

Gemeinsame Häufigkeiten & Wahrscheinlichkeiten

Deskriptive Statistik

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

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

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

Relative Häufigkeiten (Anteile):

	Zustimmung (Z)		Sum.
	ja	nein	
m	0.3	0.2	0.5
w	0.1	0.4	0.5
Sum.	0.4	0.6	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): bedingt auf Geschlecht: (Z|G)

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

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

Bedingte relative Häufigkeiten: bedingt auf Zustimmung: (G|Z)

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

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

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)		Sum.
	ja	nein	
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)		Sum.
	ja	nein	
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

medskip

Die Zustimmungsteile 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!

⇒ Wenn *alle* gemeinsamen relativen Häufigkeiten das Produkt der beiden relativen Randhäufigkeiten sind, dann sind die Merkmale *empirisch unabhängig!*

Induktive Statistik

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

	Zustimmung (Z)		$f_G(G)$
	ja	nein	
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)		$f_G(G)$
	ja	nein	
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)		$f_G(G)$
	ja =1	nein =0	
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: $E(Z) = 1 \times 0.4 + 0 \times 0.6 = 0.4$ (d.h. 40% der Personen stimmen 'ja').

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

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

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

	Zustimmung (Z)		
	ja =1	nein =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

$$E(Z|G = m) = \sum_i Z_i \Pr(Z_i|G = 0) = 1 \times 0.6 + 0 \times 0.4 = 0.6$$

$$E(Z|G = w) = \sum_i Z_i \Pr(Z_i|G = 1) = 1 \times 0.2 + 0 \times 0.8 = 0.2$$

(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:

$$E(Z) = E_G(E(Z|G = 0) + E(Z|G = 1)) = 0.5 \times 0.6 + 0.5 \times 0.2 = 0.4 \text{ (vgl. oben).}$$

Die bedingte Varianz $\text{var}(Z|G = m)$ ist

$$\text{var}(Z|G = m) = E(Z^2|G = m) - [E(Z|G = m)]^2 = 1^2 \times 0.6 + 0^2 \times 0.4 - 0.6^2 = 0.6 - 0.6^2 = 0.24$$

und analog für $\text{var}(Z|G = w)$

Die Kovarianz ist

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

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

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

$\text{cov}(Z, G) = E(ZG) - E(G)E(Z) = 0.1 - 0.4 \times 0.5 = -0.15$.

Damit können wir auch die *Lineare Regressionsfunktion* berechnen:

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

mit

$$\beta_2 = \frac{\text{cov}(Z, G)}{\text{var}(G)} = \frac{-0.15}{0.25} = -0.6$$

$$\beta_1 = E(Z) - \beta_2 E(G) = 0.4 - (-0.6)0.5 = 0.7$$

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