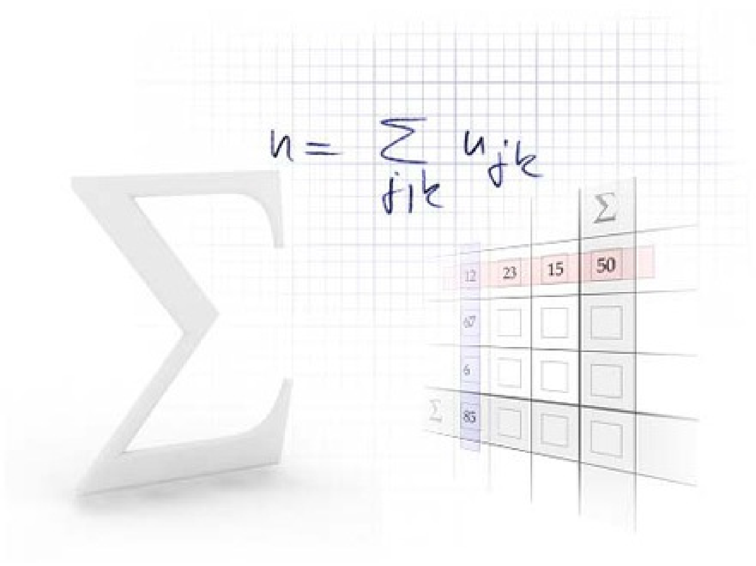
KTA - Kontingenztafel



Lernziele und Überblick

In dieser Lerneinheit werden Sie die Darstellung zweidimensionaler Merkmale als Häufigkeitstabellen kennen lernen.

Lernziele

Nach dem Durcharbeiten dieser Lerneinheit sollen Sie

- Kontingenztafeln erstellen und beschreiben können,
- Häufigkeiten in Kontingenztafeln bestimmen und interpretieren können,
- relative Häufigkeiten berechnen können,
- bedingte Häufigkeiten bestimmen und interpretieren können.

In dieser Lerneinheit werden die Begriffe

- mehrdimensionales Merkmal,
- Kontingenztafel,
- Vierfeldertafel,
- bedingte Häufigkeiten

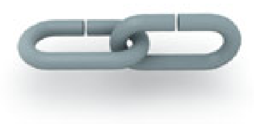
eingeführt.

Gliederung der Lerneinheit

Zeitbedarf und Umfang

Für die Durcharbeitung dieser Lerneinheit benötigen Sie ca. 120 Minuten und für die Übungen mit der Statistiksoftware **R** ca. 90 Minuten.

1 Einleitung



Für die statistische Beschreibung von nominalskalierten Merkmalen müssen wir einfach zählen, wie oft die verschiedenen möglichen Merkmalswerte auftreten. Wir haben dies schon in der Lerneinheit DHV, Abschnitt 2 gesehen, Diese Technik können wir auch dann anwenden, wenn wir es mit der simultanen Betrachtung von mehr als einem Merkmal zu tun haben.

Die ermittelten Häufigkeiten werden am sinnvollsten in Tabellenform so dargestellt, dass die verschiedenen Merkmalskombinationen direkt ablesbar sind. Dieses im Grunde sehr einfache Vorgehen sollten Sie nach dem Durcharbeiten dieser Lerneinheit beherrschen.

Die tabellarische Darstellung von Zusammenhängen ist eine wichtige Technik und häufig der Anfang einer systematischen Untersuchung von Sachverhalten.

Beispiel

KFZ-Versicherung - Schadensklassen

So will zum Beispiel eine KFZ-Versicherung wissen, ob es einen Zusammenhang zwischen der Häufigkeit von Schadensmeldungen und dem versicherten Fahrzeugtyp gibt. Die Häufigkeiten der Schadensfälle werden in Abhängigkeit vom Fahrzeugtyp tabellarisch dargestellt. Die Analyse des Zusammenhangs wird dann verwendet, um die Prämien für verschiedene Fahrzeugklassen festzulegen.

Wir beschäftigen uns in dieser Lerneinheit mit Kontingenztafeln. Die Kontingenztafel ist das Gegenstück zur Häufigkeitstabelle bei multivariaten bzw. bivariaten Datensätzen. Die Kontingenztafel ist die einfachste Darstellungsform eines Zusammenhangs.

1.1 Häufigkeiten in Kontingenztafeln

Zunächst machen wir uns mit Häufigkeitstabellen auf sportliche Weise vertraut. Sie sollen am folgenden Beispiel erkennen, wie Sie mit den Häufigkeiten in Kontingenztafeln umgehen können.

Beispiel

272 ?

Sport in Berlin

Wir haben uns einen Überblick über die Mitgliedschaften von Sportvereinen verschafft. Dabei konnten wir auf eine zufällige Auswahl von 262 Personen aus Berliner Vereinen der Bereiche Fußball, Turnen, Schwimmen und Tennis zurückgreifen. Andere Geschlechter als männlich und weiblich wurden nicht erfasst.

Die Aufschlüsselung nach Sportarten und Geschlecht.

	Geschlecht		
Sportart	**Männlich**	**Weiblich**	**Summe**
Fußball	105	15	**120**
Turnen	25	75	**100**
Schwimmen	6	6	**12**
Tennis	24	16	**40**
Summe	**160**	**112**	**272**

Tab.: Sportarten und Geschlecht eines Vereins, **absolute** Häfigkeit

Was können wir dieser Aufstellung entnehmen?

Am häufigsten sind die Fußballer vertreten. Und dabei dominieren, wen wundert das, die Männer. Beim Turnen verhält es sich fast umgekehrt: es gibt dreimal so viele Turnerinnen wie Turner.

Damit wir einen noch besseren Überblick erhalten, gehen wir zu relativen Häufigkeiten über. Am einfachsten geht das beim Turnen: 25 % der befragten Mitglieder von Turnvereinen sind männlich, 75 % sind weiblich.

So lässt sich die Geschlechterverteilung je Sportart tabellarisch wie folgt darstellen:

	Geschlecht		
Sportart	**Männlich**	**Weiblich**	**Summe**
Fußball	87.5	12.5	**100**
Turnen	25	75	**100**
Schwimmen	50	50	**100**
Tennis	60	40	**100**
Summe	**59**	**41**	**100**

Tab.: Sportarten und Geschlecht eines Vereins, **relative** Häufigkeit

Man nennt diese Art der Darstellung bedingte Verteilung für die Variable Geschlecht, gegeben die Variable Sportart.

Wir haben zwar jetzt einen besseren Überblick über das Geschlechterverhältnis bei den verschiedenen Sparten, können aber keine Aussagen mehr über die Beliebtheit der unterschiedlichen Sportarten treffen. Wenn wir die Beliebtheit untersuchen würden, hätten wir die Summen der männlichen und der weiblichen Befragten jeweils auf 100 % setzen müssen. Sie sollten das selbst ausprobieren.

Auf jeden Fall wird aus dieser Darstellung deutlich, dass die ausgewählte Sportart mit dem Geschlecht zusammenhängt: Männer bevorzugen Fußball, Frauen das Turnen, beim Schwimmen und beim Tennis ist das Verhältnis eher ausgeglichen.

1.2 Erzeugen einer bivariaten Verteilung

Wie kommen wir zu Kontingenztafeln, wenn wir mit Einzelbeobachtungen zweier Merkmale beginnen? Das Vorgehen wird Ihnen am folgenden Beispiel sicher deutlich.

Beispiel

Haar- und Augenfarbe von weiblichen Models aus Modemagazinen

Was hat es mit blauäugigen Blondinen auf sich?

Wir finden in einigen aktuellen Modemagazinen folgende Zahlen für insgesamt 17 weibliche Models:

- 10 sind blond,
- 10 haben blaue Augen,
- 8 sind blond und blauäugig.

Damit können wir schon unsere Tabelle konstruieren.

Berechnen

Übung KTA-01

Haar- und Augenfarbe von weiblichen Models

Sie können die fehlenden Felder leicht ergänzen, indem Sie auf die Einhaltung der angegebenen Summenwerte achten.

