5. ZUSAMMENHANG MEHRERER MERKMALE

Bisher haben wir meist einzelne kategorielle oder numerische Merkmale betrachtet. In den meisten statistischen Problemen geht es aber um den Zusammenhang verschiedener Größen, also um multivariate Verfahren.

Beispiel (Zusammenhang von Haar- und Augenfarbe)
Eine Stichprobe liefert die folgende Häufigkeitsverteilung für verschiedene Kombinationen von Haar- und Augenfarbe. Was lässt sich daraus über den Zusammenhang dieser beiden Merkmale schließen?

		Augenfarbe			
		Braun	Blau	Hazel	Grün
Haarfarbe	Schwarz	68	20	15	5
	Braun	119	84	54	29
	Rot	26	17	14	14
	Blond	7	94	10	16

5.1. BINÄRE MERKMALE: CHANCENQUOTIENTEN UND VIERFELDERTAFELN

Lit. [Dümbgen, Kapitel 7]

Wir betrachten zunächst den Fall, dass es für jedes Merkmal nur zwei mögliche Werte gibt (0 und 1).
Im Beispiel oben könnte wir etwa die Fragestellung darauf reduzieren, ob es einen Zusammenhang zwischen roten Haaren und grünen Augen gibt, und dementsprechend bei der Haar- und Augenfarbe nur zwischen rot / nicht rot sowie grün / nicht grün unterscheiden.

Ist A ein Ereignis mit $\mathbb{P}[A]<1$, dann definieren wir

$$\text{Chancen}[A] := \frac{\mathbb{P}[A]}{\mathbb{P}[A^c]} = \frac{\mathbb{P}[A]}{1-\mathbb{P}[A]} \in [0,\infty)$$

z. B. gilt für $\mathbb{P}[A]=\frac{1}{4}$: $\text{Chancen}[A] = 1:3$.

Da die Funktion $p \mapsto \frac{p}{1-p}$ von $[0,1)$ nach $[0,\infty)$

streng monoton wachsend und damit bijektiv ist, ist Chancen[A] einfach eine nichtlineare Transformation von $\mathbb{P}[A]$.

Vierfeldertafel und Chancenquotient

Seien nun $X, Y: \Omega \to \{0,1\}$ binäre Zufallsvariablen, und seien

$$p_{kl} = \mathbb{P}[X=k, Y=l], \quad k,l \in \{0,1\},$$

die Gewichte der gemeinsamen Verteilung $p = (p_{kl})$.

Die Randverteilungen von X und Y haben dann die Gewichte

$$p_{k+} = p_{k0} + p_{k1} = \mathbb{P}[X=k]$$

$$p_{+l} = p_{0l} + p_{1l} = \mathbb{P}[Y=l]$$

p_{11}	p_{10}	p_{1+}	Vert. von X
p_{01}	p_{00}	p_{0+}	
p_{+1}	p_{+0}	1	
Vert. von Y			

Vierfeldertafel der theoretischen Verteilung

Wir setzen im folgenden $p_{10} \neq 0$ und $p_{01} \neq 0$ voraus.

<u>DEF. 5.1</u> (Chancenquotient) Der Quotient

$$\rho := \frac{p_{11}\, p_{00}}{p_{10}\, p_{01}} = \frac{\mathbb{P}[X=1, Y=1] \cdot \mathbb{P}[X=0, Y=0]}{\mathbb{P}[X=1, Y=0] \cdot \mathbb{P}[X=0, Y=1]}$$

heißt <u>Chancenquotient</u>.

<u>BEM.</u> Es gilt

$$\rho = \frac{\mathbb{P}[X=1 \mid Y=1]}{\mathbb{P}[X=0 \mid Y=1]} \Bigg/ \frac{\mathbb{P}[X=1 \mid Y=0]}{\mathbb{P}[X=0 \mid Y=0]} = \frac{\text{Chancen}[X=1 \mid Y=1]}{\text{Chancen}[X=1 \mid Y=0]},$$

d.h. ρ misst die Abhängigkeit der Chancen für $X=1$ von Y, bzw. umgekehrt die Abhängigkeit der Chancen für $Y=1$ von X.

<u>LEMMA 5.2</u> Es gilt

$$\rho = 1 \iff X, Y \text{ unabhängig} \iff p_{kl} = p_{k+}\, p_{+l} \quad \forall k, l.$$

Achtung: Diese Aussage gilt nur für binäre Merkmale!

