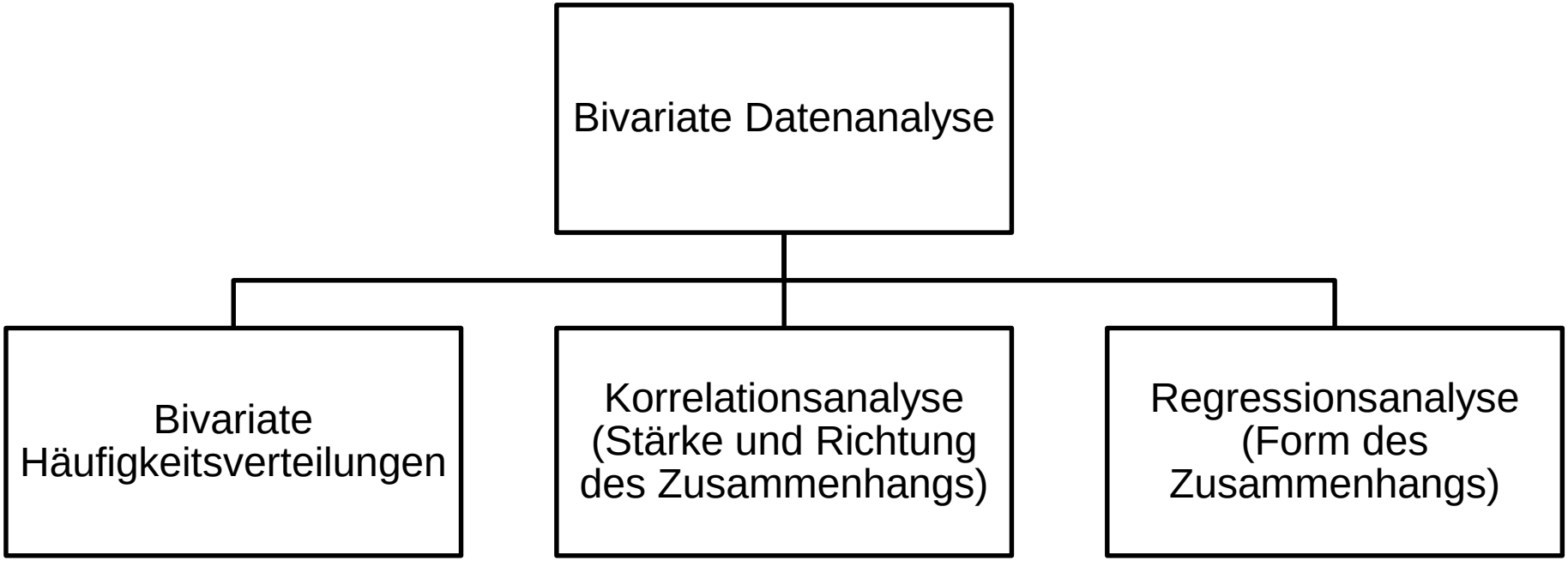
7. Bivariate Häufigkeitsverteilungen

7.1 Gemeinsame Häufigkeiten und Randhäufigkeiten

- Bisher: Betrachtung eines Merkmals
- Bei Erhebungen werden aber in der Regel mehrere Merkmale bei statistischen Einheiten gleichzeitig erfasst
- Bivariate Datenanalyse: Analyse des Zusammenhangs zwischen zwei Merkmalen



Bivariate Daten: Verbundene Beobachtungen zweier Merkmale X und Y

i	1	2	3	4	5	6	7	…	n
x_i	x_1	x_2	x_3	x_4	x_5	x_6	x_7	…	x_n
y_i	y_1	y_2	y_3	y_4	y_5	y_6	y_7	…	y_n

Beispiele:

- Ein Unternehmen könnte interessieren, ob zwischen den Merkmalen Schulabschluss und Kaufhäufigkeit ein Zusammenhang besteht. Falls höher gebildete Personen ein Produkt häufiger kaufen, könnte der Marketing-Mix (z.B. Werbung) auf diese Personengruppe zugeschnitten werden.
- Eine Autoversicherung will ihren Kundinnen und Kunden Tarife anbieten, die dem Unfallrisiko entsprechen. Ein Kunde mit einem höheren Unfallrisiko soll eine höhere Prämie zahlen. Zur Berechnung des Tarifs werden Unfallbeteiligte (statistische Einheiten) nach Alter und Geschlecht ausgewertet. Aufgrund des höheren Unfallrisikos von jüngeren Männern wird diesen ein höherer Tarif auferlegt.

Beispiel: Geschlecht und Kaufhäufigkeit

Ein Unternehmen möchte wissen, ob ein Zusammenhang zwischen Geschlecht und Kaufhäufigkeit eines Produktes besteht. Hierfür hat es 16 Kund:innen befragt. Der erste Befragte hat beispielsweise zweimal gekauft und war männlich. Alle 16 verbundenen Beobachtungswerte sind in folgender Tabelle dargestellt:

i	x_i (Kaufhäufigkeit)	y_i (Geschlecht)	i	x_i (Kaufhäufigkeit)	y_i (Geschlecht)
1	2	männlich	9	2	männlich
2	3	männlich	10	1	weiblich
3	2	weiblich	11	3	weiblich
4	3	weiblich	12	2	weiblich
5	1	weiblich	13	1	männlich
6	2	männlich	14	3	männlich
7	3	weiblich	15	3	weiblich
8	2	männlich	16	1	weiblich

Strukturierung der verbundenen Beobachtungen

- Merkmal X: r Merkmalsausprägungen (oder repräsentative Klassenwerte)

$$x_j^*, j = 1, 2, 3, \cdots, r$$

- Merkmal Y: c Merkmalsausprägungen (oder repräsentative Klassenwerte)

$$y_k^*, k = 1, 2, 3, \cdots, c$$

- Die **absolute gemeinsame Häufigkeit $\boldsymbol{n_{jk}}$** gibt die Anzahl der statistischen Einheiten mit der j-ten Merkmalsausprägung beim Merkmal X und der k-ten Merkmalsausprägung beim Merkmal Y an.