Bitte füllen Sie die Kontingenztafel aus.

	Blauäugig	Nicht blauäugig	Summe
Blond	8	2	**10**
Nicht blond	2	5	7
Summe	**10**	7	**17**

Lösung (Siehe Anhang)

Bearbeitungszeit: 2 Minuten

Die zweidimensionale Darstellung der Daten entspricht am besten der Datenstruktur, eine Liste ist hier weniger anschaulich (Haar- und Augenfarbe, Liste der Häufigkeiten).

Haarfarbe	Augenfarbe	Ausprägungen
blond	blauäugig	8
blond	nicht blauäugig	2
nicht blond	blauäugig	2
nicht blond	nicht blauäugig	5
Summe		**17**

Tab.: Beispiel der Häufigkeit von Haarfarbe und Augenfarbe

Diese Liste enthält dieselbe Information wie die Tabelle, allerdings weniger übersichtlich.

2 Grundbegriffe

Wir haben in unseren Beispielen gesehen, dass die Häufigkeiten der Merkmalskombinationen zweier Merkmale am besten als Matrix dargestellt werden. Die einzelnen Elemente einer Matrix werden durch die Angabe der zugehörigen Zeile und Spalte identifiziert. Dazu werden Indizes verwendet.

Definition

Kontingenztafel

Die Kontingenztafel ist die Zusammenstellung der Häufigkeiten der Ausprägungskombinationen zweier Merkmale in Form einer Matrix.

Wir betrachten n Untersuchungseinheiten, an denen die Merkmale X und Y beobachtet werden. Die verschiedenen Merkmalsausprägungen von X und von Y werden folgendermaßen bezeichnet:

Ausprägungen von X : $(a_1, a_2, \ldots, a_j, \ldots, a_r)$
Anzahl Zeilen: r von row (englisch Zeile)

Ausprägungen von Y : $(b_1, b_2, \ldots, b_k, \ldots, b_c)$
Anzahl Spalten: c von column (englisch Spalte).

Das zweidimensionale Merkmal (X, Y), hat dann die Ausprägungen (a_j, b_k).
Beobachten wir nun simultan Werte für X und für Y können wir die Häufigkeiten der beobachteten Merkmalskombinationen (a_j, b_k) bestimmen.

Zur Darstellung verwenden wir eine der folgenden Tabelle entsprechende Form:

	Spalten-Nr.	1	2	..	k	..	c	Σ
Zeilen-Nr.	$X \backslash Y$	b_1	b_2	..	b_k	..	b_c	
1	a_1	n_{11}	n_{12}	..	n_{1k}	..	n_{1c}	$n_{1\bullet}$
2	a_2	n_{21}	n_{22}	..	n_{2k}	..	n_{2c}	$n_{2\bullet}$
:	:	:	:	:	:	:	:	:
j	a_j	n_{j1}	n_{j2}	..	n_{jk}	..	n_{jc}	$n_{j\bullet}$
:	:	:	:	:	:	:	:	:
r	a_r	n_{r1}	n_{r2}	..	n_{rk}	..	n_{rc}	$n_{r\bullet}$
Σ		$n_{\bullet 1}$	$n_{\bullet 2}$	..	$n_{\bullet k}$	..	$n_{\bullet c}$	n

Tab.: Kontingenztafel der absoluten Häufigkeiten n_{jk}

Die Häufigkeit der Merkmalskombination(j, k) also $n(X = a_j, Y = b_k)$ wird mit $n_j k$ bezeichnet. Diese Häufigkeit wird in Zeile j und Spalte k unserer Kontingenztafel eingetragen. Für die relativen Häufigkeiten verwenden wir wie in Lerneinheit DHV den Buchstaben h.

Für die Summenhäufigkeiten hat sich folgende Schreibweise bewährt: den Index über den summiert wird, ersetzt man durch einen Punkt.
Damit erhalten wir

$$n_{j\bullet} = n(X = a_j) = \sum_{k=1}^{c} n_{jk}$$

sowie

$$n_{\bullet k} = n(Y = b_k) = \sum_{j=1}^{r} n_{jk}$$

Die univariaten Verteilungen der Merkmale X und Y stehen am Rand der Tabelle, sie werden deshalb auch als Randverteilungen bezeichnet.

2.1 Die 2 × 2-Felder-Tafel (2 × 2- Kontingenztafel)

Damit wir überhaupt eine zweidimensionale Verteilung sehen können, müssen mindestens zwei Zeilen und zwei Spalten vorhanden sein. (Klar?!)

Wir können auch für kardinale Daten eine Darstellung als Kontingenztafel erhalten, wenn wir die ursprünglichen Beobachtungen in Klassen einteilen.

Beispiel

Studenten, Gewicht und Größe

Wir betrachten noch einmal den Datensatz zu Körpergröße und Gewicht von Studierenden aus Lerneinheit *„ZHA - Zusammenhänge“*, Abschnitt 4.3.

Das Gewicht und die Größe der Studierenden teilen wir in mehrere Klassen ein, um die Ergebnisse in einer Kontingenztafel (r = 2: Zeilenanzahl, c = 2: Spaltenanzahl, n = 26) darstellen zu können. Für jedes Merkmal (X = Größe und Y = Gewicht) haben wir eine recht grobe Einteilung gemacht: nur zwei Klassen pro Merkmal. Die Häufigkeiten der 26 Studierenden sind in der folgenden Vierfeldertafel (2 × 2 -Kontingenztafel) dargestellt.

Größe (X)	Gewicht (Y)		Summe
	Y≤ 70	Y > 70	
X≤ 170	$n_{11} = 13$	$n_{12} = 4$	$n_{1\bullet} = 17$
X > 170	$n_{21} = 2$	$n_{22} = 7$	$n_{2\bullet} = 9$
Summe	$n_{\bullet 1} = 15$	$n_{\bullet 2} = 11$	$n = 26$

Tab.: Kontingenztafel für das Gewicht und die Größe von Studenten

Aus der Tabelle kann man erkennen, dass zwischen Gewicht und Größe ein Zusammenhang besteht: Studierende mit einer Körpergröße über 170 cm scheinen zumeist schwerer als 70 kg zu sein (Verhältnis leicht : schwer wie 2 : 7), während kleinere Studierende in der Regel wohl leichter sind (Verhältnis 13 : 4).

2.2 Bedingte relative Häufigkeiten

Wir betrachten noch einmal den Zusammenhang zwischen Haar- und Augenfarbe, um den Unterschied zwischen der unbedingten und der bedingten Häufigkeitsverteilung zu ermitteln.

Wir verfügen jedoch nun über etwas detailliertere Angaben der 26 Studierenden.

Beispiel

Haar- und Augenfarbe von Studierenden

Haarfarbe	Augenfarbe			Summe
	blau	grün/grau	braun	
blond	8	2	0	**10**
braun	2	4	1	**7**
rot	1	0	1	**2**
schwarz	0	1	6	**7**
Summe	**11**	**7**	**8**	**26**

Tab.: Häufigkeit und Summe von Haarfarbe und Augenfarbe

Auch wenn die Gesamtzahl der erfassten Studierenden relativ klein ist, wollen wir zur Veranschaulichung zu relativen Häufigkeiten, also zu Angaben in Prozent übergehen.

Betrachten wir zunächst nur die blonden blauäugigen Studierenden:
Unbedingte Häufigkeit: 8/26 = 30,8 %.
Bedingte Häufigkeit, für blonde Haare | Bedingung blaue Augen: 8/11= 72,7 %
Bedingte Häufigkeit, für blaue Augen | Bedingung blonde Haare: 8/10 = 80 %.

Wir sehen damit deutlich einen Zusammenhang zwischen Augen- und Haarfarbe, während nur etwa 30 % der erfassten Studierenden blond und blauäugig sind, vertreten die Blonden mehr als 70 % aller Blauäugigen, bzw. vertreten die Blauäugigen 80 % der Blonden.

In den folgenden Tabellen können Sie alle Möglichkeiten der Berechnung relativer Häufigkeiten verfolgen.