<u>Beweis</u> $\rho = 1$ ist äquivalent zu

$$(*) \qquad \text{Chancen}[X=1 \mid Y=1] = \text{Chancen}[X=1 \mid Y=0],$$

was gleichbedeutend ist mit

$$(**) \quad \mathbb{P}[X=1 \mid Y=1] = \mathbb{P}[X=1 \mid Y=0],$$

also der Unabhängigkeit von $\{X=1\}$ und Y.

Da X nur die Werte 0 oder 1 annimmt, ist diese äquivalent zur Unabhängigkeit von X und Y. $\square$

<u>Schätzer und Konfidenzintervall für ρ</u>

Wir wollen nun den Chancenquotienten benutzen, um zu testen ob X und Y unabhängig sind. Dazu schätzen wir zunächst ρ. Seien $(X_1,Y_1),\dots,(X_N,Y_N)$ unabhängige Kopien von (X,Y), und sei

$$H_{kl} = \#\{1 \le i \le N : X_i = k, Y_i = l\}$$

die Häufigkeit von (k,l) in der Stichprobe.

H_{11}	H_{10}	H_{1+}
H_{01}	H_{00}	H_{0+}
H_{+1}	H_{+0}	N

Häufigkeit der Werte von X

Vierfeldertafel der empirischen Häufigkeiten

Häufigkeit der Werte von Y

Als Plug-in Schätzer für ρ erhalten wir den empirischen Chancenquotienten

$$\hat{\rho} = \frac{\frac{H_{11}}{N} \cdot \frac{H_{00}}{N}}{\frac{H_{10}}{N} \cdot \frac{H_{01}}{N}} = \frac{H_{11} \cdot H_{00}}{H_{10} \cdot H_{01}} .$$

Als nächsten Schritt wollen wir Konfidenzintervalle für ρ herleiten, die auf den Statistiken H_{kl}, $k,l \in \{0,1\}$, basieren. Die gemeinsame Verteilung dieser Statistiken ist eine Multinomialverteilung:

$$(H_{11}, H_{10}, H_{01}, H_{00}) \sim \mathrm{Mult}(N, p_{11}, p_{10}, p_{01}, p_{00}) .$$

Uns interessieren aber nicht alle Parameter p_{kl}, sondern nur der Chancenquotient ρ. Wenn wir ρ kennen, dann können wir die bedingte Verteilung von H_{11} gegeben H_{1+} und H_{+1} berechnen. Umgekehrt können wir ein Konfidenzintervall für ρ erhalten, das auf H_{11}, H_{1+} und H_{+1} basiert. Das Vorgehen ist analog zur Herleitung von Fishers exaktem Test in Kapitel 1.

<u>LEMMA 5.3</u> Es existiert eine Konstante $C_{n,\ell,N,\rho} > 0$, die nicht von x abhängt, so dass für alle $\ell, n \in \{0, \dots, N\}$ und $x \in \{0, \dots, \min(\ell, n)\}$

$$(*) \quad \mathbb{P}[H_{11} = x \mid H_{+1} = \ell, H_{1+} = n] = C^{-1}_{n,\ell,N,\rho} \binom{n}{x}\binom{N-n}{\ell-x} \rho^x$$

<u>Beweis</u> Es gilt

$$\mathbb{P}[H_{11} = x \mid H_{+1} = \ell, H_{1+} = n] = \frac{\mathbb{P}[H_{11} = x, H_{01} = \ell - x, H_{10} = n - x, H_{00} = N - \ell - n + x]}{\mathbb{P}[H_{+1} = \ell, H_{1+} = n]}$$

$$\propto \frac{N!}{x!\,(\ell-x)!\,(n-x)!\,(N-\ell-n+x)!} \, p_{11}^{x} \, p_{01}^{\ell-x} \, p_{10}^{n-x} \, p_{00}^{N-\ell-n+x}$$

$$\propto \binom{n}{x}\binom{N-n}{\ell-x} \left(\frac{p_{11}\,p_{00}}{p_{10}\,p_{01}}\right)^x = \binom{n}{x}\binom{N-n}{\ell-x} \rho^x .$$

