bestimmen Sie **zwei** Reihen von y-Werten $y_1, \ldots, y_5$, so dass der Korrelationskoeffizient der Wertepaare $(x_1, y_1), \ldots, (x_5, y_5)$ gleich $+1$ ist.

   (b) Bestimmen Sie **zwei** Reihen von y-Werten $y_1, \ldots, y_5$, so dass der Korrelationskoeffizient der Wertepaare $(x_1, y_1), \ldots, (x_5, y_5)$ gleich -1 ist.

   (c) Bestimmen Sie **zwei** Reihen von y-Werten $y_1, \ldots, y_5$, so dass die Kovarianz der Wertepaare $(x_1, y_1), \ldots, (x_5, y_5)$ gleich 0 ist.

   (d) Die y-Werte seien durch die Funktion $y_i = x_i^n$ für $i = 1, \ldots, 5$ und eine beliebige gerade Zahl n definiert. Es besteht also ein funktionaler Zusammenhang. Berechnen Sie die Kovarianz der Wertepaare $(x_1, y_1), \ldots, (x_5, y_5)$.

Resultate einiger Übungsaufgaben

1. (a) Direkte und mühsame Rechnung ergibt die folgenden Werte:
   $\bar{x} = 5.702$, $\bar{y} = 380.5$, $var(X) = 1.44$, $var(Y) = 1466.92$, $cov(X,Y) = -36.49$ und $kor(X,Y) = -0.793$.
   Diese Aufgabe ist auch dazu gedacht, dass Sie Ihre Rechenfähigkeiten überprüfen können. Tun Sie das auch mindestens einmal. Natürlich werden wir solche Aufgaben später mit einer Software lösen.

   (b) Nein, das darf man nie.

2. $$R = 1 - \frac{6\sum_{i=1}^{n}(R_i - R_i')^2}{n(n^2-1)} = 1 - \frac{6 \cdot 124}{12(12^2-1)} = 0.5664$$

   Es existiert ein positiver, wenn auch nicht starker, Zusammenhang.

3. $$\begin{aligned}\chi^2 &= \frac{\left(47 - \frac{206\cdot 148}{500}\right)^2}{\frac{206\cdot 148}{500}} + \frac{\left(159 - \frac{206\cdot 352}{500}\right)^2}{\frac{206\cdot 352}{500}} + \frac{\left(101 - \frac{294\cdot 148}{500}\right)^2}{\frac{294\cdot 148}{500}} + \frac{\left(193 - \frac{294\cdot 352}{500}\right)^2}{\frac{294\cdot 352}{500}} \\ &= 3.2034 + 1.3469 + 2.2445 + 0.9437 = 7.7385\end{aligned}$$

   $$K = \sqrt{\frac{7.7385}{500 + 7.7385}} = 0.12345$$

   $$K_{korr} = 0.12345 \Big/ \sqrt{\frac{2-1}{2}} = 0.43 \cdot \sqrt{\frac{2}{2-1}} = 0.17458$$

   Es besteht ein schwacher Zusammenhang.

4. (a) Sie **müssen** die Aussage von Satz 2 verstehen! Dann können Sie diese Aufgabe leicht (und ohne grosse Rechnung) lösen.

   (b) Sie **müssen** die Aussage von Satz 2 verstehen! Dann können Sie diese Aufgabe leicht (und ohne grosse Rechnung) lösen.

   (c) Ein ganz einfaches Beispiel wären z.B. die (konstanten) y-Werte $y_1 = 2$, $y_2 = 2$, $y_3 = 2$, $y_4 = 2$ und $y_5 = 2$. Prüfen Sie das! Hätte die entsprechende Punktwolke auch einen Korrelationskoeffizienten von 0?

   (d)

   $$\begin{aligned}cov(X,Y) &= \frac{1}{5}\sum_{i=1}^{5}(x_i - \bar{x}) \cdot (y_i - \bar{y}) \\ &= \ldots \\ &= 0.\end{aligned}$$