

Inhaltsverzeichnis

(Ausschnitt)

4 Zweidimensionale Daten

- Häufigkeitsverteilungen unklassierter Daten
- Häufigkeitsverteilungen klassierter Daten
- Bedingte Häufigkeitsverteilungen und Unabhängigkeit
- Abhängigkeitsmaße

Auswertungsmethoden für mehrdimensionale Daten II

- Isolierte Betrachtung der einzelnen Merkmale kann allerdings Abhängigkeiten zwischen mehreren Merkmalen nicht erkennbar machen!
 - Zur Untersuchung von Abhängigkeiten mehrerer Merkmale „simultane“ Betrachtung der Merkmale erforderlich.
 - Gemeinsame Betrachtung von mehr als 2 Merkmalen allerdings technisch schwierig.
- ↪ Spezielle Methoden für **zweidimensionale** Daten (2 Merkmale simultan)

Auswertungsmethoden für mehrdimensionale Daten I

- Werden zu einer statistischen Masse mehrere Merkmale erhoben, so können diese natürlich individuell mit den Methoden für einzelne Merkmale ausgewertet werden.
- Eine Menge von Kennzahlen in den Spalten kann zum Beispiel gegen eine Menge von Merkmalen in den Zeilen tabelliert werden:

BMW.DE	$x_{(1)}$	$x_{0.5}$	$x_{(n)}$	\bar{x}	s	IQA	Schiefe	Kurt.
Preise	17.610	28.040	35.940	27.967	4.974	8.015	-0.383	1.932
log-Preise	2.868	3.334	3.582	3.314	0.189	0.286	-0.618	2.258
Renditen	-0.078	-0.001	0.148	0.002	0.030	0.034	0.672	5.941
log-Renditen	-0.081	-0.001	0.138	0.001	0.029	0.034	0.484	5.396

- Liegen die Merkmalswerte jeweils in ähnlichen Wertebereichen, ist auch ein Box-Plot verschiedener Merkmale nützlich.

Häufigkeitsverteilungen zweidimensionaler Daten I

- Im Folgenden wird angenommen, dass den Merkmalsträgern zu **zwei** Merkmalen X und Y Merkmalswerte zugeordnet werden, also ein **zweidimensionales Merkmal** (X, Y) vorliegt.
- Analog zum eindimensionalen Fall geht man davon aus, auch vor der Erhebung schon Mengen M_1 bzw. M_2 angeben zu können, die alle vorstellbaren Merkmalswerte des Merkmals X bzw. Y enthalten.
- Die Urliste der Länge n (zur statistischen Masse der Mächtigkeit n) besteht nun aus den n Paaren

$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$$

mit $x_m \in M_1$ und $y_m \in M_2$ bzw. $(x_m, y_m) \in M_1 \times M_2$ für $m \in \{1, \dots, n\}$.

Häufigkeitsverteilungen zweidimensionaler Daten II

- *Unverzichtbare Eigenschaft der Urliste ist, dass die Paare von Merkmalswerten jeweils demselben Merkmalsträger zuzuordnen sind!*
- Wie im eindimensionalen Fall wird der Merkmalsraum zu X mit $A = \{a_1, \dots, a_k\}$ bezeichnet, darüberhinaus der Merkmalsraum zu Y mit $B = \{b_1, \dots, b_l\}$.
- Es muss nicht jede der $k \cdot l$ Kombinationen (a_i, b_j) in der Urliste auftreten!
- Geeignetes Mittel zur Aggregation der Merkmalswerte, wenn sowohl $k = \#A$ als auch $l = \#B$ „klein“ sind: **Häufigkeitsverteilungen**

Häufigkeitsverteilungen zweidimensionaler Daten IV

- Natürlich gilt auch hier $\sum_{i=1}^k \sum_{j=1}^l h(a_i, b_j) = n$ und $\sum_{i=1}^k \sum_{j=1}^l r(a_i, b_j) = 1$.
- Tabellarische Darstellung zweidimensionaler Häufigkeitsverteilungen in **Kontingenztafeln**:

$X \setminus Y$	b_1	b_2	\dots	b_l
a_1	h_{11}	h_{12}	\dots	h_{1l}
a_2	h_{21}	h_{22}	\dots	h_{2l}
\vdots	\vdots	\vdots	\ddots	\vdots
a_k	h_{k1}	h_{k2}	\dots	h_{kl}

- Statt absoluter Häufigkeiten h_{ij} hier auch relative Häufigkeiten r_{ij} üblich.