Hiebei bedeutet "$\propto$" dass beide Seiten sich nur um eine multiplikative Konstante unterscheiden, die von n, ℓ, N und p aber nicht von x abhängt. Daraus ergibt sich $(*)$, wobei $C_{n,\ell,N,\rho}$ die Normierungskonstante der bedingten Verteilung ist. $\square$

Mithilfe des Lemmas erhalten wir <u>Konfidenzintervalle</u> für den Chancenquotienten ρ. Sei dazu $\alpha \in (0,1)$, und sei $\overline{F}_{\rho,N,\ell,n}$ die Verteilungsfunktion der W'verteilung mit den Gewichten aus (*). Nach Lemma 1.5 gilt

$$\mathbb{P}\left[\overline{F}_{\rho,N,H_{+1},H_{1+}}(H_{11}) \le \alpha \mid H_{+1}=\ell, H_{1+}=n\right]$$
$$= \mathbb{P}\left[\overline{F}_{\rho,N,\ell,n}(H_{11}) \le \alpha \mid H_{+1}=\ell, H_{1+}=n\right] \le \alpha$$

für alle $\ell, n \in \{0,\dots,N\}$. Nach dem Satz von der totalen W'keit folgt

$$(**) \quad \mathbb{P}\left[\overline{F}_{\rho,N,H_{+1},H_{1+}}(H_{11}) \le \alpha\right] \le \alpha .$$

Wir betrachten nun die Funktion

$$\rho \mapsto \widetilde{F}_{N,H_{+1},H_{1+},H_{11}}(\rho) := \overline{F}_{\rho,N,H_{+1},H_{1+}}(H_{11}) .$$

Diese ist streng monoton fallend (Übung, siehe [Dümbgen, Lemma 2.5]), also invertierbar. Daher folgt aus (**) unmittelbar die erste Aussage des folgende Sätze:

SATZ 5.4 (Konfidenzschranken für Chancenquotienten)

1) $b_\alpha := \tilde{F}^{-1}_{N, H_{\cdot 1}, H_{1\cdot}, H_{11}}(\alpha)$ ist eine obere Konf.schranke für ρ zum Niveau $1-\alpha$.

2) $a_\alpha := \tilde{F}^{-1}_{N, H_{\cdot 1}, H_{1\cdot}, H_{11}-1}(1-\alpha)$ ist eine untere Konf.schranke für ρ zum Niveau $1-\alpha$.

__Beweis__ Nach (**) gilt $\mathbb{P}[\rho \geq b_\alpha] \leq \alpha$. Die zweite Aussage folgt auf ähnliche Weise. □

__BEM.__ (Fishers exakter Test) Da wir Konfidenzschranken erhalten, können wir auch die entsprechenden Testprobleme betrachten, z.B. $H_0: \rho = 1$ vs. $A_1: \rho > 1$. Wir verwerfen H_0 falls $\hat{\rho}$ auffällig groß ist. Dies ist der Fall wenn $a_\alpha > 1$ gilt, d.h.

$$1 - \alpha < \overline{F}_{1, N, H_{\cdot 1}, H_{1\cdot}}(H_{11} - 1).$$

Dies ist genau der rechtsseitige p-Wert bei Fishers exakten Test.

Vierfeldertafeln für modifiziertes Modell

Auch in einer etwas anderen Situation kann man analog vorgehen:

Beispiel (Randomisierte Studie, siehe Abschnitt 1.1)

Eine randomisierte Studie für ein neues Medikament liefert folgende Ergebnisse

	Besserung	keine Besserung	
Medikament	H_{11}	H_{12}	n_1
Placebo	H_{21}	H_{22}	n_2
	H_{+1}	H_{+2}	N

Im Unterschied zu oben betrachteten Situation sind n_1 und n_2 fest vorgegeben.

MODELL: $H_{11} \sim Bin(n_1, p_1)$, $H_{21} \sim Bin(n_2, p_2)$ unabhängig,

$$H_{12} = n_1 - H_{11}, \quad H_{22} = n_2 - H_{21}.$$

Die theoretische Verteilung ist nun

$$\begin{array}{cc} p_1 & 1-p_1 \\ p_2 & 1-p_2 \end{array}$$

Dementsprechend betrachten wir

$$\rho := \frac{p_1 (1-p_2)}{p_2 (1-p_1)}$$

Wie oben können wir schätzen durch

$$\hat{\rho} = \frac{H_{11} H_{22}}{H_{12} H_{21}}.$$

Man kann zeigen, dass Lemma 5.3 auch in dieser Situation gilt, siehe [Dümbgen, Lemma 7.2].