Versuchen Sie die Ergebnisse bitte selbst zu interpretieren.

Haarfarbe	Augenfarbe			Summe
	blau	grün/grau	braun	
blond	30,8	7,7	0	**38,5**
braun	7,7	15,4	3,8	**26,9**
rot	3,8	0	3,8	**7,7**
schwarz	0	3,8	23,1	**26,9**
Summe	**42,3**	**26,9**	**30,8**	**100**

Tab.: Relative Häufigkeiten, Haar- und Augenfarbe

Haarfarbe	Augenfarbe			Summe
	blau	grün/grau	braun	
blond	80	20	0	**100**
braun	29	57	14	**100**
rot	50	0	50	**100**
schwarz	0	14	86	**100**

Tab.: Bedingte Häufigkeiten, Bedingung Haarfarbe

Die Unterschiede in den bedingten Häufigkeiten zeigen den Zusammenhang zwischen Haar- und Augenfarbe.

Haarfarbe	Augenfarbe		
	blau	grün/grau	braun
blond	73	29	0
braun	18	57	12
rot	9	0	12
schwarz	0	14	75
Summe	**100**	**100**	**100**

Tab.: Bedingte Häufigkeiten, Bedingung Augenfarbe

3 Berechnungsformeln für Kontingenztafeln

Wir haben Sie bisher relativ schonend behandelt, was den Umgang mit mathematischen Formeln angeht. Das war auch Absicht, denn es ist viel wichtiger, dass Sie den Inhalt und die Bedeutung der Begriffe erfassen, als formale Schreibweisen zu lernen. Allerdings ist die mathematische Notation ungeheuer effektiv und oft die einzige Möglichkeit, Sachverhalte wirklich exakt zu formulieren.

Sie haben jetzt die Gelegenheit, dieses Vorgehen anhand einiger Formeln zur Berechnung relativer und bedingter Häufigkeiten vertieft kennenzulernen.

Hat man einmal den Formelapparat entwickelt, kann man leicht die Berechnungen in Computerprogramme übertragen.

In den folgenden Abschnitten führen wir Schritt für Schritt mathematische Notation, Definitionen und Berechnungsformeln ein.

3.1 Klassen, Häufigkeiten und Verteilungen

Die Kontingenztafeln werden in der Regel (in der deskriptiven Statistik) nur für nominale oder ordinale Merkmale erstellt und analysiert.

Definition

Zweidimensionales Merkmal

Werden zwei Merkmale X und Y simultan erhoben, dann heißt das Paar (X, Y) zweidimensionales Merkmal.

Anmerkungen

Besitzt X die möglichen Merkmalsausprägungen $a_1, a_2, \ldots, a_j, \ldots, a_r$ und Y die Ausprägungen $b_1, b_2, \ldots, b_k, \ldots, b_c$, dann gibt es für das zweidimensionale Merkmal $(X, Y) r \cdot c$ verschiedene Merkmalswerte der Form (a_j, b_k).

Alle Werte einer Datenreihe $(x_1, y_1), (x_2, y_2)$, für die$(x_i, y_i) = a_j, b_k)$ gilt, gehören zur Klasse (j, k).

Definition

Häufigkeit der Klasse (j, k)

Die Anzahl der Elemente in der Klasse (j, k) heißt Häufigkeit der Klasse (j, k).

Mit $n_j k = n(a_j, b_k)$ bzw.$h_j k = h(a_j, b_k)$ werden die absoluten bzw. die relativen Häufigkeiten bezeichnet.

Anmerkungen

Wenn in einer Kontingenztafel die Anzahl von Ausprägungen $r = c$ gleich ist, ist diese Kontingenztafel quadratisch, ansonsten rechteckig. Ein Spezialfall ist eine Kontingenztafel für dichotome bzw. dichotomisierte Merkmale, die wegen $r = c = 2$ und $r \cdot c = 4$ auch 2 × 2-Felder-Tafel genannt wird.

Definition

Bivariate Häufigkeitsverteilung

Die Menge aller $r \cdot c$ Ausprägungspaare und die zugehörigen absoluten Häufigkeiten $n_j k = n(a_j, b_k)$ bzw. relativen Häufigkeiten $h_j k = h(a_j, b_k)$ heißt zweidimensionale oder bivariate Häufigkeitsverteilung.

Da durch die verschiedenen Klassen alle beobachteten Werte erfasst werden, gilt offensichtlich:

$$\sum_{j=1}^{r} \sum_{k=1}^{c} n_{jk} = n \text{ und } \sum_{j=1}^{r} \sum_{k=1}^{c} h_{jk} = 1$$

Die Summe der Häufigkeiten einer Zeile oder einer Spalte geben uns Informationen über den zugehörigen einzelnen Merkmalswert (wir erinnern uns, 10 blonde Studierende).
Wohin damit?

Am einfachsten notiert man diese Zahlen am Rand der Tabelle. Der rechte oder untere Rand wird hier verwendet. Schon haben wir die Randhäufigkeiten.

Im zweidimensionalen Schema der n_{jk} sind partielle Summierungen möglich:
Die Zeilensummen sind durch die absolute Häufigkeit der Ordnung j am Rand der Kontingenztafel gegeben und heißen absolute Randhäufigkeit der Merkmalsausprägung a_j:

$$n_{j\bullet} = \sum_{k=1}^{c} n_{jk} \text{ , } (j = 1, 2, \ldots, r).$$

Die Spaltensummen sind durch die absolute Häufigkeit der Ordnung am Rand der Kontingenztafel gegeben und heißen absolute Randhäufigkeit der Merkmalsausprägung b_k:

$$n_{\bullet k} = \sum_{j=1}^{r} n_{jk}\ ,\ (k = 1, 2, \dots, c)$$

Anmerkung

Punktsymbol ●

Das Punktsymbol ● im Index dient einer vereinfachten Schreibweise und kennzeichnet jeweils die Summe über alle j bzw. k.

3.2 Die Vierfeldertafel in allgemeiner Schreibweise

Zum Einüben der eingeführten Begriffe haben Sie jetzt die Gelegenheit die Bezeichnungen im Detail an einer $2 \cdot 2$ - Feldertafel nachzuvollziehen.

Berechnen

Übung KTA-02

Vierfeldertafel in Symbolen

Die Tabelle zeigt eine Vierfeldertafel mit absoluten Häufigkeiten.

X \ Y	b_1	b_2	Summe
a_1	n_{11}	n_{12}	$n_{1\bullet} = n_{11} + n_{12}$
a_2	n_{21}	n_{22}	$n_{2\bullet} = n_{21} + n_{22}$
Summe	$n_{\bullet 1} = n_{11} + n_{21}$	$n_{\bullet 2} = n_{12} + n_{22}$	n

Wenn Sie die folgenden Fragen beantworten können, haben Sie die eingeführten Begriffe parat.

1. Welche Werte kann das zweidimensionale Merkmal (X,Y) annehmen?
2. Wieviele Klassen gibt es?
3. Was ist die Häufigkeit der Klasse 1, 2?
4. Was sind die Randverteilungen?
5. Welcher Rand (rechts oder unten) enthält die Häufigkeitsverteilung von X?

Lösung (Siehe Anhang)

Bearbeitungszeit: 5 Minuten

Wir zeigen Ihnen der Vollständigkeit halber auch die zugehörigen relativen Häufigkeiten in der Vierfeldertafel.

X \ Y	b_1	b_2	Summe
a_1	h_{11}	h_{12}	$h_{1\bullet} = h_{11} + h_{12}$
a_2	h_{21}	h_{22}	$h_{2\bullet} = h_{21} + h_{22}$
Summe	$h_{\bullet 1} = h_{11} + h_{21}$	$h_{\bullet 2} = h_{12} + h_{22}$	1

Tab.: Vierfeldertafel der relativen Häufigkeiten

Hier gibt es eigentlich keine neuen Fragen, es werden die absoluten Häufigkeiten einfach durch die relativen Häufigkeiten ersetzt.