Häufigkeitsverteilungen zweidimensionaler Daten III

- Zur Erstellung einer Häufigkeitsverteilung: Zählen, wie oft jede Kombination (a_i, b_j) der Merkmalsausprägung a_i von X und b_j von Y , $i \in \{1, \dots, k\}$, $j \in \{1, \dots, l\}$, in der Urliste $(x_1, y_1), \dots, (x_n, y_n)$ vorkommt.
 - ▶ Die **absoluten Häufigkeiten** $h_{ij} := h(a_i, b_j)$ geben für die Kombination (a_i, b_j) , $i \in \{1, \dots, k\}$, $j \in \{1, \dots, l\}$, die (absolute) Anzahl der Einträge der Urliste mit der Ausprägung (a_i, b_j) an, in Zeichen

$$h_{ij} := h(a_i, b_j) := \#\{m \in \{1, \dots, n\} \mid (x_m, y_m) = (a_i, b_j)\}.$$

- ▶ Die **relativen Häufigkeiten** $r_{ij} := r(a_i, b_j)$ geben für die Kombination (a_i, b_j) , $i \in \{1, \dots, k\}$, $j \in \{1, \dots, l\}$, den (relativen) Anteil der Einträge der Urliste mit der Ausprägung (a_i, b_j) an der gesamten Urliste an, in Zeichen

$$r_{ij} := r(a_i, b_j) := \frac{h(a_i, b_j)}{n} = \frac{\#\{m \in \{1, \dots, n\} \mid (x_m, y_m) = (a_i, b_j)\}}{n}.$$

Häufigkeitsverteilungen zweidimensionaler Daten V

- Zu den absoluten Häufigkeiten h_{ij} und relativen Häufigkeiten r_{ij} definiert man die **absoluten Randhäufigkeiten**

$$h_{i\cdot} := \sum_{j=1}^l h_{ij} \text{ für } i \in \{1, \dots, k\} \quad \text{und} \quad h_{\cdot j} := \sum_{i=1}^k h_{ij} \text{ für } j \in \{1, \dots, l\}$$

sowie die **relativen Randhäufigkeiten**

$$r_{i\cdot} := \sum_{j=1}^l r_{ij} \text{ für } i \in \{1, \dots, k\} \quad \text{und} \quad r_{\cdot j} := \sum_{i=1}^k r_{ij} \text{ für } j \in \{1, \dots, l\}.$$

- Diese Randhäufigkeiten stimmen offensichtlich (!) mit den (eindimensionalen) individuellen Häufigkeitsverteilungen der Merkmale X bzw. Y überein.

Häufigkeitsverteilungen zweidimensionaler Daten VI

- Kontingenztafeln werden oft durch die Randhäufigkeiten, die sich dann als Zeilen- bzw. Spaltensummen ergeben, in der Form

$X \setminus Y$	b_1	b_2	\dots	b_l	$h_{i\cdot}$
a_1	h_{11}	h_{12}	\dots	h_{1l}	$h_{1\cdot}$
a_2	h_{21}	h_{22}	\dots	h_{2l}	$h_{2\cdot}$
\vdots	\vdots	\vdots	\ddots	\vdots	\vdots
a_k	h_{k1}	h_{k2}	\dots	h_{kl}	$h_{k\cdot}$
$h_{\cdot j}$	$h_{\cdot 1}$	$h_{\cdot 2}$	\dots	$h_{\cdot l}$	n

oder

$X \setminus Y$	b_1	b_2	\dots	b_l	$r_{i\cdot}$
a_1	r_{11}	r_{12}	\dots	r_{1l}	$r_{1\cdot}$
a_2	r_{21}	r_{22}	\dots	r_{2l}	$r_{2\cdot}$
\vdots	\vdots	\vdots	\ddots	\vdots	\vdots
a_k	r_{k1}	r_{k2}	\dots	r_{kl}	$r_{k\cdot}$
$r_{\cdot j}$	$r_{\cdot 1}$	$r_{\cdot 2}$	\dots	$r_{\cdot l}$	1