Damit erhält man wieder analoge Konfidenzintervalle wie oben.

5.2. Test auf Unabhängigkeit

Wir betrachten nun zwei allgemeine kategorielle Merkmale mit Werten in endlichen Mengen $S=\{a_1,\dots,a_K\}$ und $T=\{b_1,\dots,b_L\}$ mit $|S|=K$ und $|T|=L$. Seien

$$X_1,\dots,X_N : \Omega \to S \quad \text{und} \quad Y_1,\dots,Y_N : \Omega \to T$$

jeweils unabhängig und identisch verteilt mit gemeinsamer Verteilung $p=(p_{kl})$ von X_i und Y_i, d.h.

$$p_{kl} = \mathbb{P}[X_i=a_k,\, Y_i=b_l] \quad \text{für } 1\le k\le K,\ 1\le l\le L.$$

Dann sind $X=(X_1,\dots,X_N)$ und $Y=(Y_1,\dots,Y_N)$ genau dann unabhängig, wenn die gemeinsame Verteilung das Produkt der Randverteilungen ist, d.h. wenn

$$p_{kl} = p_{k+}\,p_{+l} =: \bar{p}_{kl} \quad \text{für alle } k,l.$$

Wir wollen nun diese Nullhypothese testen:

$$H_0 : \quad p = \bar{p}$$

Multiple Tests

Eine Möglichkeit ist es, Fishers exakten Test für jeden festen Wert von k und l durchzuführen. Dies liefert sehr genaue Informationen über den Zusammenhang von X und Y, erfordert aber eine große Anzahl von Stichproben um ein vorgegebenes Niveau zu erreichen, da wir K·L Tests durchführen (Bonferroni-Korrektur).

Chiquadrat-Test auf Unabhängigkeit

Wir betrachten nun eine andere Option analog zu Anpassungstests: Wir schätzen die Chiquadrat-Divergenz

$$\chi^2(p|\bar{p}) = \sum_{k,l} \left(\frac{p_{kl}}{\bar{p}_{kl}} - 1 \right)^2 \bar{p}_{kl} = \sum_{k,l} \frac{p_{kl}^2}{\bar{p}_{kl}} - 1$$

zwischen der Verteilung und dem Produkt der Randverteilungen.

Statt der Chiquadrat-Divergenz kann man auch wieder die relative Entropie betrachten. Ein Schätzer für $N \cdot \chi^2(p|\bar{p})$ ist die entsprechende

Chiquadrat-Divergenz der empirischen Verteilungen:

$$T(X,Y) = N \cdot \chi^2\left(L_N^{X,Y} \,\|\, L_N^X \otimes L_N^Y\right)$$

$$= \sum_{k,l} \left(\frac{H_{kl}(X,Y)}{\overline{H}_{kl}(X,Y)} - 1\right)^2 \overline{H}_{kl}(X,Y) = \sum_{k,l} \frac{H_{kl}(X,Y)^2}{\overline{H}_{kl}(X,Y)} - N,$$

wobei $H_{kl}(X,Y) = \sum_{i=1}^{N} 1_k(X_i)\, 1_l(Y_i)$ die Häufigkeit von (k,l), und

$$\overline{H}_{kl}(X,Y) := \frac{H_k(X) \cdot H_l(Y)}{N}$$

ist. Unter der Nullhypothese sollte für große N gelten:

$$\frac{H_{kl}(X,Y)}{N} \overset{N \text{ groß}}{\approx} p_{kl} \overset{H_0}{=} p_{k+}\, p_{+l} \overset{N \text{ groß}}{\approx} \frac{H_k(X)}{N} \cdot \frac{H_l(Y)}{N} = \frac{\overline{H}_{kl}(X,Y)}{N}.$$

Die Chiquadrat-Statistik $T(X,Y)$ misst die mittlere quadratische Abweichung dieser Größen, welche wir in einer Kontingenztafel darstellen können:

Unabhängigkeit

	b_1	$\cdots$	b_L	
a_1	H_{11}	$\cdots$	H_{1L}	$H_1(X)$
$\vdots$	$\vdots$		$\vdots$	$\vdots$
a_K	H_{K1}	$\cdots$	H_{KL}	$H_K(X)$
	$H_1(Y)$	$\cdots$	$H_L(Y)$	N