3.3 Randverteilung

Schauen wir bei einer zweidimensionalen Häufigkeitstabelle nur auf die Ränder, erhalten wir die Randverteilungen. Diese entsprechen natürlich den bekannten univariaten Häufigkeitsverteilungen der einzelnen Variablen X und Y.
Die Randverteilungen von X und Y sind in den folgenden Tabellen dargestellt:

Tab.: Randverteilung von X

Merkmal X	absolute Häufigkeit	relative Häufigkeit
a_1	n_1	h_1
:	:	:
a_j	n_j	h_j
:	:	:
a_r	n_r	h_r
Summe	**n**	**1**

Tab.: Randverteilung von Y

Merkmal Y	absolute Häufigkeit	relative Häufigkeit
b_1	n_1	h_1
:	:	:
b_k	n_k	h_k
:	:	:
b_c	n_c	h_c
Summe	**n**	**1**

Definition

Randhäufigkeit

Für die Randhäufigkeiten der Merkmale X und Y gilt stets:

$$n = \sum_{j=1}^{r} n_{j\bullet} = \sum_{k=1}^{c} n_{\bullet k} = \sum_{j=1}^{r}\sum_{k=1}^{c} n_{jk}$$

3.4 Bedingte relative Häufigkeit

Definition

Bedingte relative Häufigkeit

Die relative Häufigkeit, mit der das Merkmal X (bzw. Y) die Ausprägung a_j (bzw. b_k) unter der Bedingung annimmt, dass das Merkmal Y (bzw. X) die Ausprägung b_k (bzw. a_j) besitzt, heißt bedingte (oder konditionale, lat.: conditio - Bedingung) relative Häufigkeit.

Die bedingte relative Häufigkeit von Y

Wählt man $X = a_j$ als festen Wert, so ergibt sich die bedingte relative Häufigkeit von Y unter der Bedingung $X = a_j$ (kurz $Y \mid X = a_j$) durch

$$h(Y = b_k|X = a_j) = \frac{n(X = a_j\,,\,Y = b_k)}{n(X = a_j)} = \frac{n_{jk}}{n_{j\bullet}} = h_{k|j}$$

(Dabei ist $n\,(X = a_j) > 0$ vorausgesetzt).
Also, die bedingte relative Häufigkeit $h\,(Y = b_k \mid X = a_j)$ gibt die relative Häufigkeit der Beobachtung b_k in dem durch Vorkommen des Wertes a_j eingeschränkten Datensatz an.

Auf Grund der Beziehung: $\frac{n_{jk}}{n_{j\bullet}} = \frac{\frac{n_{jk}}{n}}{\frac{n_{j\bullet}}{n}} = \frac{h_{jk}}{h_{j\bullet}}$

erhalten wir: $h(Y = b_k|X = a_j) = \frac{h_{jk}}{h_{j\bullet}}$

Die bedingte relative Häufigkeit von X

Die bedingte relative Häufigkeit von X wird analog bestimmt. Es gilt:

$$h(X = a_j|Y = b_k) = \frac{n(X = a_j\,,\,Y = b_k)}{n(Y = b_k)} = \frac{n_{jk}}{n_{\bullet k}} = h_{j|k}$$

sowie:

$$h(X = a_j|Y = b_k) = \frac{h_{jk}}{h_{\bullet k}}$$

Auch bedingte Häufigkeiten ergeben den Wert 1 bei Summation:

$$\sum_{k=1}^{c} h(Y = b_k|X = a_j) = 1 \; und \; \sum_{j=1}^{r} h(X = a_j|Y = b_k) = 1$$

Zum Abschluss dieses Formelkrams zeigen wir Ihnen noch einmal ausführlich am Beispiel der Kontingenztafel der Haar- und Augenfarbe, wie bedingte Häufigkeiten mit Hilfe der angegebenen Formeln bestimmt werden.
Eigentlich können Sie das schon.

Beispiel

Bedingte Häufigkeiten

Von der Häufigkeitsverteilung des Beispiels „Studenten Haar- und Augenfarbe“ interessiert uns der Anteil der blonden StudentInnen (Bedingung $X = a_1$) , die blaue, grau/grüne oder braune Augen haben ($Y = b_1$, $Y = b_2$ oder $Y = b_3$).

Die folgende Tabelle zeigt den relevanten Ausschnitt der Kontingenztafel.

Zeilen-Nr.	Spalten-Nr.	1	2	3	Summe
	X \ Y	b_1	b_2	b_3	
1	a_1	8	2	0	**10**

Tab.: Relevanter Ausschnitt der Kontingenztafel

Wir interessieren uns speziell für blonde Studenten, so werden wir uns auf deren Anzahl $n_1 = n(a_1) = 10$ beziehen und erhalten damit die bedingten relativen Häufigkeiten unter der Bedingung a_1 .
Also, die bedingte relative Häufigkeitsverteilung von Y unter der Bedingung $X = a_1$ sieht so aus:

$$h(Y = 8|X = a_1) = \frac{n(X = a_1\ , Y = b_1)}{n(X = a_1)} = \frac{n_{11}}{n_{1\bullet}} = \frac{8}{10} = 0{,}8$$

$$h(Y = 2|X = a_1) = \frac{n(X = a_1\ , Y = b_2)}{n(X = a_1)} = \frac{n_{12}}{n_{1\bullet}} = \frac{2}{10} = 0{,}2$$

$$h(Y = 0|X = a_1) = \frac{n(X = a_1\ , Y = b_3)}{n(X = a_1)} = \frac{n_{13}}{n_{1\bullet}} = \frac{0}{10} = 0$$

80 % der blonden Studenten haben blaue Augen.
20 % der blonden Studenten haben grau-grüne Augen.
Keiner der blonden Studenten hat braune Augen.

4 Übungen zur Kontingenztafel

Berechnen

Übung KTA-03

Trocken oder halbtrocken?

Die Inhaberin des Weinfachgeschäftes Maestro hat letzte Woche mit einem Weintransport 500 Flaschen verschiedener Proseccos bekommen. Von diesen 500 Flaschen sind 200 von der Sorte „Trocken“, 20 davon stammen aus Umbrien. 300 der 500 Flaschen stammen aus Abruzzen, davon sind 200 „Halbtrocken“.
150 Flaschen Prosecco stammen aus Latium.

Aufgaben

1. Benennen Sie den Merkmalsträger, die Gesamtheit und deren Umfang, die Identifikationsmerkmale und die Erhebungsmerkmale. Wie sind die Erhebungsmerkmale skaliert?
2. Erstellen Sie die Kontingenztafel.
3. Bestimmen Sie (in Form einer Tabelle) die absolute und relative Randverteilung von der Sorte des Prosecco (X). Erklären Sie anhand der erstellten Tabelle den Begriff Randverteilung des Merkmals X.
4. Bestimmen Sie (in Form einer Tabelle) die absolute und relative Randverteilung von dem Herkunftsort (Y). Erklären Sie anhand der erstellten Tabelle den Begriff Randverteilung des Merkmals Y.
5. Bestimmen Sie die bedingten relativen Häufigkeitsverteilungen mit Hilfe einer Tabelle. Erklären Sie anhand der erstellten Tabelle den Begriff Häufigkeitsverteilungen.

Lösung (Siehe Anhang)

Bearbeitungszeit: 25 Minuten

Berechnen

Übung KTA-04

Untergang der Titanic

Beim Untergang der Titanic haben von 885 Besatzungsmitgliedern 212 überlebt. Unter den Überlebenden waren 20 der 23 weiblichen Besatzungsmitglieder.

Quelle:
British Board of Trade (1990), Report on the Loss of the Titanic (S.S.).
British Board of Trade Inquiry Report , Gloucester, UK: Allan Sutton Publishing.

Seit dem berühmten Film gibt es ausführliche Informationen auch im Internet, z. B. bei Wikipedia wikipedia.org/wiki/RMS_Titanic

Aufgaben

1. Bestimmen Sie die zugehörige Vierfeldertafel. Berechnen Sie die bedingte Verteilung für die Variablen „Überleben“, gegeben „Geschlecht“.
2. Interpretieren Sie das Ergebnis.