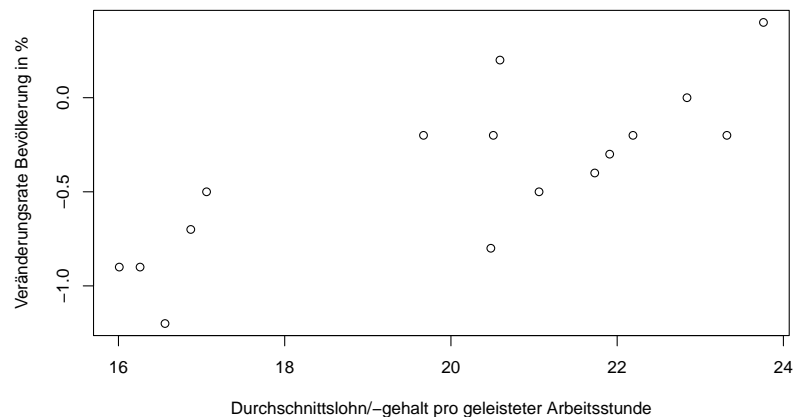
ergänzt.

- Zur besseren Abgrenzung von Randhäufigkeiten nennt man h_{ij} bzw. r_{ij} oft auch **gemeinsame absolute** bzw. **relative Häufigkeiten**.

- Zur Visualisierung zweidimensionaler Daten mit (überwiegend) paarweise verschiedenen Ausprägungen (z.B. bei zwei stetigen Merkmalen):

Streudiagramm bzw. Scatter-Plot

Durchschnittslohn vs. Bevölkerungswachstum nach Bundesländern 2009



- Bei mehr als zwei Merkmalen: Paarweise Streudiagramme üblich.

Beispiel (Kontingenztafel)

- Merkmal X: Mathematiknote,
Merkmal Y: Physiknote,

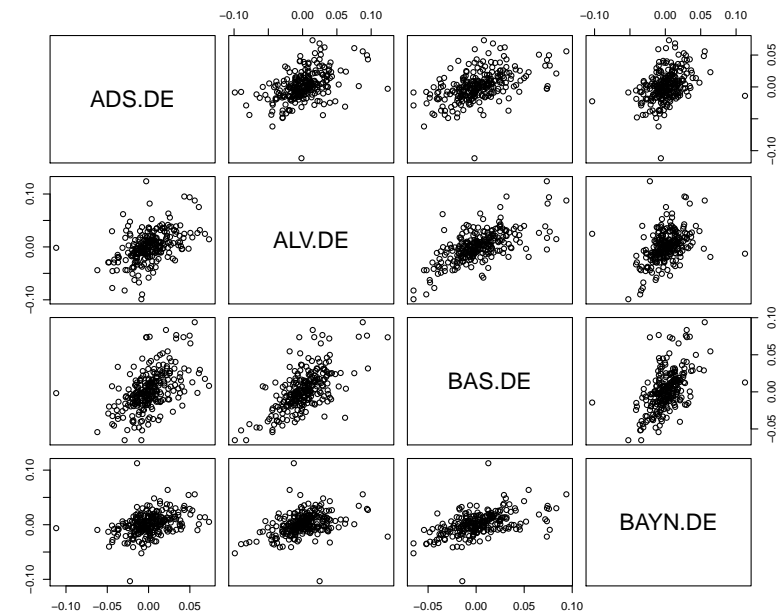
Urliste zum zweidimensionalen Merkmal (X, Y) :

$(2, 2), (2, 3), (3, 3), (5, 3), (2, 3), (5, 4), (5, 5), (4, 2), (4, 4), (1, 2), (2, 3),$
 $(1, 3), (4, 4), (2, 3), (4, 4), (3, 4), (4, 2), (5, 4), (2, 3), (4, 4), (5, 4), (2, 3),$
 $(4, 3), (1, 1), (2, 1), (2, 2), (1, 1), (2, 3), (5, 4), (2, 2)$

- Kontingenztafel (mit Randhäufigkeiten)

$X \setminus Y$	1	2	3	4	5	$h_{i\cdot}$
1	2	1	1	0	0	4
2	1	3	7	0	0	11
3	0	0	1	1	0	2
4	0	2	1	4	0	7
5	0	0	1	4	1	6
$h_{\cdot j}$	3	6	11	9	1	30