- Die **absolute Randhäufigkeit $\boldsymbol{n_{j\bullet}}$** gibt an, bei wie vielen statistischen Einheiten die j-te Merkmalsausprägung des Merkmals X gemessen wurde (das Merkmal Y bleibt unberücksichtigt).

$$n_{j\bullet} = \sum_{k=1}^{c} n_{jk} \qquad \text{Summierung über alle Spalten}$$

- Entsprechend gibt die **absolute Randhäufigkeit $\boldsymbol{n_{\bullet k}}$** die Anzahl der statistischen Einheiten mit der k-ten Merkmalsausprägung des Merkmals Y an.

$$n_{\bullet k} = \sum_{j=1}^{r} n_{jk} \qquad \text{Summierung über alle Zeilen}$$

Zweidimensionale absolute Häufigkeitsverteilung

X \ Y	y_1^*	y_2^*	…	y_c^*	$\sum_{k=1}^{c} n_{jk}$
x_1^*	n_{11}	n_{12}	…	n_{1c}	$n_{1\bullet} = \sum_{k=1}^{c} n_{1k}$
x_2^*	n_{21}	n_{22}	…	n_{2c}	$n_{2\bullet} = \sum_{k=1}^{c} n_{2k}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
x_r^*	n_{r1}	n_{r2}	…	n_{rc}	$n_{r\bullet} = \sum_{k=1}^{c} n_{rk}$
$\sum_{j=1}^{r} n_{jk}$	$n_{\bullet 1} = \sum_{j=1}^{r} n_{j1}$	$n_{\bullet 2} = \sum_{j=1}^{r} n_{j2}$	…	$n_{\bullet c} = \sum_{j=1}^{r} n_{jc}$	$n = \sum_{j=1}^{r}\sum_{k=1}^{c} n_{jk} = \sum_{j=1}^{r} n_{j\bullet} = \sum_{k=1}^{c} n_{\bullet k}$

Beispiel (Fortsetzung): Geschlecht und Kaufhäufigkeit (I)

Auch bei zweidimensionalen Häufigkeitsverteilungen bietet sich oft die Verwendung von Strichlisten zur Auszählung an:

X (Kaufhäufigkeit) \ Y (Geschlecht)	y_1^* (männlich)	y_2^* (weiblich)
x_1^* (einmal)	I	III
x_2^* (zweimal)	IIII	II
x_3^* (dreimal)	II	IIII

Beispiel (Fortsetzung): Geschlecht und Kaufhäufigkeit (II)

Zweidimensionale Häufigkeitsverteilung absoluter Häufigkeiten:

X \ Y	y_1^* (männlich)	y_2^* (weiblich)	$\sum_{k=1}^{2} n_{jk}$
x_1^* (einmal)	$n_{11} = 1$	$n_{12} = 3$	$n_{1\bullet} = \sum_{k=1}^{2} n_{1k} = 4$
x_2^* (zweimal)	$n_{21} = 4$	$n_{22} = 2$	$n_{2\bullet} = \sum_{k=1}^{2} n_{2k} = 6$
x_3^* (dreimal)	$n_{31} = 2$	$n_{32} = 4$	$n_{3\bullet} = \sum_{k=1}^{2} n_{3k} = 6$
$\sum_{j=1}^{3} n_{jk}$	$n_{\bullet 1} = \sum_{j=1}^{3} n_{j1} = 7$	$n_{\bullet 2} = \sum_{j=1}^{3} n_{j2} = 9$	$n = \sum_{j=1}^{3} \sum_{k=1}^{2} n_{jk} = 16$

Beispiel (Fortsetzung): Geschlecht und Kaufhäufigkeit (III)

Interpretation:

- $n_{11} = 1$: Eine Person hat das Produkt einmal gekauft und ist männlich.
- $n_{32} = 4$: Vier Personen haben das Produkt dreimal gekauft und sind weiblich.
- $n_{\bullet 1} = 7$: Sieben Personen sind männlich.
- $n_{1\bullet} = 4$: Vier Personen haben das Produkt einmal gekauft.