Lösung (Siehe Anhang)

Bearbeitungszeit: 10 Minuten

Zusammenfassung

- Simultan erhobene Merkmale können als mehrdimensionale Merkmale aufgefasst werden.
- Die Häufigkeitsverteilung zweidimensionaler nominaler Merkmale wird in Form einer Kontingenztafel dargestellt.
- Die Randhäufigkeiten einer Kontingenztafel entsprechen den eindimensionalen Häufigkeiten der einzelnen Merkmale.
- Bedingte Verteilungen erhält man durch die Einschränkung der Beobachtungen auf die ausgewählten Werte des bedingenden Merkmals.
- Unterschiede in den bedingten Verteilungen weisen auf einen Zusammenhang zwischen den Merkmalen hin.
- Merkmale mit nur zwei Ausprägungen heißen dichotom.
- Die Kontingenztafel zweier dichotomer Merkmale heißt Vierfeldertafel.

Sie sind am Ende dieser Lerneinheit angelangt. Auf der folgenden Seite finden Sie noch die Übungen zur Wissensüberprüfung, weitere Übungen und wichtige Formeln.

Wissensüberprüfung

Multiple Choice

Übung KTA-05

Welche Aussagen sind falsch und welche richtig?

	Richtig	Falsch	Auswertung
Kontingenztabellen werden in der Regel nur für nominale oder ordinale Merkmale erstellt und analysiert.	○	○	
In einer Vierfeldtafel gilt: $n_{1\bullet} = n_{11} + n_{21}$	○	○	
Gegeben ist folgender Ausschnitt aus einer Kontingenztabelle: $\begin{array}{\|c\|c\|c\|c\|} & b_1 & b_2 & b_3 \\ \hline a_1 & 4 & 3 & 3 \end{array}$ Die bedingte relative Häufigkeit von Y=b_2 unter der Bedingung x = a_1 ist: $h = (Y = b_2 \| X = a_1) = \frac{n(X=a_1, Y=b_2)}{n(X=a_1)} = \frac{n_{12}}{n_{j\bullet}} = \frac{3}{10} = 0{,}3$.	○	○	
Für die relativen Randhäufigkeiten einer beliebigen Vierfeldtafel gilt: $h_{\bullet 1} + h_{\bullet 2} = h_{\bullet 1} + h_{\bullet 2} = 1$	○	○	

Übungen mit der Statistiksoftware R

Die in der Lerneinheit behandelten Themen können Sie anhand der folgenden Übungsaufgaben mit der Statistiksoftware **R** bearbeiten. Um die Übungen zu bearbeiten, muss die Software „**R**" auf Ihrem Rechner installiert sein.

Installationshinweise [Manuals | **R** Installation and Administration]

Für einige Übungen stehen auch Musterlösungen für das Programm Excel bereit.

Berechnen

Übung KTA-06a

Dichotom

Aus Datensätzen mit zwei Merkmalen lässt sich leicht eine 4-Feldertafel erstellen indem die Daten dichotomisiert werden. D. h. die beiden Merkmale werden in je zwei Klassen unterteilt und die dazugehörigen Häufigkeiten in eine 2 x 2-Kontingenztabelle eingetragen.

Die Datei **`dichotom.txt`** enthält die Daten zu Größe und Gewicht der Gruppe Studierender die hier in der Tabelle aufgeführt sind.

i	1	2	3	4	5	6	7	8	9	10	11	12	13
Größe	158	160	163	165	165	168	168	168	169	170	171	171	172
Gewicht	48	59	102	57	80	53	58	58	66	87	70	79	68

i	14	15	16	17	18	19	20	21	22	23	24	25	26
Größe	173	174	174	174	177	177	178	178	186	190	191	192	194
Gewicht	73	76	63	83	65	77	72	85	67	80	95	90	72

`dichotom.txt` (2 KB)

Aufgabe

1. Erstellen Sie für den bekannten Datensatz der Größe und des Gewichts von Studierenden eine 4-Feldertafel. Wählen sie dabei für die Größe die Klassen
   X_1 für Größe < Median (Größe) und
   X_2 für Größe $\geq$ Median (Größe).

   Die Klassierung des Gewichts erfolgt analog in Y_1 und Y_2.

Lösung mit R und Excel (Siehe Anhang)

Bearbeitungszeit: 20 Minuten

Berechnen

Übung KTA-06b

Sport

Wir wiederholen hier unser Beispiel aus dem Kapitel 1.1. Die Tabelle zeigt den Überblick über die Mitgliedschaften in Sportvereinen von 262 Personen. Betrachtet werden die Bereiche: Fußball, Turnen, Schwimmen und Tennis. Andere Geschlechter als männlich und weiblich wurden nicht erfasst.

	Geschlecht	
Sportart	**Männlich**	**Weiblich**
Fußball	105	15
Turnen	25	75
Schwimmen	6	6
Tennis	24	16

Aufgabe

1. Erstellen Sie eine Kontingenztabelle mit den dazugehörigen Randverteilungen und bestimmen Sie die relativen bedingten Häufigkeiten.

Lösung mit R und Excel (Siehe Anhang)

Bearbeitungszeit: 20 Minuten

Berechnen

Übung KTA-06c

Prosecco Lieferung

Auch unsere Übung „Trocken oder halbtrocken?“ wiederholen wir hier, um die Lösung mit R und mit Excel zu zeigen.

Die Inhaberin des Weinfachgeschäftes Maestro hat letzte Woche mit einem Weintransport 500 Flaschen verschiedener Proseccos bekommen. Von diesen 500 Flaschen sind 200 von der Sorte „Trocken“, 20 davon stammen aus Umbrien. 300 der 500 Flaschen stammen aus Abruzzen, davon sind 200 „Halbtrocken“. 150 Flaschen Prosecco stammen aus Latium.

Aufgaben

1. Lesen Sie die Daten ein, erstellen Sie die absoluten Häufigkeiten mit den Randverteilungen
2. Erstellen Sie eine Kontingenztabelle mit den relativen Häufigkeiten.
3. Berechnen Sie die bedingte relative Häufigkeit des Herkunftsortes und interpretieren Sie diese.

Lösung mit R und Excel (Siehe Anhang)

Bearbeitungszeit: 20 Minuten

Zusätzliche Übungsaufgaben

Zum Vertiefen der vorgestellten Inhalte finden Sie hier zusätzliche Übungsaufgaben zur freiwilligen Bearbeitung.

Berechnen

Übung KTA-07

Preise und Verpackung

Bitte füllen Sie die Kontingenztafel aus.

X/Y	normaler Preis	erhöhter Preis	Σ
einfache Verpackung	10		18
aufwendige Verpackung			
Σ	24		40

Lösung (Siehe Anhang)

Bearbeitungszeit: 3 Minuten

Berechnen

Übung KTA-08

Verkehrsunfälle

Von 1000 Verkehrsunfällen waren 280 mit tödlichen Ausgang. Davon ereigneten sich 80 bei einer Geschwindigkeit von mehr als 150 km/h. Insgesamt ereigneten sich 900 Verkehrsunfälle bei einer niedrigeren Geschwindigkeit.

Aufgaben

1. Benennen Sie den Merkmalsträger, die Gesamtheit und deren Umfang, und die Erhebungsmerkmale.
2. Erstellen Sie die Kontingenztafel.
3. Bestimmen Sie (in Form einer Tabelle) die absolute und relative Randverteilung des Merkmals X.
4. Bestimmen Sie (in Form einer Tabelle) die absolute und relative Randverteilung des Merkmals Y.
5. Bestimmen Sie die bedingten relativen Häufgkeitsverteilungen mit Hilfe einer Tabelle.

Lösung (Siehe Anhang)

Bearbeitungszeit: 15 Minuten

Berechnen

Übung KTA-09

Studentenstreik

200 zufällig ausgewählte Studierende der Berliner Hochschule für Technik wurden im November 2022 nach ihrem Studienfach (X) und nach ihrer Einstellung zum Bildungsstreik (Y) befragt.