Tagesrenditen (2008) verschiedener DAX-Papiere



Klassierung mehrdimensionaler Daten

- Analog zu eindimensionalen Daten: Häufigkeitstabellen schlecht geeignet für Merkmale mit „vielen“ verschiedenen Ausprägungen, also insbesondere stetige Merkmale.
- Genauso wie bei eindimensionalen Daten möglich: **Klassierung**
- Klassierung für mehrdimensionale Daten wird hier nicht mehr im Detail ausgeführt
- Allgemeine „Anleitungen“ für Klassierungen mehrdimensionaler Daten:
 - ▶ Oft werden nicht alle Merkmale klassiert, sondern nur einzelne.
 - ▶ Klassierung erfolgt individuell für jedes zu klassierende Merkmal.
 - ▶ Anwendung von Verfahren für nominalskalierte und ordinalskalierte Merkmale klar.
 - ▶ Bei Verfahren für kardinalskalierte Daten: Annahme gleichmäßiger Verteilung der Merkmalswerte auf Klassen, ggf. Klassenmitte als Näherung für die Merkmalsausprägungen (wie bisher!)

Bedingte Häufigkeitsverteilungen II

- Für festes $j \in \{1, \dots, l\}$ entsprechen die bedingten relativen Häufigkeiten $r(a_i|Y = b_j)$ also den relativen Häufigkeiten von Merkmal X bei *Einschränkung der statistischen Masse auf die Merkmalsträger, für die das Merkmal Y die Ausprägung b_j annimmt.*
- Umgekehrt entsprechen für festes $i \in \{1, \dots, k\}$ die bedingten relativen Häufigkeiten $r(b_j|X = a_i)$ den relativen Häufigkeiten von Merkmal Y bei *Einschränkung der statistischen Masse auf die Merkmalsträger, für die das Merkmal X die Ausprägung a_i annimmt.*
- Man nennt die Merkmale X und Y *unabhängig*, wenn diese Einschränkungen keinen Effekt auf die relativen Häufigkeiten haben, d.h. alle bedingten relativen Häufigkeiten mit den zugehörigen relativen Randhäufigkeiten übereinstimmen.

Bedingte Häufigkeitsverteilungen I

- Ziel einer gemeinsamen Betrachtung von mehreren Merkmalen: *Untersuchung von Abhängigkeiten und Zusammenhängen*
- **Zentrale Frage:** Hängen die jeweils angenommenen Merkmalswerte eines Merkmals X mit denen eines anderen Merkmals Y zusammen?
- Untersuchungsmöglichkeit auf Grundlage gemeinsamer Häufigkeiten zu den Merkmalen X mit Merkmalsraum $A = \{a_1, \dots, a_k\}$ und Y mit Merkmalsraum $B = \{b_1, \dots, b_l\}$: Bilden der **bedingten relativen Häufigkeiten**
 - ▶ $r(a_i|Y = b_j) := \frac{h_{ij}}{h_{.j}}$
 - ▶ $r(b_j|X = a_i) := \frac{h_{ij}}{h_{i.}}$
 für $i \in \{1, \dots, k\}$ und $j \in \{1, \dots, l\}$.

Beispiel (bedingte Häufigkeitsverteilungen)

- Von Folie 106: Merkmal X : Mathematiknote, Merkmal Y : Physiknote, Kontingenztabelle

$X \setminus Y$	1	2	3	4	5	$h_{i.}$
1	2	1	1	0	0	4
2	1	3	7	0	0	11
3	0	0	1	1	0	2
4	0	2	1	4	0	7
5	0	0	1	4	1	6
$h_{.j}$	3	6	11	9	1	30

- Tabelle mit bedingten Häufigkeitsverteilungen für $Y|X = a_i$:

b_j	1	2	3	4	5	Σ
$r(b_j X = 1)$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$	0	0	1
$r(b_j X = 2)$	$\frac{1}{11}$	$\frac{3}{11}$	$\frac{7}{11}$	0	0	1
$r(b_j X = 3)$	0	0	$\frac{1}{2}$	$\frac{1}{2}$	0	1
$r(b_j X = 4)$	0	$\frac{2}{7}$	$\frac{1}{7}$	$\frac{4}{7}$	0	1
$r(b_j X = 5)$	0	0	$\frac{1}{6}$	$\frac{2}{3}$	$\frac{1}{6}$	1