Relative gemeinsame Häufigkeiten

- Die **relativen gemeinsamen Häufigkeiten $\boldsymbol{h_{jk}}$** geben den Anteil der statistischen Einheiten mit der Merkmalskombination x_j^* und y_k^* an.
- **Relative Randhäufigkeit** von X:

$$h_{j\bullet} = \frac{n_{j\bullet}}{n} \text{ oder } h_{j\bullet} = \sum_{k=1}^{c} h_{jk} \qquad \text{Summierung über alle Spalten}$$

- **Relative Randhäufigkeit** von Y:

$$h_{\bullet k} = \frac{n_{\bullet k}}{n} \text{ oder } h_{\bullet k} = \sum_{j=1}^{r} h_{jk} \qquad \text{Summierung über alle Zeilen}$$

Zweidimensionale relative Häufigkeitsverteilung

X \ Y	y_1^*	y_2^*	…	y_c^*	$\sum_{k=1}^{c} h_{jk}$
x_1^*	h_{11}	h_{12}	…	h_{1c}	$h_{1\bullet} = \sum_{k=1}^{c} h_{1k}$
x_2^*	h_{21}	h_{22}	…	h_{2c}	$h_{2\bullet} = \sum_{k=1}^{c} h_{2k}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
x_r^*	h_{r1}	h_{r2}	…	h_{rc}	$h_{r\bullet} = \sum_{k=1}^{c} h_{rk}$
$\sum_{j=1}^{r} h_{jk}$	$h_{\bullet 1} = \sum_{j=1}^{r} h_{j1}$	$h_{\bullet 2} = \sum_{j=1}^{r} h_{j2}$	…	$h_{\bullet c} = \sum_{j=1}^{r} h_{jc}$	$1 = \sum_{j=1}^{r} \sum_{k=1}^{c} h_{jk} = \sum_{j=1}^{r} h_{j\bullet} = \sum_{k=1}^{c} h_{\bullet k}$

Beispiel (Fortsetzung): Geschlecht und Kaufhäufigkeit (I)

Zweidimensionale Häufigkeitsverteilung relativer Häufigkeiten:

X \ Y	y_1^* (männlich)	y_2^* (weiblich)	$\sum_{k=1}^{2} h_{jk}$
x_1^* (einmal)	$h_{11} = 1/16$	$h_{12} = 3/16$	$h_{1\bullet} = \sum_{k=1}^{2} h_{1k} = 4/16$
x_2^* (zweimal)	$h_{21} = 4/16$	$h_{22} = 2/16$	$h_{2\bullet} = \sum_{k=1}^{2} h_{2k} = 6/16$
x_3^* (dreimal)	$h_{31} = 2/16$	$h_{32} = 4/16$	$h_{3\bullet} = \sum_{k=1}^{2} h_{3k} = 6/16$
$\sum_{j=1}^{3} h_{jk}$	$h_{\bullet 1} = \sum_{j=1}^{3} h_{j1} = 7/16$	$h_{\bullet 2} = \sum_{j=1}^{3} h_{j2} = 9/16$	$\sum_{j=1}^{3}\sum_{k=1}^{2} h_{jk} = 1$

Beispiel (Fortsetzung): Geschlecht und Kaufhäufigkeit (II)

Interpretation:

- $h_{11} = 1/16 = 0{,}0625$: 6,25% der Personen haben das Produkt einmal gekauft und sind männlich.
- $h_{32} = 4/16 = 0{,}25$: 25% der Personen haben das Produkt dreimal gekauft und sind weiblich.
- $h_{\bullet 1} = 7/16 = 0{,}4375$: 43,75% der Personen sind männlich.
- $h_{1\bullet} = 4/16 = 0{,}25$: 25% der Personen haben das Produkt einmal gekauft.

7.2 Bedingte Häufigkeiten und statistische Unabhängigkeit

- **Bedingte (relative) Häufigkeitsverteilung**:
  Verteilung eines Merkmals bei gegebener Ausprägung eines anderen Merkmals

- Bedingte (relative) Häufigkeiten von y_k^* gegeben x_j^*:

$$h\left(y_k^* \middle| x_j^*\right) = \frac{n_{jk}}{n_{j\bullet}} = \frac{h_{jk}}{h_{j\bullet}}, \qquad j = 1, 2, \cdots, r; \;\; k = 1, 2, \cdots, c$$

  Dabei werden nun nicht alle statistischen Einheiten, sondern nur diejenigen mit der Merkmalsausprägung x_j^* als Bezugsbasis gewählt.

- Bedingte (relative) Häufigkeiten von x_j^* gegeben y_k^*:

$$h\left(x_j^* \middle| y_k^*\right) = \frac{n_{jk}}{n_{\bullet k}} = \frac{h_{jk}}{h_{\bullet k}}, \qquad j = 1, 2, \cdots, r; \;\; k = 1, 2, \cdots, c$$

  Dabei werden nun nicht alle statistischen Einheiten, sondern nur diejenigen mit der Merkmalsausprägung y_k^* als Bezugsbasis gewählt.

Bedingte Häufigkeitsverteilung von Y gegeben x_j^*

X \ Y	y_1^*	y_2^*	…	y_c^*	$\sum_{k=1}^{c}$
Bedingung: x_1^*	$h(y_1^*\|x_1^*)$	$h(y_2^*\|x_1^*)$	…	$h(y_c^*\|x_1^*)$	1
Bedingung: x_2^*	$h(y_1^*\|x_2^*)$	$h(y_2^*\|x_2^*)$	…	$h(y_c^*\|x_2^*)$	1
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
Bedingung: x_r^*	$h(y_1^*\|x_r^*)$	$h(y_2^*\|x_r^*)$	…	$h(y_c^*\|x_r^*)$	1

Bedingte Häufigkeitsverteilung von X gegeben y_k^*

X \ Y	Bedingung: y_1^*	Bedingung: y_2^*	…	Bedingung: y_c^*
x_1^*	$h(x_1^*\|y_1^*)$	$h(x_1^*\|y_2^*)$	…	$h(x_1^*\|y_c^*)$
x_2^*	$h(x_2^*\|y_1^*)$	$h(x_2^*\|y_2^*)$	…	$h(x_2^*\|y_c^*)$
$\vdots$	$\vdots$	$\vdots$		$\vdots$
x_r^*	$h(x_r^*\|y_1^*)$	$h(x_r^*\|y_2^*)$	…	$h(x_r^*\|y_c^*)$
$\sum_{j=1}^{r}$	1	1	…	1

Beispiel (Fortsetzung): Geschlecht und Kaufhäufigkeit (I)

Zunächst werden die Verteilungen des Geschlechts (Y) unter der Bedingung der einzelnen Kaufkategorien (x_j^*) bestimmt.