Dabei ergaben sich folgende relative Häufigkeiten:

X/Y	positiv	negativ	neutral	
Naturwissenschaften	0,2		0,15	
Geisteswissenschaften		0,05	0,05	0,2
Wirtschaftswissenschaften		0,2		0,4
Σ	0,4			

Ergänzen Sie obige Tabelle und bestimmen Sie mit Hilfe einer Tabelle die drei bedingten
relativen Häufigkeitsverteilungen von Y unter der Bedingung X = Naturwissenschaften, X = Geisteswissenschaften und X = Wirtschaftswissenschaften.

Lösung (Siehe Anhang)

Bearbeitungszeit: 5 Minuten

Berechnen

Übung KTA-10

Wirkung von Alkohol auf die Reaktionszeit

In einem Experiment zur Wirkung von Alkohol auf die Reaktionszeit wurden insgesamt 400 Versuchspersonen zufällig in zwei Gruppen aufgeteilt. Eine der beiden Gruppen erhielt dabei eine standardisierte Menge Alkohol.

Abschließend ergab sich die folgende Kontingenztafel:

Gruppe/Reaktion	gut	mittel	stark verzögert	Σ
ohne Alkohol	120	60	20	
mit Alkohol	60	100	40	
Σ				

Lösung (Siehe Anhang)

Bearbeitungszeit: 5 Minuten

Appendix

Lösung Übung KTA-01

Haar- und Augenfarbe von weiblichen Models

Komplette Kontingenztafel:

	Blauäugig	Nicht blauäugig	Summe
Blond	8	2	**10**
Nicht blond	2	5	**7**
Summe	10	7	**17**

Lösung für Übung KTA-02

Vierfeldertafel in Symbolen

Antwort zu Frage 1:
$(a_1, b_1), (a_1, b_2), (a_2, b_1), (a_2, b_2)$

Antwort zu Frage 2:
Es gibt $2 \cdot 2 = 4$ Klassen

Antwort zu Frage 3:
Klasse 1, 2: n_{12}

Antwort zu Frage 4:
$(n_{1.}, n_{2.}), (n_{.1}, n_{.2})$

Antwort zu Frage 5:
Der rechte Rand

Lösung für Übung KTA-03

Trocken oder halbtrocken?

Antwort 1:
Merkmalsträger: Flasche Prosecco.
Gesamtheit: 500 Flaschen Prosecco
Identifikationsmerkmale: Proseccoflaschen (sachlich), Weinfachgeschäft Maestro (örtlich), die letzte Woche (zeitlich).
Erhebungsmerkmale: Sorte des Prosecco (X) und Herkunftsort (Y).
Skalierung der Erhebungsmerkmale: jeweils nominal skaliert.

Antwort 2:
In der folgenden 2 × 3-Kontingenztafel wird die Beziehung von Merkmalen, Sorte des Prosecco (X) und Herkunftsort (Y) dargestellt.

X\ Y	Abruzzen	Latium	Umbrien	Summe
Halbtrocken	200	70	30	**300**
Trocken	100	80	20	**200**
Summe	**300**	**150**	**50**	**500**

Antwort 3:
Die absolute und relative Randverteilung des Merkmals Sorte des Prosecco (X) wird wie folgt abgebildet.

Sorte des Prosecco (X)	absolute Randverteilung $n_{j\bullet}$	relative Randverteilung $h_{j\bullet}$
Halbtrocken	300	0,6
Trocken	200	0,4
Summe	**500**	**1**

Die für die Sorte des Proseccos relative Randverteilung sieht so aus: {0,6; 0,4}; demnach sind 60 % aller Flaschen Halbtrocken und 40 % aller Flaschen Trocken.

Antwort 4:
Die absolute und relative Randverteilung des Merkmals Herkunftsort (Y) wird ebenfalls wie folgt abgebildet.

Herkunftsort (Y)	absolute Randverteilung $n_{j\bullet}$	relative Randverteilung $h_{j\bullet}$
Abruzzen	300	0,6
Latium	150	0,3
Umbrien	50	0,1
Summe	**500**	**1**

Die für den Herkunftsort relative Randverteilung sieht so aus: {0,6; 0,3; 0,1}; demnach stammen 60 % aller Flaschen aus Abruzzen, 30 % aus Latium und 10 % aus Umbrien.

Antwort 5:
Die bedingte relative Häufigkeitsverteilung von Herkunftsort (Y) unter der Bedingung Sorte des Proseccos $X = a_j$ wird hier gezeigt.

X\ Y	Abruzzen	Latium	Umbrien
Halbtrocken	200/300	70/300	0,1
Trocken	0,5	0,4	0,1

Die bedingte relative Häufigkeitsverteilung von Sorte des Krimsekts (X) unter der Bedingung Herkunftsort Y = b_k ist hier dargestellt.

X\ Y	Abruzzen	Latium	Umbrien
Halbtrocken	200/300	70/300	0,6
Trocken	100/300	80/150	0,4

Aus der Sorte des Proseccos lassen sich z. B. zwei Häufigkeitsverteilungen {200/300; 70/300; 0,1} und {0,5; 0,4; 0,1} ableiten. Da sich beide Häufigkeitsverteilungen voneinander unterscheiden, ist bereits hier die Kontingenz angezeigt; während z. B. 66,7 % aller Flaschen der Sorte Halbtrocken aus Abruzzen stammen, waren es bei allen Flaschen der Sorte Trocken immerhin 50 %.

Lösung für Übung KTA-04

Untergang der Titanic

In der folgenden Vierfeldertafel wird die Besatzung der Titanic zahlenmäßig abgebildet.

	männlich	weiblich	Summe
Überlebt	192	20	**212**
Nicht Überlebt	670	3	**673**
Summe	**862**	**23**	**885**

Die bedingten Häufigkeiten der Besatzung werden in dieser Vierfeldertafel gezeigt.

	männlich	weiblich
Überlebt	22	87
Nicht Überlebt	78	13
Summe	**100**	**100**

Bei den Besatzungsmitgliedern der Titanic hatten die Männer eine wesentlich geringere Überlebenschance (22 %) als die Frauen (87 %).

Lösung für Übung KTA-06a

Dichotom

Kontingenztafel

	Y_1	Y_2
X_1	9	4
X_2	3	10

Lösung mit R

dichotom_loesung.R

```r
# Einlesen der Daten aus der Datei dichotom.txt
daten<-read.table("dichotom.txt",sep="\t",header=TRUE)

#  Position X1,Y1
pos11 <-
  length(subset(daten, Groesse < median(Groesse) &
                Gewicht < median(Gewicht))$Groesse)

#  Position X1,Y2
pos12 <-
  length(subset(daten, Groesse < median(Groesse) &
                Gewicht >= median(Gewicht))$Groesse)

#  Position X2,Y1
pos21 <-
  length(subset(daten, Groesse >= median(Groesse) &
                Gewicht < median(Gewicht))$Groesse)

#  Position X2,Y2
pos22 <-
  length(subset(daten, Groesse >= median(Groesse) &
                Gewicht >= median(Gewicht))$Groesse)

#  VierFelderTafel
vierfeldertafel <- matrix(c(pos11, pos12, pos21, pos22),
                          ncol = 2,
                          byrow = T)
dimnames(vierfeldertafel) <- list(c("X1", "X2"), c("Y1", "Y2"))

# Ausgabe der Vierfeldertafel
vierfeldertafel
```

Lösung mit Excel

WMS_KTA_06_Dichotom.xlsx (10 KB)

Aufgabe 1: *Erstellen Sie eine 4-Feldertafel für den bekannten Datensatz der Größe und des Gewichts von Studierenden. Wählen Sie dabei für die Größe die Klassen X1 für Größe < Median(Größe) und X2 für Größe ≥ Median(Größe). Die Klassierung des Gewichts erfolgt analog in Y1 und Y2.*

Als erstes werden die Mediane der Größe und des Gewichts berechnet.

