1 Deskriptive Statistik: Gemeinsame Häufigkeiten (Kontingenztabellen)

Gruppierte und klassierte Daten*

Wenn nur wenige Merkmalsausprägungen vorliegen können Daten *gruppiert* werden, d.h. es kann ausgezählt werden, wie häufig jede Merkmalsausprägung vorliegt (bei ordinal skalierten Daten werden diese auch noch nach Merkmalsausprägung sortiert). In R geschieht dies mit dem `table()` Befehl.

Bei metrisch skalierten Variablen, die häufig sehr viele Ausprägungen haben (bzw. stetig sind), müssen vorher eine überschaubare Anzahl von Intervallen gebildet werden (Klassen), und anschließend wird ausgezählt, wie oft jede Merkmalsausprägung in eine der Klassen fällt. Das Ergebnis sind *klassierte* Daten. Der R-Befehl für eine Klassierung ist `cut()`. Genau genommen entstehen durch die Klassenbildung ordinal skalierte Daten, da die Abstände nicht mehr exakt interpretierbar sind.

Beispiel: Urliste mit Merkmalsausprägungen von 8 U'objekten: $\{5, 3, 6, 5, 2, 6, 5, 3\}$

Gruppiert:

x	n_j
2	1
3	2
5	3
6	2

Klassiert: (2 Klassen)

x	n_j
(2,4]	3
(4,6]	5

1.1 Beispiel mit gruppierten Daten:

Stellen wir uns vor, 10 Personen beiderlei Geschlechts werden befragt, ob sie einer Maßnahme zustimmen oder nicht.

Urliste:

i	1	2	3	4	5	6	7	8	9	10
Geschlecht:	m	m	m	m	m	w	w	w	w	w
Zustimmung:	ja	ja	ja	nein	nein	ja	nein	nein	nein	nein

Absolute Häufigkeiten: nach Geschlecht (m/w) und Zustimmung zu einer Maßnahme (ja/nein)

Beispiel:

	Zustimmung (Z)		
	ja	nein	Sum.
m	3	2	5
w	1	4	5
Sum.	4	6	10

Allgemein:

	Zustimmung (Z)		
	ja	nein	Sum.
m	n_{11}	n_{12}	$n_{1\bullet}$
w	n_{21}	n_{22}	$n_{2\bullet}$
Sum.	$n_{\bullet 1}$	$n_{\bullet 2}$	n

mit $n_{1\bullet} := \sum_{j=1}^{2} n_{1j}$; $\quad n_{\bullet 2} := \sum_{i=1}^{2} n_{i2}$ (i ist der Zeilen- und j der Spaltenindex)

Relative Häufigkeiten (Anteile):

Beispiel:

	Zustimmung (Z)		
	ja	nein	Sum.
m	0.3	0.2	0.5
w	0.1	0.4	0.5
Sum.	0.4	0.6	1

Allgemein:

	Zustimmung (Z)		
	ja	nein	Sum.
m	n_{11}/n	n_{12}/n	$n_{1\bullet}/n$
w	n_{21}/n	n_{22}/n	$n_{2\bullet}/n$
Sum.	$n_{\bullet 1}/n$	$n_{\bullet 2}/n$	1

Zum Beispiel sind 30% aller Personen Männer, die zustimmen. Die Randhäufigkeiten sagen uns, dass jeweils 50% der Befragten Männer bzw. Frauen sind, und 40% aller Personen stimmen der Maßnahme zu, 60% lehnen sie ab.

Bedingte relative Häufigkeiten (bedingte Anteile):

a) *bedingt auf Geschlecht:* $(Z|G)$

Beispiel:

	Zustimmung (Z)		
	ja	nein	Sum.
m	$\frac{0.3}{0.5} = 0.6$	$\frac{0.2}{0.5} = 0.4$	$\frac{0.5}{0.5} = 1$
w	$\frac{0.1}{0.5} = 0.2$	$\frac{0.4}{0.5} = 0.8$	$\frac{0.5}{0.5} = 1$

Allgemein:

	Zustimmung (Z)		
	ja	nein	Sum.
m	$\frac{n_{11}}{n_{1\bullet}}$	$\frac{n_{12}}{n_{1\bullet}}$	1
w	$\frac{n_{21}}{n_{2\bullet}}$	$\frac{n_{22}}{n_{2\bullet}}$	1