X \ Y	y_1^* (männlich)	y_2^* (weibllich)	$\sum_{k=1}^{2}$
Bedingung: x_1^* (einmal)	$h(y_1^*\|x_1^*) = \frac{n_{11}}{n_{1\bullet}} = \frac{1}{4} = \frac{h_{11}}{h_{1\bullet}} = \frac{0{,}0625}{0{,}25} = 0{,}25$	$h(y_2^*\|x_1^*) = \frac{n_{12}}{n_{1\bullet}} = \frac{3}{4} = \frac{h_{12}}{h_{1\bullet}} = \frac{0{,}1875}{0{,}25} = 0{,}75$	$0{,}25 + 0{,}75 = 1$
Bedingung: x_2^* (zweimal)	$h(y_1^*\|x_2^*) = \frac{n_{21}}{n_{2\bullet}} = \frac{4}{6} = \frac{h_{21}}{h_{2\bullet}} = \frac{0{,}25}{0{,}375} = 0{,}667$	$h(y_2^*\|x_2^*) = \frac{n_{22}}{n_{2\bullet}} = \frac{2}{6} = \frac{h_{22}}{h_{2\bullet}} = \frac{0{,}125}{0{,}375} = 0{,}333$	$0{,}667 + 0{,}333 = 1$
Bedingung: x_3^* (dreimal)	$h(y_1^*\|x_3^*) = \frac{n_{31}}{n_{3\bullet}} = \frac{2}{6} = \frac{h_{31}}{h_{2\bullet}} = \frac{0{,}125}{0{,}375} = 0{,}333$	$h(y_2^*\|x_3^*) = \frac{n_{32}}{n_{3\bullet}} = \frac{4}{6} = \frac{h_{32}}{h_{3\bullet}} = \frac{0{,}25}{0{,}375} = 0{,}667$	$0{,}333 + 0{,}667 = 1$

Beispiel (Fortsetzung): Geschlecht und Kaufhäufigkeit (II)

Bedingte Verteilungen der Kaufhäufigkeit (X) in den Gruppen Männer (y_1^*) und Frauen (y_2^*).

X \ Y	Bedingung: y_1^* (männlich)	Bedingung: y_2^* (weibllich)
x_1^* (einmal)	$h(x_1^*\|y_1^*) = \frac{n_{11}}{n_{\bullet 1}} = \frac{1}{7} = 0{,}143$	$h(x_1^*\|y_2^*) = \frac{n_{12}}{n_{\bullet 2}} = \frac{3}{9} = 0{,}333$
x_2^* (zweimal)	$h(x_2^*\|y_1^*) = \frac{n_{21}}{n_{\bullet 1}} = \frac{4}{7} = 0{,}571$	$h(x_2^*\|y_2^*) = \frac{n_{22}}{n_{\bullet 2}} = \frac{2}{9} = 0{,}222$
x_3^* (dreimal)	$h(x_3^*\|y_1^*) = \frac{n_{31}}{n_{\bullet 1}} = \frac{2}{7} = 0{,}286$	$h(x_3^*\|y_2^*) = \frac{n_{32}}{n_{\bullet 2}} = \frac{4}{9} = 0{,}444$
$\sum_{j=1}^{3}$	$\frac{1}{7} + \frac{4}{7} + \frac{2}{7} = 1$	$\frac{3}{9} + \frac{2}{9} + \frac{4}{9} = 1$

Beispiel (Fortsetzung): Geschlecht und Kaufhäufigkeit (III)

Interpretation der bedingten relativen Häufigkeiten bei gegebenem x_j^*:

- $h(y_1^*|x_1^*) = 1/4 = 0{,}25$: Von den Personen, die das Produkt einmal gekauft haben, sind 25% Männer.
- $h(y_2^*|x_3^*) = 2/3 = 0{,}667$: Von den Personen, die das Produkt dreimal gekauft haben, sind 66,7% Frauen.

Interpretation der bedingten relativen Häufigkeiten bei gegebenem y_k^*:

- $h(x_1^*|y_1^*) = 1/7 = 0{,}143$: Von den Männern haben 14,3% das Produkt einmal gekauft.
- $h(x_2^*|y_2^*) = 2/9 = 0{,}222$: Von den Frauen haben 22,2% das Produkt zweimal gekauft.

Statistische Unabhängigkeit

- Statistische Unabhängigkeit liegt vor, wenn die bedingten (relativen) Häufigkeiten eines Merkmals für alle Teilgesamtheiten des anderen Merkmals übereinstimmen. Sie sind dann gleich der entsprechenden Randhäufigkeit.

- Andernfalls liegt eine statistische Abhängigkeit zwischen den beiden Merkmalen X und Y vor.