Unabhängigkeit von zwei Merkmalen I

Definition 4.1 (Unabhängigkeit von zwei Merkmalen)

Die Merkmale X mit Merkmalsraum $A = \{a_1, \dots, a_k\}$ und Y mit Merkmalsraum $B = \{b_1, \dots, b_l\}$ eines zweidimensionalen Merkmals (X, Y) zu einer Urliste der Länge n heißen **unabhängig**, falls

$$r(a_i|Y = b_j) = \frac{h_{ij}}{h_{.j}} \stackrel{!}{=} \frac{h_{i.}}{n} = r(a_i)$$

bzw. (gleichbedeutend)

$$r(b_j|X = a_i) = \frac{h_{ij}}{h_{i.}} \stackrel{!}{=} \frac{h_{.j}}{n} = r(b_j)$$

für alle $i \in \{1, \dots, k\}$ und $j \in \{1, \dots, l\}$ gilt.

Abhängigkeitsmaße

- Je nach Skalierungsniveau der Merkmale X und Y können verschiedene Verfahren zur Messung der Abhängigkeit verwendet werden, das niedrigste Skalierungsniveau (nominal \prec ordinal \prec kardinal) ist dabei für die Einschränkung der geeigneten Verfahren maßgeblich:
 - ▶ Verfahren für ordinalskalierte Merkmale können nur dann eingesetzt werden, wenn beide Merkmale X und Y mindestens ordinalskaliert sind.
 - ▶ Verfahren für kardinalskalierte Merkmale können nur dann eingesetzt werden, wenn beide Merkmale X und Y kardinalskaliert sind.
- Trotz unterschiedlicher Wertebereiche der Abhängigkeitsmaße besteht die Gemeinsamkeit, dass die Abhängigkeit von X und Y stets mit dem Wert 0 gemessen wird, falls X und Y unabhängig gemäß Definition 4.1 sind.
- *Vorsicht beim Ableiten von Kausalitätsbeziehungen (Wirkungsrichtungen) aus entdeckten Abhängigkeiten!*

Unabhängigkeit von zwei Merkmalen II

- Die Bedingungen in Definition 4.1 sind offensichtlich genau dann erfüllt, wenn $h_{ij} = \frac{h_{i.} \cdot h_{.j}}{n}$ bzw. $r_{ij} = r_{i.} \cdot r_{.j}$ für alle $i \in \{1, \dots, k\}$ und $j \in \{1, \dots, l\}$ gilt, die gemeinsamen relativen Häufigkeiten sich also als Produkt der relativen Randhäufigkeiten ergeben.
- Unabhängigkeit im Sinne von Definition 4.1 ist eher ein idealtypisches Konzept und in der Praxis kaum erfüllt.
- Interessant sind daher Maße, die vorhandene Abhängigkeiten zwischen zwei Merkmalen näher quantifizieren.

Kardinalskalierte Merkmale

Definition 4.2 (emp. Kovarianz, Pearsonscher Korrelationskoeffizient)

Gegeben sei das zweidimensionale Merkmal (X, Y) mit der Urliste $(x_1, y_1), \dots, (x_n, y_n)$ der Länge n , X und Y seien kardinalskaliert. Mit $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ bzw. $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$ seien wie üblich die arithmetischen Mittelwerte von X bzw. Y bezeichnet, mit

$$s_X = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2} \quad \text{bzw.} \quad s_Y = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2}$$

die jeweiligen empirischen Standardabweichungen. Dann heißen

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

die **empirische Kovarianz** von X und Y und

$$r_{X,Y} := \frac{s_{X,Y}}{s_X \cdot s_Y} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$$

der (**Bravais-**)**Pearsonsche Korrelationskoeffizient** von X und Y .