Kontingenztafel der Häufigkeiten

	b_1	$\cdots$	b_L	
a_1	$\overline{H}_{11}$	$\cdots$	$\overline{H}_{1L}$	$H_1(X)$
$\vdots$	$\vdots$		$\vdots$	$\vdots$
a_K	$\overline{H}_{K1}$	$\cdots$	$\overline{H}_{KL}$	$H_K(X)$
	$H_1(Y)$	$\cdots$	$H_L(Y)$	N

erwartete Häufigkeiten im Produktfall (genaue Aussage siehe Lemma 5.5)

Analog zum χ^2-Anpassungstest betrachten wir nun den folgenden Unabhängigkeitstest, wobei $c > 0$ ein vorzugebender Schwellenwert ist.

<u>χ^2-Test auf Unabhängigkeit</u>: Verwerfe H_0, falls $T(X,Y) \geq c$.

Sei $t = T(x,y)$ der beobachtete Wert der χ^2-Statistik.
Eine exakte Berechnung des entsprechenden p-Werts ist normalerweise nicht möglich. Stattdessen bieten sich zwei approximative Verfahren an:

a) <u>Bootstrap</u> Unter H_0 gilt $p_{kl} = p_{k+} p_{+l} \approx \frac{H_{k+}}{N} \frac{H_{+l}}{N}$ für große N. Daher kann man wie folgt vorgehen, um den p-Wert zu schätzen:

- Simuliere $B \cdot N$ Stichproben vom Produkt der empirischen Verteilungen von X und Y.
- Erstelle daraus B Simulationen von Kontingenztafeln, und berechne Werte $t_1, \dots, t_B$ der χ^2-Statistik.
- Berechne den Monte-Carlo-p-Wert

$$\hat{p}_{MC} = \frac{\#\{i \in \{1, \dots, B\} : t_i \geq t\} + 1}{B+1}$$

b) Chiquadrat-Approximation (N groß)

Man kann zeigen, dass die Verteilung von $T(X,Y)$ unter der Nullhypothese für $N \to \infty$ gegen eine χ^2-Verteilung mit $(K-1)\cdot(L-1)$ Freiheitsgraden konvergiert.

Damit ergibt sich ein approximativer Test, der H_0 zum Niveau α verwirft, falls

$$T(X,Y) \geq q_{1-\alpha;\,\chi^2((K-1)(L-1))}$$

gilt. Der approximative p-Wert ist

$$p_r \approx 1 - F_{\chi^2((K-1)(L-1))}\left(T(X,Y)\right).$$

Der vollständige Beweis der Konvergenz gegen die χ^2-Verteilung findet sich z.B. in [Georgii, Satz 11.17].

Wir motivieren hier nur anschaulich, warum die Anzahl der Freiheitsgrade gleich $(K-1)(L-1)$ ist.

<u>Heuristik</u> In diesem Fall ist der Parameterraum Θ die Menge aller W'verteilungen auf $\{a_1, \dots, a_K\} \times \{b_1, \dots, b_L\}$. Diese Menge ist ein Simplex im $\mathbb{R}^{K \cdot L}$:

$$\Theta = \{ p = (p_{kl}) : p_{kl} \geq 0 \ \forall k, l, \ \sum p_{kl} = 1 \},$$

hat also Dimension $K \cdot L - 1$. Die Parametermenge der Nullhypothese ist

$$\Theta_0 = \{ P_X \otimes P_Y : P_X \in WV(\{a_1, \dots, a_K\}), P_Y \in WV(\{b_1, \dots, b_L\}) \}$$

Diese Menge hat Dimension $K-1+L-1$. Die Anzahl der im orthogonalen Komplement von Θ_0 verbleibenden Freiheitsgrade ist also

$$\dim(\Theta) - \dim(\Theta_0) = KL - 1 - K - L + 2 = (K-1) \cdot (L-1)$$

Hat man dies realisiert, dann kann man die Konvergenz gegen die Verteilung $\chi^2((K-1) \cdot (L-1))$ auf ähnliche Weise zeigen wie beim χ^2-Anpassungstest, siehe [Georgii, Satz 11.17]

Die Nullhypothese der Unabhängigkeit kann man noch etwas abschwächen.