- Gleichheit der bedingten Häufigkeiten von Y bei gegebenem x_j^*:

$$h(y_k^*|x_1^*) = h(y_k^*|x_2^*) = \cdots = h(y_k^*|x_r^*) = h_{\bullet k}, \qquad \text{für alle } k = 1, 2, \cdots, c$$

$\rightarrow$ statistische Unabhängigkeit

- Gleichheit der bedingten Häufigkeiten von X bei gegebenem y_k^*:

$$h\left(x_j^*\middle|y_1^*\right) = h\left(x_j^*\middle|y_2^*\right) = \cdots = h\left(x_j^*\middle|y_c^*\right) = h_{j\bullet}, \qquad \text{für alle } j = 1, 2, \cdots, r$$

$\rightarrow$ statistische Unabhängigkeit

Beispiel (Fortsetzung): Geschlecht und Kaufhäufigkeit (I)

Die bedingten Häufigkeiten für die gegebenen x_j^* oder die gegebenen y_k^* können berechnet werden, um eine statistische Abhängigkeit zu prüfen. Wir gehen zunächst vom erstgenannten Fall aus.

X \ Y	y_1^* (männlich)	y_2^* (weibllich)	$\sum_{k=1}^{2}$
Bedingung: x_1^* (einmal)	$h(y_1^*\|x_1^*) = \frac{n_{11}}{n_{1\bullet}} = \frac{1}{4}$	$h(y_2^*\|x_1^*) = \frac{n_{12}}{n_{1\bullet}} = \frac{3}{4}$	1
Bedingung: x_2^* (zweimal)	$h(y_1^*\|x_2^*) = \frac{n_{21}}{n_{2\bullet}} = \frac{4}{6}$	$h(y_2^*\|x_2^*) = \frac{n_{22}}{n_{2\bullet}} = \frac{2}{6}$	1
Bedingung: x_3^* (dreimal)	$h(y_1^*\|x_3^*) = \frac{n_{31}}{n_{3\bullet}} = \frac{2}{6}$	$h(y_2^*\|x_3^*) = \frac{n_{32}}{n_{3\bullet}} = \frac{4}{6}$	1
Alle Personen	$h_{\bullet 1} = \frac{7}{16}$	$h_{\bullet 2} = \frac{9}{16}$	1

Gleichheit der Werte in den **Spalten** nicht gegeben → keine statistische Unabhängigkeit!

Beispiel (Fortsetzung): Geschlecht und Kaufhäufigkeit (II)

X \ Y	Bedingung: y_1^* (männlich)	Bedingung: y_2^* (weiblich)	Alle Personen
x_1^* (einmal)	$h(x_1^*\|y_1^*) = \frac{n_{11}}{n_{\bullet 1}} = \frac{1}{7} = 0{,}143$	$h(x_1^*\|y_2^*) = \frac{n_{12}}{n_{\bullet 2}} = \frac{3}{9} = 0{,}333$	$h_{1\bullet} = \frac{4}{16}$
x_2^* (zweimal)	$h(x_2^*\|y_1^*) = \frac{n_{21}}{n_{\bullet 1}} = \frac{4}{7} = 0{,}571$	$h(x_2^*\|y_2^*) = \frac{n_{22}}{n_{\bullet 2}} = \frac{2}{9} = 0{,}222$	$h_{2\bullet} = \frac{6}{16}$
x_3^* (dreimal)	$h(x_3^*\|y_1^*) = \frac{n_{31}}{n_{\bullet 1}} = \frac{2}{7} = 0{,}286$	$h(x_3^*\|y_2^*) = \frac{n_{32}}{n_{\bullet 2}} = \frac{4}{9} = 0{,}444$	$h_{3\bullet} = \frac{6}{16}$
$\sum_{j=1}^{3}$	1	1	1

Gleichheit der Werte in den **Zeilen** nicht gegeben → keine statistische Unabhängigkeit!

7.3 Bei Unabhängigkeit zu erwartende Häufigkeiten

- Bedingte (relative) Häufigkeiten von y_k^* gegeben x_j^*:

$$h\left(y_k^* \middle| x_j^*\right) = \frac{n_{jk}}{n_{j\bullet}} = \frac{h_{jk}}{h_{j\bullet}}, \qquad j = 1, 2, \cdots, r; \;\; k = 1, 2, \cdots, c$$

- Bei statistischer Unabhängigkeit gilt die Gleichheit der bedingten Häufigkeiten von Y bei gegebenem x_j^*:

$$h(y_k^*|x_1^*) = h(y_k^*|x_2^*) = \cdots = h(y_k^*|x_r^*) = h_{\bullet k}, \qquad \text{für alle } k = 1, 2, \cdots, c$$

- Daraus folgt:

$$\frac{h_{jk}}{h_{j\bullet}} = h_{\bullet k}$$

- Bei Auflösen nach h_{jk}:

$$\tilde{h}_{jk} = h_{\bullet k} \cdot h_{j\bullet} = \frac{n_{jk}}{n} = \frac{n_{j\bullet}}{n} \cdot \frac{n_{\bullet k}}{n}$$