Bemerkungen I

- $s_{X,Y}$ kann meist einfacher gemäß $s_{X,Y} = \overline{xy} - \bar{x} \cdot \bar{y}$ mit

$$\overline{xy} := \frac{1}{n} \sum_{i=1}^n x_i \cdot y_i$$

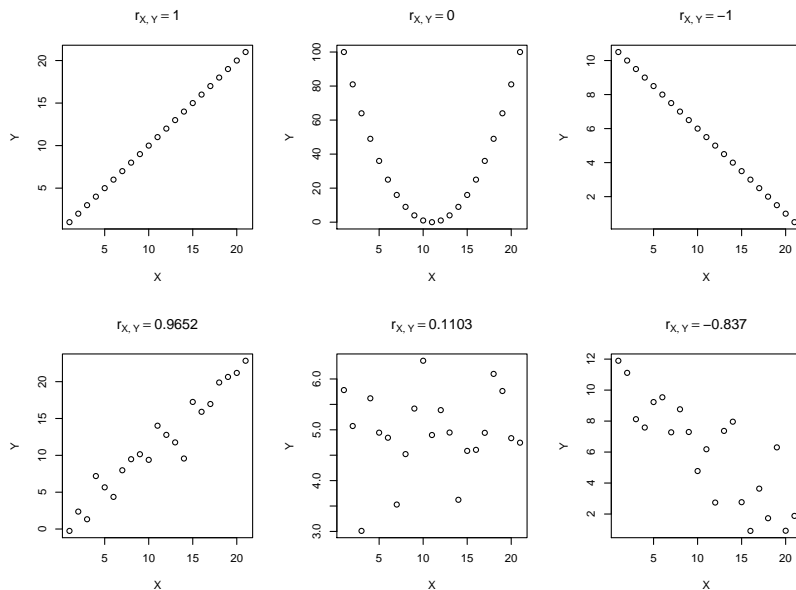
berechnet werden.

- Bei Vorliegen der Häufigkeitsverteilung kann \overline{xy} einfacher gemäß

$$\overline{xy} = \frac{1}{n} \sum_{i=1}^k \sum_{j=1}^l a_i b_j \cdot h_{ij} = \sum_{i=1}^k \sum_{j=1}^l a_i b_j \cdot r_{ij}$$

berechnet werden (\bar{x} und \bar{y} werden zur Berechnung von $s_{X,Y}$ dann zweckmäßigerweise ebenfalls mit Hilfe der Häufigkeitsverteilungen berechnet, siehe dazu Folie 67).

Beispiel: Pearsonscher Korrelationskoeffizient



Bemerkungen II

- Es gilt **stets** $-1 \leq r_{X,Y} \leq 1$.
- $r_{X,Y}$ misst **lineare** Zusammenhänge, spezieller gilt
 - ▶ $r_{X,Y} > 0$ bei positiver „Steigung“ („X und Y sind **positiv korreliert**“),
 - ▶ $r_{X,Y} < 0$ bei negativer „Steigung“ („X und Y sind **negativ korreliert**“),
 - ▶ $|r_{X,Y}| = 1$, falls alle (x_i, y_i) auf einer Geraden (mit Steigung $\neq 0$) liegen.
- $r_{X,Y}$ ist nur definiert, wenn X und Y jeweils mindestens zwei verschiedene Merkmalsausprägungen besitzen.

(Mindestens) ordinalskalierte Merkmale I

- Messen *linearer* Zusammenhänge bei Ordinalskala nicht (mehr) möglich, stattdessen: Messen *monotoner* Zusammenhänge.
- Hierzu für X und Y erforderlich: Bilden der **Ränge** der Merkmalswerte (gemäß der vorgegebenen Ordnung).
- Aus den Merkmalen X und Y mit Merkmalswerten x_1, \dots, x_n bzw. y_1, \dots, y_n werden dabei neue Merkmale $rg(X)$ und $rg(Y)$ mit Merkmalswerten $rg(X)_1, \dots, rg(X)_n$ bzw. $rg(Y)_1, \dots, rg(Y)_n$.
- Bilden der Ränge wird exemplarisch für Merkmal X beschrieben (Bilden der Ränge für Y ganz analog).

(Mindestens) ordinalskalierte Merkmale II

- 1 Einfacher Fall: *Alle n Merkmalswerte sind verschieden.*
 \rightsquigarrow Ränge von 1 bis n werden den Merkmalswerten nach der Position in der gemäß der vorgegebenen Ordnung sortierten Urliste zugewiesen:

$$x_{(1)} \mapsto 1, \dots, x_{(n)} \mapsto n$$

- 2 Komplizierter Fall: Es existieren mehrfach auftretende Merkmalswerte (sog. *Bindungen*), d.h. es gilt $x_i = x_j$ für (mindestens) ein Paar (i, j) mit $i \neq j$.
 \rightsquigarrow Prinzipielle Vorgehensweise wie im einfachen Fall, Ränge übereinstimmender Merkmalswerte müssen aber (arithmetisch) gemittelt werden.