60% *aller Männer* haben 'ja' und 40% *aller Männer* haben 'nein' gestimmt;

20% *aller Frauen* haben 'ja' gestimmt und 80% *aller Frauen* haben 'nein' gestimmt;

d.h. das Abstimmungsverhalten unterscheidet sich zwischen Männern und Frauen

b) *bedingt auf Zustimmung*: $(G|Z)$

Beispiel

	Zustimmung (Z)	
	ja	nein
m	$\frac{0.3}{0.4} = 0.75$	$\frac{0.2}{0.6} = 1/3$
w	$\frac{0.1}{0.4} = 0.25$	$\frac{0.4}{0.6} = 2/3$
	$\frac{0.4}{0.4} = 1$	$\frac{0.6}{0.6} = 1$

Allgemein:

	Zustimmung (Z)	
	ja	nein
m	$\frac{n_{11}}{n_{\bullet 1}}$	$\frac{n_{12}}{n_{\bullet 1}}$
w	$\frac{n_{21}}{n_{\bullet 1}}$	$\frac{n_{22}}{n_{\bullet 2}}$
	1	1

75% *aller 'ja' stimmenden Personen* sind Männer und 25% *aller 'ja' stimmenden Personen* sind Frauen;

Von *allen 'nein' stimmenden Personen* sind 1/3 Männer und 2/3 Frauen.

d.h. das Abstimmungsverhalten unterscheidet sich zwischen Männern und Frauen.

Achtung: $Z|G \neq G|Z$!!!

1.1.1 Empirische Unabhängigkeit

Frage: wie müssten die Anteile aussehen, wenn sich das Abstimmungsverhalten *nicht* unterscheiden sollte?

Die bedingten relativen Häufigkeiten (bedingten Anteile) erhielten wir durch Division durch die relativen Rand-Häufigkeiten.

Wenn die bedingten Anteile genau das *Produkt der Randhäufigkeiten* sind, dann unterscheiden sich die bedingten Anteile nicht zwischen den Gruppen, denn diese erhielten wir ja durch Division.

In diesem Fall (d.h., wenn die bedingten Anteile gleich dem Produkt der Randanteile sind), sagen wir Geschlecht (G) und Zustimmung (Z) sind *empirisch unabhängig*!

Hypothetische relative Häufigkeiten *bei empirischer Unabhängigkeit*:

Die gemeinsamen relativen Häufigkeiten (= Anteile) sind das Produkt der relativen Rand-Häufigkeiten (Gruppen-Anteile):

	Zustimmung (Z)		
	ja	nein	Sum.
m	$0.4 \times 0.5 = 0.2$	$0.6 \times 0.5 = 0.3$	0.5
w	$0.4 \times 0.5 = 0.2$	$0.6 \times 0.5 = 0.3$	0.5
Sum.	0.4	0.6	1

Bei empirischer Unabhängigkeit unterscheiden sich die *bedingten* Anteile nicht zwischen den Gruppen! Wenn wir bei empirischer Unabhängigkeit können wir aus Gruppenzugehörigkeit nichts lernen. Wenn z.B. das Geschlecht und Abstimmungsverhalten empirisch unabhängig wären, könnten wir aus der Kenntnis des Geschlechts keine Rückschlüsse über das Abstimmungsverhalten ziehen.

Für obenstehende Tabelle mit empirisch unabhängigem Verhalten erhalten wir natürlich:

Bedingt auf G: $Z|G$:

	Zustimmung (Z)		
	ja	nein	Sum.
m	$0.2/0.5 = 0.4$	$0.3/0.5 = 0.6$	1
w	$0.2/0.5 = 0.4$	$0.3/0.5 = 0.6$	1

Die Zustimmungsanteile unterscheiden sich *nicht* nach Geschlecht!

D.h., aus dem Geschlecht können wir nichts über das Abstimmungsverhalten lernen!

Bedingt auf Z: $G|Z$:

	Zustimmung (Z)	
	ja	nein
m	$0.2/0.4 = 0.5$	$0.3/0.6 = 0.5$
w	$0.2/0.4 = 0.5$	$0.3/0.6 = 0.5$
Sum.	1	1

Die Geschlechteranteile unterscheiden sich *nicht* nach Zustimmungsverhalten!
D.h., aus dem Abstimmungsverhalten können wir nichts über das Geschlecht lernen!