- Multipliziert mit n:

$$\tilde{n}_{jk} = \frac{n_{j\bullet} \cdot n_{\bullet k}}{n}$$

Multiplikationsregel

- Erwartete relative Häufigkeiten:

$$\tilde{h}_{jk} = h_{\bullet k} \cdot h_{j\bullet}$$

- Erwartete absolute Häufigkeiten:

$$\tilde{n}_{jk} = \frac{n_{j\bullet} \cdot n_{\bullet k}}{n}$$

- Bei **statistischer Unabhängigkeit** entsprechen die zu erwartenden Häufigkeiten den beobachteten Häufigkeiten:

$$\tilde{h}_{jk} = h_{jk}, \qquad \text{für alle } j = 1, 2, \cdots, r; k = 1, 2, \cdots, c$$

bzw.

$$\tilde{n}_{jk} = n_{jk}, \qquad \text{für alle } j = 1, 2, \cdots, r; k = 1, 2, \cdots, c$$

Beispiel (Fortsetzung): Geschlecht und Kaufhäufigkeit (I)

Durch den Vergleich der bei Unabhängigkeit erwarteten und beobachteten absoluten Häufigkeiten sieht man, dass in einigen Zellen Unterschiede bestehen. So weicht die erwartete Häufigkeit bei den Männern mit einer zweifachen Kaufhäufigkeit mit 2,625 bspw. deutlich von dem beobachteten Wert mit 4 ab. Eine gleich große Abweichung besteht bei den Frauen mit einer Kaufhäufigkeit von zwei.

X \ Y	y_1^* (männlich)	y_2^* (weiblich)	$\sum_{k=1}^{2} n_{jk}$
x_1^* (einmal)	$\tilde{n}_{11} = \frac{4 \cdot 7}{16} = 1{,}75$ $(n_{11} = 1)$	$\tilde{n}_{12} = \frac{4 \cdot 9}{16} = 2{,}25$ $(n_{12} = 3)$	$n_{1\bullet} = 4$
x_2^* (zweimal)	$\tilde{n}_{21} = \frac{6 \cdot 7}{16} = 2{,}625$ $(n_{21} = 4)$	$\tilde{n}_{22} = \frac{6 \cdot 9}{16} = 3{,}375$ $(n_{22} = 2)$	$n_{2\bullet} = 6$
x_3^* (dreimal)	$\tilde{n}_{31} = \frac{6 \cdot 7}{16} = 2{,}625$ $(n_{31} = 2)$	$\tilde{n}_{32} = \frac{6 \cdot 9}{16} = 3{,}375$ $(n_{32} = 4)$	$n_{3\bullet} = 6$
$\sum_{j=1}^{3} n_{jk}$	$n_{\bullet 1} = 7$	$n_{\bullet 2} = 9$	$n = 16$

Beispiel (Fortsetzung): Geschlecht und Kaufhäufigkeit (II)

X \ Y	y_1^* (männlich)	y_2^* (weiblich)	$\sum_{k=1}^{2} h_{jk}$
x_1^* (einmal)	$\tilde{h}_{11} = 0{,}25 \cdot 0{,}4375 = 0{,}1094$ $(h_{11} = 0{,}0625)$	$\tilde{h}_{12} = 0{,}25 \cdot 0{,}5625 = 0{,}1406$ $(h_{12} = 0{,}1875)$	$h_{1\bullet} = 0{,}25$
x_2^* (zweimal)	$\tilde{h}_{21} = 0{,}375 \cdot 0{,}4375 = 0{,}1641$ $(h_{21} = 0{,}25)$	$\tilde{h}_{22} = 0{,}375 \cdot 0{,}5625 = 0{,}1406$ $(h_{22} = 0{,}125)$	$h_{2\bullet} = 0{,}375$
x_3^* (dreimal)	$\tilde{h}_{31} = 0{,}375 \cdot 0{,}4375 = 0{,}1641$ $(h_{31} = 0{,}125)$	$\tilde{h}_{32} = 0{,}375 \cdot 0{,}5625 = 0{,}2109$ $(h_{32} = 0{,}25)$	$h_{3\bullet} = 0{,}375$
$\sum_{j=1}^{3} h_{jk}$	$h_{\bullet 1} = 0{,}4375$	$h_{\bullet 2} = 0{,}5625$	1

Starke Abweichungen:

- Männer sind überproportional häufig zweifache Käufer.
- Frauen weisen dagegen eine größere einfache und dreifache Kaufhäufigkeit auf.