32	Median Größe:
33	=MEDIAN(C2:C27)
34	Median Gewicht:
35	=MEDIAN(D2:D27)

Median Größe:
172,5
Median Gewicht:
72

Nun kann von allen Werten überprüft werden, ob sie kleiner oder größer/gleich dem Median sind. Dann kann von allen Personen überprüft werden, ob deren Größe und Gewicht kleiner oder größer/gleich dem jeweiligen Median sind. Es ergeben sich also vier Möglichkeiten. Wie viele Personen diese Möglichkeiten erfüllen, zählen wir und tragen es in die Tabelle ein.

Das alles wird mit der Funktion `ZÄHLEWENNS` gemacht. Im Argument der Funktion steht zunächst der erste Kriterienbereich. Dieser besteht aus den Zellen C2 bis C27, also den Größen. Dann kommt die Bedingung für die Größe, also ob sie kleiner oder größer/gleich dem Median der Größe ist. Dann folgt der zweite Kriterienbereich. Das sind die Gewichte in den Zellen D2 bis D27. Als letztes kommt die Bedingung, die für die Gewichte geprüft wird. Also auch, ob ein Gewicht jeweils kleiner oder größer/gleich dem Median des Gewichts ist.

So sehen die Einträge in die jeweiligen Zellen aus:

B40 =ZÄHLENWENNS(C2:C27;"<" & A33;D2:D27;"<" & A35)

	A	B	C
31			
32	Median Größe:		
33	=MEDIAN(C2:C27)		
34	Median Gewicht:		
35	=MEDIAN(D2:D27)		
36			
37	Kontingenztafel:		
38			
39		Y1	Y2
40	X1	=ZÄHLENWENNS(C2:C27;"	=ZÄHLENWENNS(C2:C27;"<" & A33;D2:D27;">=" & A35)
41	X2	=ZÄHLENWENNS(C2:C27;"	=ZÄHLENWENNS(C2:C27;">" & A33;D2:D27;">=" & A35)

Damit erhalten wir folgendes Ergebnis:

Kontingenztafel:		
	Y1	Y2
X1	9	4
X2	3	10

Lösung für Übung KTA-06b

Sport

Kontingenztabelle mit den dazugehörigen Randverteilungen

	Geschlecht		
Sportart	**Männlich**	**Weiblich**	**Summe**
Fußball	105	15	**120**
Turnen	25	75	**100**
Schwimmen	6	6	**12**
Tennis	24	16	**40**
Summe	**160**	**112**	**272**

Kontingenztabelle mit bedingten relativen Häufigkeiten „Geschlecht“

	Geschlecht		
Sportart	**Männlich**	**Weiblich**	**Summe**
Fußball	0,66	0,13	0,44
Turnen	0,16	0,67	0,37
Schwimmen	0,04	0,05	0,15
Tennis	0,15	0,14	0,15
Summe	**100**	**100**	**100**

Kontingenztabelle mit bedingten relativen Häufigkeiten „Sportart“

	Geschlecht		
Sportart	**Männlich**	**Weiblich**	**Summe**
Fußball	87,5	12,5	**100**
Turnen	25	75	**100**
Schwimmen	50	50	**100**
Tennis	60	40	**100**
Summe	**59**	**41**	**100**

Lösung mit R

sport_loesung.R

```
#  Sport - Aufgabe

# Einlesen der Daten
sport <-
  data.frame(
    Sportart = c("Fussball", "Turnen", "Schwimmen", "Tennis"),
    M = c(105, 25, 6, 24),
    W = c(15, 75, 6, 16)
  )

#  Kontingenztabelle mit Randverteilungen
sport.m <-
  rbind(sport, data.frame(
    Sportart = "insgesamt",
    M = sum(sport$M),
    W = sum(sport$W)
  ))
sport.mR <- cbind(sport.m, Summe = sport.m$M + sport.m$W)
sport.mR

#  Kontingenztabelle mit bedingten relativen Häufigkeiten Geschlecht
sport.bH1 <-
  data.frame(
    Sportart = sport.mR$Sportart,
    M = sport.mR$M / sport.mR$Summe,
    W = sport.mR$W / sport.mR$Summe,
    Summe = rep(1, 5)
  )
# Ausgabe
sport.bH1

#  Kontingenztabelle mit bedingten relativen Häufigkeiten Sportart
sport.bH2 <-
  data.frame(
    Sportart = sport.mR$Sportart,
    M = sport.mR$M / sport.mR$M[5],
    W = sport.mR$W / sport.mR$W[5],
    Summe = sport.mR$Summe / sport.mR$Summe[5]
  )
# Ausgabe
sport.bH2
```

Lösung mit Excel

WMS_KTA_06_Sport.xlsx (10 KB)

***Aufgabe 1**: Erstellen Sie eine Kontingenztabelle mit den dazugehörigen Randverteilungen und bestimmen Sie die relativen bedingten Häufigkeiten.*

Um die Kontingenztabelle mit den Randverteilungen zu bestimmen, berechnen wir mit der Funktion `SUMME` einfach die jeweiligen Summen.

D14 | fx =SUMME(B14:C14)

	A	B	C	D
10	Kontingenztabelle			
11				
12	Sportart	M	W	Summe
13	Fußball	105	15	=SUMME(B13:C13)
14	Turnen	25	75	=SUMME(B14:C14)
15	Schwimmen	6	6	=SUMME(B15:C15)
16	Tennis	24	16	=SUMME(B16:C16)
17	Gesamt	=SUMME(B13:B16)	=SUMME(C13:C16)	=SUMME(B17:C17)

Sportart	M	W	Summe
Fußball	105	15	120
Turnen	25	75	100
Schwimmen	6	6	12
Tennis	24	16	40
Gesamt	160	112	272

Um die bedingten relativen Häufigkeiten nach der Sportart zu ermitteln, müssen wir für jede Sportart herausfinden, wie viele Menschen (getrennt nach Frauen und Männern) sie jeweils spielen und diese Anzahl durch Gesamtzahl der Menschen teilen, die diese Sportart betreiben. Die Zeilensummen müssen also immer 1 ergeben.

D23 fx =SUMME(B23:C23)

	A	B	C	D
22	Sportart	M	W	Summe
23	Fußball	=B13/D13	=C13/D13	=SUMME(B23:C23)
24	Turnen	=B14/D14	=C14/D14	=SUMME(B24:C24)
25	Schwimmen	=B15/D15	=C15/D15	=SUMME(B25:C25)
26	Tennis	=B16/D16	=C16/D16	=SUMME(B26:C26)
27	Gesamt	=B17/D17	=C17/D17	=SUMME(B27:C27)

Sportart	M	W	Summe
Fußball	0,88	0,13	1
Turnen	0,25	0,75	1
Schwimmen	0,50	0,50	1
Tennis	0,60	0,40	1
Gesamt	0,59	0,41	1

Um die bedingten relativen Häufigkeiten nach Geschlecht herauszufinden, gehen wir fast genauso vor. Wir betrachten die beiden Geschlechter und berechnen, welchen Anteil welche Sportart hat. Hier müssen die Spaltensummen jeweils 1 ergeben.