Allgemeiner:
Wenn wir zur Vereinfachung die relativen Häufigkeiten als $n_{ij}/n := a_{ij}$ schreiben sind die relativen Häufigkeiten *bei empirischer Unabhängigkeit*

	Zustimmung (Z)		
	ja	nein	Sum.
m	$a_{1\bullet} \times a_{\bullet 1}$	$a_{1\bullet} \times a_{\bullet 2}$	$a_{1\bullet}$
w	$a_{2\bullet} \times a_{\bullet 1}$	$a_{2\bullet} \times a_{\bullet 2}$	$a_{2\bullet}$
Sum.	$a_{\bullet 1}$	$a_{\bullet 2}$	1

Also sind die bedingten relativen Häufigkeiten *bei empirischer Unabhängigkeit*:
bedingt auf Geschlecht: $Z|G$

	Zustimmung (Z)		
	ja	nein	Sum.
m	$a_{\bullet 1}$	$a_{\bullet 2}$	1
w	$a_{\bullet 1}$	$a_{\bullet 2}$	1

bedingt auf Zustimmung: $G|Z$

	Zustimmung (Z)	
	ja	nein
m	$a_{1\bullet}$	$a_{1\bullet}$
w	$a_{2\bullet}$	$a_{2\bullet}$
Sum.	1	1

$\Rightarrow$ Wenn *alle* gemeinsamen relativen Häufigkeiten das Produkt der beiden relativen Randhäufigkeiten sind, dann und nur dann sind die Merkmale *empirisch unabhängig*!

1.2 Kontingenzmaße*

1.2.1 χ^2 Koeffizient

Zur Erinnerung, die tatsächliche Kontigenztabelle (KTab) und die entsprechende hypothetische Kontigenztabelle bei empirischer Unabhängigkeit sind:

Tatsächliche KTab:

	Zustimmung (Z)		
	ja	nein	Sum.
m	0.3	0.2	0.5
w	0.1	0.4	0.5
Sum.	0.4	0.6	1

Hyp. KTab bei empir. Unabhängigkeit:

	Zustimmung (Z)		
	ja	nein	Sum.
m	0.2	0.3	0.5
w	0.2	0.3	0.5
Sum.	0.4	0.6	1

Grundidee: Vergleich der tatsächlich beobachteten Häufigkeiten mit den hypothetischen Häufigkeiten bei empirischer Unabhängigkeit mit Hilfe einer Kennzahl $\rightarrow$ χ^2 - Koeffizient (gesprochen: chi-quadrat).

$$\chi^2 = \sum_{j=1}^{J}\sum_{l=1}^{L} \frac{(n_{jl} - n^u_{jl})^2}{n^u_{jl}} = n \sum_{j=1}^{J}\sum_{l=1}^{L} \frac{(a_{ij} - a^u_{ij})^2}{a^u_{ij}}$$

mit

$$n^u_{jl} = \frac{n_{j\bullet} n_{\bullet l}}{n} \qquad \text{und} \qquad a_{jl} := \frac{n_{jl}}{n}$$

In diesem Beispiel hat unsere Kontingenztabelle zwei Zeilen und zwei Spalten, also $J = L = 2$, und $n = 10$.

Also

$$\begin{aligned}\chi^2 &= 10\left(\frac{(0.3-0.2)^2}{0.2}+\frac{(0.2-0.3)^2}{0.3}+\right.\\ &\quad \left.\frac{(0.1-0.2)^2}{0.2}+\frac{(0.4-0.3)^2}{0.3}\right)\\ &= 1.666667\end{aligned}$$

1.2.2 Kontingenzkoeffizient nach Pearson (K)

Der Kontingenzkoeffizient nach Pearson (*Contingency Coeff*) ist

$$K = \sqrt{\frac{\chi^2}{\chi^2+n}} = \sqrt{\frac{1.666667}{1.666667+10}} = 0.3779645$$

und nimmt Werte zwischen 0 und $K_{\max} = \sqrt{\frac{\min(J,L)-1}{\min(J,L)}} < 1$ an.