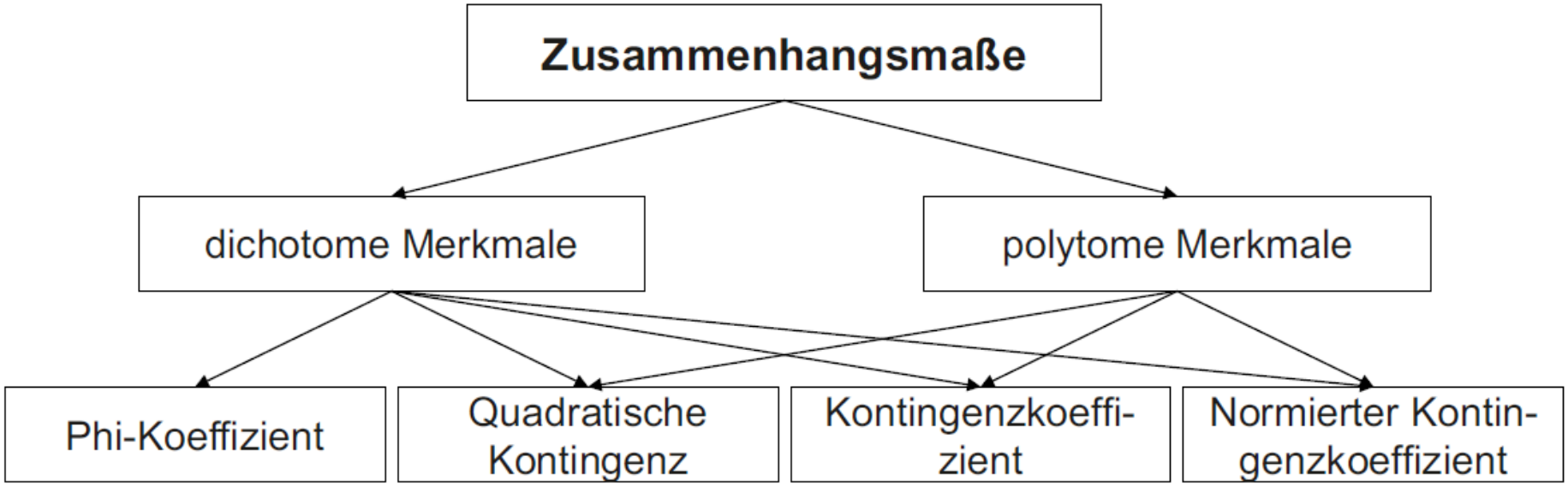
7.4 Phi-Koeffizient und Kontingenzkoeffizient

- **Fragestellung**: Wie lassen sich die in einer bivariaten Häufigkeitstabelle enthaltenen Informationen über den Zusammenhang zwischen den beiden Merkmalen X und Y (Abhängigkeit der beiden Merkmale X und Y voneinander) zu einer Kenngröße zusammenfassen?

- Speziell interessieren wir uns hier für ein Abhängigkeitsmaß zweier Merkmale X und Y einer Kontingenztabelle, in der **mindestens eines der Merkmale X oder Y nominalskaliert** ist.

- Der **Phi-Koeffizient**, die **quadratische Kontingenz** und der **Kontingenzkoeffizient** sind Maße für die Stärke der Abhängigkeit zwischen zwei Merkmalen einer Kontingenztabelle. Man bezeichnet sie als Assoziationskoeffizienten (oder Zusammenhangsmaße). Sie setzen nur ein nominales Skalenniveau voraus.

Übersicht verschiedener Zusammenhangsmaße



Dichotome Kodierung

- Kodierung der beiden Merkmalsausprägungen: 1 und 0
- Falls eine Merkmalsausprägung besonders im Blickfeld steht, wird sie mit 1 kodiert.
- Bei gleichwertigen Merkmalsausprägungen ist die Kodierung mit 0 oder 1 beliebig.

Vierfeldertafel

Darstellung der gemeinsamen Häufigkeitsverteilung dichotomer Merkmale mit ihren Randverteilungen:

X \ Y	1	0	Σ
1	a	b	a + b
0	c	d	c + d
Σ	a + c	b + d	n = a + b + c + d

Phi-Koeffizient:

$$\phi = \frac{a \cdot d - b \cdot c}{\sqrt{(a+b) \cdot (c+d) \cdot (a+c) \cdot (b+d)}}$$

- Der Phi-Koeffizient gibt die Stärke des Zusammenhangs zwischen zwei dichotomen Merkmalen an.
- ϕ-Werte nahe bei ± 1 belegen eine starken Zusammenhang, ϕ-Werte nahe bei 0 weisen aus, dass zwischen den beiden Merkmalen ein schwacher Zusammenhang besteht.
- Der Koeffizient ist auf das Intervall [-1; 1] normiert: $-1 \leq \phi \leq 1$
- Positive ϕ-Werte: Mit der gleichen Zahl kodierte Ausprägungen treten relativ häufig auf.
- Negative ϕ-Werte: Mit unterschiedlichen Zahlen kodierte Ausprägungen treten relativ häufig auf.

Beispiel: Rentner:innen und Produktinteresse

Ein Unternehmen hat bei 100 Befragten die Merkmale „Interesse am Produkt“ (X) und „Verrentung“ (Y) erhoben. Das Produktinteresse steht im Blickfeld, weshalb eine Zustimmung (ja) beim Merkmal X mit 1 kodiert wird. Außerdem ist die Zielgruppe der Rentner:innen gefragt, so dass die Bejahung einer Verrentung beim Merkmal Y ebenfalls mit 1 kodiert wird.

Die Erhebungsergebnisse sind in einer Vierfeldertafel wiedergegeben:

Y (Verrentung) / X (Interesse am Produkt)	1 (ja)	0 (nein)	Σ
1 (ja)	a = 40	b = 10	50
0 (nein)	c = 20	d = 30	50
Σ	60	40	n = 100

$$\phi = \frac{40 \cdot 30 - 10 \cdot 20}{\sqrt{(40+10) \cdot (20+30) \cdot (40+20) \cdot (10+30)}} = \frac{1.000}{\sqrt{6.000.000}} = 0{,}408$$

→ mittelstarker Zusammenhang

Quadratische Kontingenz und Kontingenzkoeffizient

- Der Phi-Koeffizient ist ein Assoziationskoeffizient, der allein bei 2x2-Kontingenztabellen anwendbar ist.
- Will man die Stärke des Zusammenhang bei beliebig großen $r \times c$-Kontingenztabellen (mit mindestens einem nominalskalierten Merkmal) messen, bietet sich die Berechnung der **quadratischen Kontingenz** und des **Kontingenzkoeffizienten** an.
- Die Maße basieren auf einem Vergleich der beobachteten gemeinsamen absoluten Häufigkeiten n_{jk} mit den bei Unabhängigkeit zu erwartenden gemeinsamen absoluten Häufigkeiten $\tilde{n}_{jk}$.