(Mindestens) ordinalskalierte Merkmale IV

- Der zweite (subtrahierte) Term $\frac{\#\{j \in \{1, \dots, n\} \mid x_j = x_i\} - 1}{2}$ bzw. $\frac{h(x_i) - 1}{2}$ in Definition 4.3 dient der Berechnung des arithmetischen Mittels bei Vorliegen von Bindungen.
- Liegen keine Bindungen vor (sind also alle Merkmalswerte verschieden), ist der zweite (subtrahierte) Term in Definition 4.3 immer gleich 0.
- Idee zur Konstruktion eines Abhängigkeitsmaßes für (mindestens) ordinalskalierte zweidimensionale Merkmale (X, Y) :
 - 1 Übergang von X zu $\text{rg}(X)$ sowie von Y zu $\text{rg}(Y)$
 - 2 Berechnung des Pearsonschen Korrelationskoeffizienten von $\text{rg}(X)$ und $\text{rg}(Y)$

(Mindestens) ordinalskalierte Merkmale III

- „Berechnungsvorschrift“ für beide Fälle in folgender Definition:

Definition 4.3 (Rang eines Merkmals X , $\text{rg}(X)_i$)

Gegeben sei ein Merkmal X mit Urliste x_1, \dots, x_n . Dann heißt für $i \in \{1, \dots, n\}$

$$\begin{aligned} \text{rg}(X)_i &:= \#\{j \in \{1, \dots, n\} \mid x_j \leq x_i\} - \frac{\#\{j \in \{1, \dots, n\} \mid x_j = x_i\} - 1}{2} \\ &= \sum_{\substack{a_j \leq x_i \\ 1 \leq j \leq n}} h(a_j) - \frac{h(x_i) - 1}{2} \\ &= n \cdot F(x_i) - \frac{h(x_i) - 1}{2} \end{aligned}$$

der **Rang** von x_i . Die Werte $\text{rg}(X)_1, \dots, \text{rg}(X)_n$ können als Urliste zu einem neuen Merkmal $\text{rg}(X)$ aufgefasst werden.

Spearmanischer Rangkorrelationskoeffizient I

Definition 4.4 (Spearmanischer Rangkorrelationskoeffizient)

Gegeben sei das zweidimensionale Merkmal (X, Y) mit der Urliste $(x_1, y_1), \dots, (x_n, y_n)$ der Länge n , X und Y seien (mindestens) ordinalskaliert. Zu X und Y seien die Ränge $\text{rg}(X)$ und $\text{rg}(Y)$ gemäß Definition 4.3 gegeben. Dann heißt

$$r_{X,Y}^{(S)} := r_{\text{rg}(X), \text{rg}(Y)} = \frac{s_{\text{rg}(X), \text{rg}(Y)}}{s_{\text{rg}(X)} \cdot s_{\text{rg}(Y)}}$$

der **Spearmanische Rangkorrelationskoeffizient** von X und Y .

- Wegen des Zusammenhangs mit dem Pearsonschen Korrelationskoeffizienten gilt offensichtlich stets

$$-1 \leq r_{X,Y}^{(S)} \leq 1.$$

Spearman'scher Rangkorrelationskoeffizient II

- Bei der Berechnung von $r_{X,Y}^{(S)}$ kann die Eigenschaft

$$\overline{\text{rg}(X)} = \overline{\text{rg}(Y)} = \frac{n+1}{2}$$

ausgenutzt werden.