$\min(J, L) = \min(2, 2) = 2$, also $K_{\max} = \sqrt{\frac{2-1}{2}} = 0.7071068$

Der **normierte (bzw. korrigierte) Kontingenzkoeffizient** nach Pearson kann nur Werte zwischen Null und Eins annehmen:

$$K^* = \frac{K}{K_{\max}} = 0.5345225$$

1.2.3 Cramers V

Das Kontingenzmaß nach Cramer V ist definiert als

$$V = \sqrt{\frac{\chi^2}{n(\min(J,L)-1)}}$$

und kann nur Werte zwischen Null und Eins annehmen.

Beispiel:

$$V = \sqrt{\frac{1.666667}{10(\min(2,2)-1)}} = \sqrt{0.1666667} = 0.4082483$$

Faustregel:

- wenn $V = 0$: kein Zusammenhang
- wenn $0 < V < 0.2$: schwacher Zusammenhang
- wenn $0.2 \leq V < 0.6$: mittlerer Zusammenhang
- wenn $0.6 \leq V < 1$: starker Zusammenhang
- wenn $V = 1$: perfekter Zusammenhang

In unserem Beispiel würden wir also von einem mittleren Zusammenhang zwischen dem Geschlecht und der beobachteten Zustimmungsrate sprechen.

In R ist das Problem natürlich einfacher zu lösen.

```
## Kontigenzmasze

G <- c("m", "m", "m", "m", "m", "w", "w", "w", "w", "w")
Z <- c("ja", "ja", "ja", "nein", "nein", "ja", "nein",
       "nein", "nein", "nein")
KTab <- table(G, Z)

# alternativ:
# KTab <- matrix(c(3, 2, 1, 4), nrow = 2, byrow = TRUE)
# rownames(KTab) <- c("m", "w")
# colnames(KTab) <- c("ja", "nein")

library(vcd)  ## package zuerst installieren
assocstats(KTab)

##                     X^2 df P(> X^2)
## Likelihood Ratio 1.7261  1  0.18891
## Pearson          1.6667  1  0.19671
##
## Phi-Coefficient   : 0.408
## Contingency Coeff.: 0.378
## Cramer's V        : 0.408
```

2 Induktive Statistik

In der induktiven Statistik interpretieren wir die relativen Häufigkeiten als Wahrscheinlichkeiten:

	Zustimmung (Z)		
	ja	nein	$f_G(G)$
m	0.3	0.2	0.5
w	0.1	0.4	0.5
$f_Z(Z)$	0.4	0.6	1

Bei *stochastischer Unabhängigkeit* würden wir folgende gemeinsamen Wahrscheinlichkeiten erwarten: (Produkt der marginalen Wahrscheinlichkeiten)

	Zustimmung (Z)		
	ja	nein	$f_G(G)$
m	0.2	0.3	0.5
w	0.2	0.3	0.5
$f_Z(Z)$	0.4	0.6	1

Kehren wir zurück zur empirischen Wahrscheinlichkeitsfunktion. Um Erwartungswerte etc. berechnen zu können müssen wir den Ausprägungen unserer Variablen Zahlen zuordnen, z.B. , z.B. 'ja' = 1 und 'nein' = 0, sowie 'm' = 0 und 'w' = 1.

	Zustimmung (Z)		
	ja	nein	
	=1	=0	$f_G(G)$
m = 0	0.3	0.2	0.5
w = 1	0.1	0.4	0.5
$f_Z(Z)$	0.4	0.6	1

Den (unbedingten) Erwartungswert von Z können wir mit Hilfe der Randwahrscheinlichkeiten (= marginalen Wahrscheinlichkeiten) einfach berechnen: $\mathrm{E}(Z) = 1 \times 0.4 + 0 \times 0.6 = 0.4$ (d.h. 40% der Personen stimmen 'ja').

Analog, $\mathrm{E}(G) = 0 \times 0.5 + 1 \times 0.5 = 0.5$ (d.h. 50% der Personen sind Frauen).