Quadratische Kontingenz

$$\chi^2 = \sum_{j=1}^{r} \sum_{k=1}^{c} \frac{\left(n_{jk} - \tilde{n}_{jk}\right)^2}{\tilde{n}_{jk}} \quad (\chi^2\text{: Chi−Quadrat})$$

- Die Abweichungen von der Unabhängigkeit gehen nicht absolut, sondern bezogen auf die bei Unabhängigkeit zu erwartenden Häufigkeiten $\tilde{n}_{jk}$ in den Kontingenzkoeffizienten ein.
- Da die Richtung des Zusammenhangs bei nominalskalierten Merkmalen allgemein nicht definiert ist, eliminiert man hierbei das Vorzeichen einer Abweichung durch Quadrierung.
- Wertebereich: $0 \leq \chi^2 \leq \infty$
- Statistische Unabhängigkeit: $\chi^2 = 0$
- Mit zunehmender Abweichung von der Unabhängigkeit wächst der χ^2−Wert.
- Mittlere quadratische Kontingenz: χ^2/n
- Bei einer Vierfeldertafel gilt: $\chi^2/n = \phi^2$

Unnormierter und normierter Kontingenzkoeffizient

- Da χ^2 mit steigender Anzahl von Beobachtungen zunimmt, ist sie in „roher“ Form noch nicht als Assoziationsmaß geeignet.
- Der Anzahleffekt kann in unterschiedlicher Form beseitigt werden.

Unnormierter Kontingenzkoeffizient $\boldsymbol{K}$:

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

- Wertebereich: $0 \leq K \leq \sqrt{(s-1)/s}$ mit $s = min\{r, c\}$

Normierter Kontingenzkoeffizient $\boldsymbol{K^*}$:

$$K^* = \frac{K}{\sqrt{(s-1)/s}}$$

- Wertebereich: $0 \leq K^* \leq 1$
- Je stärker sich der normierte Kontingenzkoeffizient K^* dem Wert eins annähert, desto stärker ist der Zusammenhang zwischen den Merkmalen X und Y. Bei einer statistischen Unabhängigkeit nimmt K^* den Wert null an.

Beispiel: Quadratische Kontingenz und Kontingenzkoeffizient (I)

Für die beiden Merkmale „Geschlecht“ und „Kaufhäufigkeit“ wurden die absoluten (beobachteten) Häufigkeiten n_{jk} und die absoluten bei Unabhängigkeit zu erwartenden Häufigkeiten $\tilde{n}_{jk}$ bereits berechnet.

X \ Y	y_1^* (männlich)	y_2^* (weiblich)	$\sum_{k=1}^{2} n_{jk}$
x_1^* (einmal)	$\tilde{n}_{11} = \frac{4 \cdot 7}{16} = 1{,}75$ $(n_{11} = 1)$	$\tilde{n}_{12} = \frac{4 \cdot 9}{16} = 2{,}25$ $(n_{12} = 3)$	$n_{1\bullet} = 4$
x_2^* (zweimal)	$\tilde{n}_{21} = \frac{6 \cdot 7}{16} = 2{,}625$ $(n_{21} = 4)$	$\tilde{n}_{22} = \frac{6 \cdot 9}{16} = 3{,}375$ $(n_{22} = 2)$	$n_{2\bullet} = 6$
x_3^* (dreimal)	$\tilde{n}_{31} = \frac{6 \cdot 7}{16} = 2{,}625$ $(n_{31} = 2)$	$\tilde{n}_{32} = \frac{6 \cdot 9}{16} = 3{,}375$ $(n_{32} = 4)$	$n_{3\bullet} = 6$
$\sum_{j=1}^{3} n_{jk}$	$n_{\bullet 1} = 7$	$n_{\bullet 2} = 9$	$n = 16$

Beispiel: Quadratische Kontingenz und Kontingenzkoeffizient (II)

- Quadratische Kontingenz:

$$\chi^2 = \sum_{j=1}^{r}\sum_{k=1}^{c}\frac{\left(n_{jk}-\tilde{n}_{jk}\right)^2}{\tilde{n}_{jk}}$$
$$= \frac{(1-1{,}75)^2}{1{,}75} + \frac{(3-2{,}25)^2}{2{,}25} + \frac{(4-2{,}625)^2}{2{,}625} + \frac{(2-3{,}375)^2}{3{,}375} + \frac{(2-2{,}625)^2}{2{,}625}$$
$$+ \frac{(4-3{,}375)^2}{3{,}375} = 2{,}116$$

- Unnormierter Kontingenzkoeffizient:

$$\mathrm{K} = \sqrt{\frac{\chi^2}{\chi^2+n}} = \sqrt{\frac{2{,}116}{2{,}116+16}} = 0{,}342$$

- Normierter Kontingenzkoeffizient: $s = min\{r, c\} = min\{3{,}2\} = 2$

$$K^* = \frac{K}{\sqrt{(s-1)/s}} = \frac{0{,}342}{\sqrt{(2-1)/2}} = 0{,}484$$

→ mittelstarker Zusammenhang