- Damit gilt für $r_{X,Y}^{(S)}$ stets:

$$r_{X,Y}^{(S)} = \frac{\frac{1}{n} \sum_{i=1}^n \text{rg}(X)_i \cdot \text{rg}(Y)_i - \frac{(n+1)^2}{4}}{\sqrt{\left[\frac{1}{n} \sum_{i=1}^n (\text{rg}(X)_i)^2 - \frac{(n+1)^2}{4} \right] \cdot \left[\frac{1}{n} \sum_{i=1}^n (\text{rg}(Y)_i)^2 - \frac{(n+1)^2}{4} \right]}}$$

- Nur** wenn $x_i \neq x_j$ und $y_i \neq y_j$ für alle $i \neq j$ gilt (also **keine** Bindungen vorliegen), gilt die wesentlich leichter zu berechnende „Formel“

$$r_{X,Y}^{(S)} = 1 - \frac{6 \cdot \sum_{i=1}^n (\text{rg}(X)_i - \text{rg}(Y)_i)^2}{n \cdot (n^2 - 1)}$$

(Mindestens) nominalskalierte Merkmale I

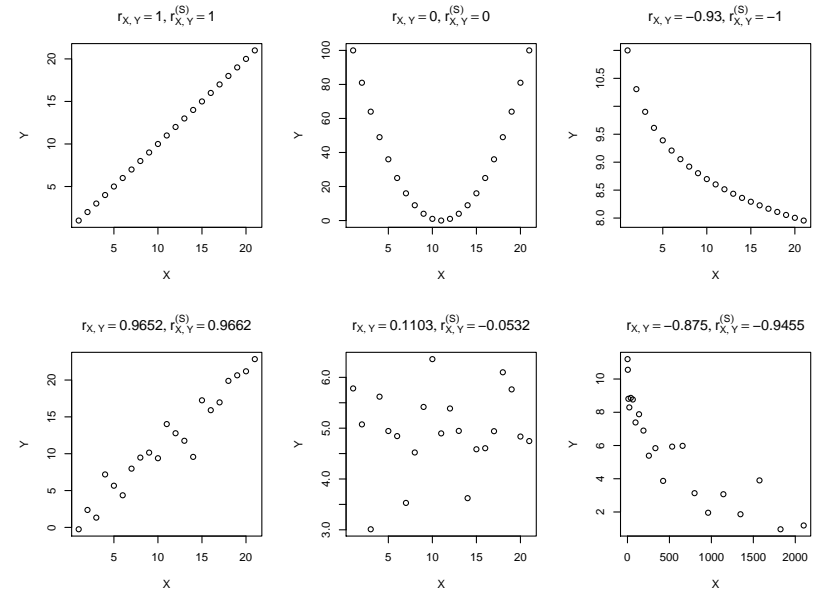
- Da bei nominalskalierten Merkmalen keine Ordnung vorgegeben ist, kann hier lediglich die *Stärke*, nicht aber die *Richtung* der Abhängigkeit zwischen X und Y gemessen werden.
- Idee zur Konstruktion eines Abhängigkeitsmaßes auf Basis der gemeinsamen Häufigkeitstabelle zu (X, Y) :
 - Bei Unabhängigkeit der Merkmale X und Y müsste nach Definition 4.1 auf Folie 113

$$h_{ij} = \frac{h_{i \cdot} \cdot h_{\cdot j}}{n} \quad \text{für alle } i \in \{1, \dots, k\}, j \in \{1, \dots, l\}$$

gelten.

- Abweichungen zwischen h_{ij} und $\frac{h_{i \cdot} \cdot h_{\cdot j}}{n}$ können also zur Messung der Abhängigkeit eingesetzt werden.
- Hier verwendetes Abhängigkeitsmaß entsteht aus geeigneter Zusammenfassung und Normierung dieser Abweichungen.

Beispiel: Spearman'scher Rangkorrelationskoeffizient



(Mindestens) nominalskalierte Merkmale II

Definition 4.5 (Pearson'scher Kontingenzkoeffizient)

Gegeben sei das zweidimensionale Merkmal (X, Y) zu einer Urliste der Länge n mit den zugehörigen absoluten gemeinsamen Häufigkeiten h_{ij} sowie den Randhäufigkeiten $h_{i \cdot}$ und $h_{\cdot j}$ für $i \in \{1, \dots, k\}, j \in \{1, \dots, l\}$.

Dann heißt

$$C_{X,Y}^{\text{korr}} := \sqrt{\frac{\min\{k, l\}}{\min\{k, l\} - 1} \cdot \frac{\chi^2}{n + \chi^2}}$$

mit

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

korrigierter Pearson'scher Kontingenzkoeffizient der Merkmale X und Y .

- Es gilt stets $0 \leq C_{X,Y}^{\text{korr}} \leq 1$.