D33 fx =D13/D17

	A	B	C	D
32	Sportart	M	W	Summe
33	Fußball	=B13/B17	=C13/C17	=D13/D17
34	Turnen	=B14/B17	=C14/C17	=D14/D17
35	Schwimmen	=B15/B17	=C15/C17	=D15/D17
36	Tennis	=B16/B17	=C16/C17	=D16/D17
37	Gesamt	=SUMME(B33:B36)	=SUMME(C33:C36)	=SUMME(D33:D36)

Sportart	M	W	Summe
Fußball	0,66	0,13	0,44
Turnen	0,16	0,67	0,37
Schwimmen	0,04	0,05	0,04
Tennis	0,15	0,14	0,15
Gesamt	1	1	1

Lösung für Übung KTA-06c

Prosecco Lieferung

1. Tabelle der absoluten Häufigkeiten mit den Randverteilungen

X\ Y	Abruzzen	Latium	Umbrien	Summe
Halbtrocken	200	70	30	**300**
Trocken	100	80	20	**200**
Summe	**300**	**150**	**50**	**500**

2. Kontingenztabelle mit den relativen Häufigkeiten

X\ Y	Abruzzen	Latium	Umbrien	Summe
Halbtrocken	0,06	0,14	0,06	0,6
Trocken	0,04	0,16	0,04	0,4
Summe	0,1	0,30	0,10	1

3. Kontingenztabelle mit den bedingten relativen Häufigkeiten nach Sorte

X\ Y	Abruzzen	Latium	Umbrien	Summe
Halbtrocken	0,67	0,23	0,1	1
Trocken	0,5	0,4	0,1	1
Summe	0,6	0,3	0,1	1

An der Kontingenztabelle ist zu erkennen wieviel Prozent welcher Sorte aus welchem Herkunftsort stammt. 10 Prozent der Sorte „Halbtrocken“ stammen beispielsweise aus Umbrien, 50 Prozent der Sorte „Trocken“ stammen aus Abruzzen und 30 Prozent des gesamten Weines stammen aus Latium

Lösung mit R

wein_loesung.R

```
#  Wein - Aufgabe

#  Einlesen der Daten in einen Data-Frame "wein"
wein <-
  data.frame(
    Sorte = c("halbtrocken", "trocken", "Summe"),
    Umb = c(30, 20, 50),
    Abr = c(200, 100, 300),
    Lat = c(70, 80, 150),
    Summe = c(300, 200, 500)
  )
# Ausgabe der absoluten Häufigkeiten
wein

#  Kontingenztabelle mit den relativen Häufigkeiten
wein.rH <-
  data.frame(
    Sorte = wein$Sorte,
    Umb = wein$Umb / 500,
    Abr = wein$Abr / 500,
    Lat = wein$Lat / 500,
    Summe = wein$Summe / 500
  )
wein.rH

#  Bedingte relative Häufigkeit des Herkunftsortes
wein.bH <-
  data.frame(
    Sorte = wein$Sorte,
    Umb = wein$Umb / wein$Summe,
    Abr = wein$Abr / wein$Summe,
    Lat = wein$Lat / wein$Summe,
    Summe = wein$Summe / wein$Summe
  )
wein.bH
```

Lösung mit Excel

WMS_KTA_06_Wein.xlsx (10 KB)

***Aufgabe 1**: Erstellen Sie einen Data Frame mit den absoluten Häufigkeiten und den Randverteilungen.*

Zunächst schreiben wir den Datensatz in Form einer Tabelle auf.

Sorte	Umbrien	Abruzzen	Latium	Summe
trocken	20	100	80	200
halbtrocken	30	200	70	300
gesamt	50	300	150	500

***Aufgabe 2**: Erstellen Sie eine Kontingenztabelle mit den relativen Häufigkeiten.*

Die Kontingenztabelle mit den Randverteilungen der relativen Häufigkeiten erhalten wir, indem wir jeden Wert durch die Gesamtanzahl der Weinflaschen teilen.

15	Sorte	Umbrien	Abruzzen	Latium	Summe
16	trocken	=B6/$E$8	=C6/$E$8	=D6/$E$8	=E6/$E$8
17	halbtrocken	=B7/$E$8	=C7/$E$8	=D7/$E$8	=E7/$E$8
18	gesamt	=B8/$E$8	=C8/$E$8	=D8/$E$8	=E8/$E$8

So sieht dann das Ergebnis aus:

Sorte	Umbrien	Abruzzen	Latium	Summe
trocken	0,04	0,2	0,16	0,4
halbtrocken	0,06	0,4	0,14	0,6
gesamt	0,1	0,6	0,3	1

***Aufgabe 3**: Berechnen Sie die bedingte relative Häufigkeit des Herkunftsortes und interpretieren Sie*
diese.

Um die bedingten relativen Häufigkeiten des Herkunftsorts (also nach Sorte) zu bestimmen, müssen wir folgendes in die jeweiligen Zellen eintragen:

25	Sorte	Umbrien	Abruzzen	Latium	Summe
26	trocken	=B6/E6	=C6/E6	=D6/E6	=SUMME(B26:D26)
27	halbtrocken	=B7/E7	=C7/E7	=D7/E7	=SUMME(B27:D27)
28	gesamt	=B8/E8	=C8/E8	=D8/E8	=SUMME(B28:D28)

Sorte	Umbrien	Abruzzen	Latium	Summe
trocken	0,10	0,50	0,40	1
halbtrocken	0,10	0,67	0,23	1
gesamt	0,10	0,60	0,30	1

Interpretation:

- 10 % der Sorte „Trocken“, 10 % der Sorte „Halbtrocken“ und 10 % insgesamt kommen aus Umbrien.
- 50 % der Sorte „Trocken“, 67 % der Sorte „Halbtrocken“ und 60 % insgesamt kommen aus Abruzzen.
- 40 % der Sorte „Trocken“, 23 % der Sorte „Halbtrocken“ und 30 % insgesamt kommen aus Latium.

Lösung für Übung KTA-07

Preise und Verpackung

X/Y	normaler Preis	erhöhter Preis	Σ
einfache Verpackung	10	8	18
aufwendige Verpackung	14	8	22
Σ	24	16	40

Lösung für Übung KTA-08

Verkehrsunfälle

Aufgabe 1:
Merkmalsträger: Verkehrsunfall.
Gesamtheit: 1000 Verkehrsunfälle.
Erhebungsmerkmale: Ausgang(X) und Geschwindigkeit(Y).

Aufgabe 2:
In der folgenden 2 x 2- Kontingenztafel wird die Beziehung der Merkmale X und Y dargestellt.

X/Y	Geschwindigkeit > 150 km/h	Geschwindigkeit < 150 km/h	Σ
tödlicher Ausgang	80	200	280
überlebt	20	700	720
Σ	100	900	1000

Aufgabe 3:
Die absolute und relative Randverteilung des Merkmals Ausgang (X) wird wie folgt abgebildet:

X	absolute Randverteilung	relative Randverteilung
tödlicher Ausgang	280	0,28
überlebt	720	0,72
Σ	1000	1

Aufgabe 4:
Die absolute und relative Randverteilung des Merkmals Geschwindigkeit (Y) wird wie folgt abgebildet:

Y	absolute Randverteilung	relative Randverteilung
> 150 km/h	100	0,1
< 150 km/h	900	0,9
Σ	1000	1

Aufgabe 5:
Die bedingte relative Häufigkeitsverteilung von Geschwindigkeit (Y).

X/Y	Geschwindigkeit > 150 km/h	Geschwindigkeit < 150 km/h
tödlicher Ausgang	80/280	200/280
überlebt	20/720	700/720

Die bedingte relative Häufigkeitsverteilung von Ausgang (X).

X/Y	Geschwindigkeit > 150 km/h	Geschwindigkeit < 150 km/h
tödlicher Ausgang	80/100	200/900
überlebt	20/100	700/900

Lösung für Übung KTA-09

Studentenstreik

Tabelle 1 (vollständig ausgefüllt)

X/Y	positiv	negativ	neutral	Σ
Naturwissenschaften	0,2	0,05	0,15	0,4
Geisteswissenschaften	0,1	0,05	0,05	0,2
Wirtschaftswissenschaften	0,1	0,2	0,1	0,4
Σ	0,4	0,3	0,3	1

Tabelle 2 (bedingte relative Häufigkeitsverteilungen von Y):

X/Y	positiv	negativ	neutral
Naturwissenschaften	0,5	0,125	0,375
Geisteswissenschaften	0,5	0,25	0,25
Wirtschaftswissenschaften	0,25	0,5	0,25

Lösung für Übung KTA-10

Wirkung von Alkohol auf die Reaktionszeit

Tabelle 1 (Vollständig ausgefüllt)

Gruppe/Reaktion	gut	mittel	stark verzögert	Σ
ohne Alkohol	120	60	20	200
mit Alkohol	60	100	40	200
Σ	180	160	60	400

Tabelle 2 (bedingte relative Häufigkeitsverteilungen):

Gruppe/Reaktion	gut	mittel	stark verzögert
ohne Alkohol	0,6	0,3	0,1
mit Alkohol	0,3	0,5	0,2