Die Varianz von G ist $\mathrm{var}(G) = \mathrm{E}(G^2) - [\mathrm{E}(G)]^2 = 0.5 - 0.25 = 0.25$

Die *auf das Geschlecht bedingten Wahrscheinlichkeiten* $\Pr(Z|G)$ sind

	Zustimmung (Z)		
	ja	nein	
	=1	=0	
m = 0	0.6	0.4	1
w = 1	0.2	0.8	1

Um die bedingten Erwartungswerte zu berechnen gewichten wir die Ausprägungen mit den bedingten Wahrscheinlichkeiten
$\mathrm{E}(Z|G=m) = \sum_i Z_i \Pr(Z_i|G=0) = 1 \times 0.6 + 0 \times 0.4 = 0.6$
$\mathrm{E}(Z|G=w) = \sum_i Z_i \Pr(Z_i|G=1) = 1 \times 0.2 + 0 \times 0.8 = 0.2$
(d.h. 60% der Männer und 20% der Frauen stimmen zu)
Dies bildet die *bedingte Erwartungswertfunktion*!

Mit Hilfe des *Gesetzes der iterativen Erwartungen* können wir aus den bedingten Erwartungswerten wieder den unbedingten Erwartungswert berechnen:
$\mathrm{E}(Z) = \mathrm{E}_G(\mathrm{E}(Z|G=0) + \mathrm{E}(Z|G=1) = 0.5 \times 0.6 + 0.5 \times 0.2 = 0.4$ (vgl. oben).

Die bedingte Varianz $\mathrm{var}(Z|G=m)$ ist
$\mathrm{var}(Z|G=m) = \mathrm{E}(Z^2|G=m) - [\mathrm{E}(Z|G=m)]^2 = 1^2 \times 0.6 + 0^2 \times 0.4 - 0.6^2 = 0.6 - 0.6^2 = 0.24$
und analog für $\mathrm{var}(Z|G=w)$

Die Kovarianz ist

$$\begin{aligned}\mathrm{cov}(Z,G) &:= \mathrm{E}[Z - \mathrm{E}(Z)][G - \mathrm{E}(G)] \\ &= \mathrm{E}[ZG - G\,\mathrm{E}(Z) - Z\,\mathrm{E}(G) + \mathrm{E}(Z)\,\mathrm{E}(G)] \\ &= \mathrm{E}(ZG) - \mathrm{E}(G)\,\mathrm{E}(Z) - \mathrm{E}(Z)\,\mathrm{E}(G) + \mathrm{E}(\mathrm{E}(Z)\,\mathrm{E}(G)) \\ &= \mathrm{E}(ZG) - 2(\mathrm{E}(G)\,\mathrm{E}(Z)) + \mathrm{E}(Z)\,\mathrm{E}(G) \\ &= \mathrm{E}(ZG) - \mathrm{E}(G)\,\mathrm{E}(Z)\end{aligned}$$

$\mathrm{E}(GZ)$ erhalten wir als Summe über alle Produkte der Merkmalsausprägungen von Z und G, jeweils multipliziert mit den gemeinsamen Wahrscheinlichkeiten.
Für unsere frühere Wahrscheinlichkeitsfunktion:

	Zustimmung (Z)		
	ja	nein	
	=1	=0	$f_G(G)$
m = 0	0.2	0.3	0.5
w = 1	0.2	0.3	0.5
$f_Z(Z)$	0.4	0.6	1

$\mathrm{E}(GZ) = 0 \times 1 \times 0.3 + 0 \times 0 \times 0.2 + 1 \times 1 \times 0.1 + 1 \times 0 \times 0.4 = 0.1$, und

$\mathrm{cov}(Z,G) = \mathrm{E}(ZG) - \mathrm{E}(G)\,\mathrm{E}(Z) = 0.1 - 0.4 \times 0.5 = -0.15.$
Damit können wir auch die *Lineare Regressionsfunktion* berechnen:

$$Z = \beta_1 + \beta_2 G + \varepsilon$$

mit

$$\begin{aligned}\beta_2 &= \frac{\mathrm{cov}(Z,G)}{\mathrm{var}(G)} = \frac{-0.15}{0.25} = -0.6 \\ \beta_1 &= \mathrm{E}(Z) - \beta_2\,\mathrm{E}(G) = 0.4 - (-0.6)0.5 = 0.7\end{aligned}$$

also ist die PRF $Z = 0.7 - 0.6G + \varepsilon$ (wenn ich mich nicht verrechnet habe :)