Verallgemeinerte Nullhypothese

(H_0') Bedingte Austauschbarkeit von Y gegeben X:

Für jede Permutation $\pi \in \mathcal{S}_N$ gilt

$$(X, {}_\pi Y) \sim (X, Y),$$

wobei ${}_\pi Y := (Y_{\pi(1)}, \ldots, Y_{\pi(N)})$

Aus (H_0) folgt (H_0'), denn, da $Y_1, \ldots, Y_N$ nach Voraussetzung unabhängig und identisch verteilt sind, gilt ${}_\pi Y \sim Y$ für alle $\pi \in \mathcal{S}_N$. Unter (H_0) ist X unabhängig von Y, und damit gilt auch $(X, {}_\pi Y) \sim (X, Y)$ für alle $\pi \in \mathcal{S}_N$. Die folgenden Aussagen gelten sogar unter der schwächeren Annahme (H_0').

LEMMA 5.5 Sei $\pi \sim \mathrm{Unif}(\mathcal{S}_N)$ eine zufällige Permutation der Indizes. Dann gilt für alle $x=(x_1, \dots, x_N)$ und $y=(y_1, \dots, y_N)$:

1) $\mathbb{E}[H_{kl}(x, \pi y)] = \overline{H}_{kl}(x, y)$ für alle k, l

2) $\mathbb{E}[T(x, \pi y)] = \frac{N}{N-1}(k-1)(L-1)$.

Zum Beweis betrachtet man die Darstellung

$$H_{kl}(x, \pi y) = \sum_{i=1}^{N} 1_k(x_i)\, 1_l(y_{\pi(i)}),$$

und berechnet damit die Erwartungswerte von $H_{kl}(x, \pi y)$ und $H_{kl}(x, \pi y)^2$, siehe Übung.

SATZ 5.6 Unter (H_0') gilt

1) $\mathbb{E}[H_{kl}(X, Y)] = \mathbb{E}[\overline{H}_{kl}(X, Y)]$ für alle k, l

2) $\mathbb{E}[T(X, Y)] = \frac{N}{N-1}(k-1)\cdot(L-1)$

3) $\mathbb{P}[T(X, Y) \le c] \xrightarrow{N \uparrow \infty} F_{\chi^2((k-1)(L-1))}(c)$ für alle $c \in \mathbb{R}$

Die Aussagen 1) und 2) folgen aus dem Lemma, siehe Übung. Zum Beweis von 3) siehe [Georgii, Satz 11.17].

5.3. PERMUTATIONSTESTS

Literatur: [Dümbgen, Abschnitt 8.1-8.3]

Wir betrachten nun ein allgemeines Verfahren um auf Symmetrien (z.B. Invarianz unter Permutationen) zu testen. Sei

$$(X,Y) = (X_1, \dots, X_n, Y_1, \dots, Y_m) : \Omega \to S^{n+m}$$

ein Stichprobenvektor mit Werten in S^{n+m}, wobei $(S, \mathcal{B})$ ein messbarer Raum ist. Weiterhin sei G ~~eine~~ endliche Gruppe von Bijektionen $g: S^{n+m} \to S^{n+m}$.

Wir wollen die Nullhypothese testen, dass die Verteilung von (X,Y) invariant unter G ist.

(H_0) <u>G-Invarianz</u>: Für alle $g \in G$ gilt

$$g(X_1,\dots,X_n,Y_1,\dots,Y_m) \sim (X_1,\dots,X_n,Y_1,\dots,Y_m)$$

In diesen allgemeinen Rahmen lassen sich viele interessante Fälle einbetten.

<u>BEISPIELE</u> 1) <u>Test auf Austauschbarkeit</u>
Sei $G = \mathcal{S}_{n+m}$ die Gruppe aller Permutationen von $\{1,\dots,n+m\}$. Wir setzen $X_{n+k} := Y_k$ für $k=1,\dots,m$ und definieren für $\pi \in G$:

$$\overline{\pi}(X,Y) = \pi(X_1,\dots,X_{n+m}) = (X_{\pi(1)},\dots,X_{\pi(n+m)}),$$

d.h. die Datenwerte X_i und Y_j werden permutiert. Die Nullhypothese lautet dann

(H_0) $X_1,\dots,X_n,Y_1,\dots,Y_m$ sind <u>austauschbar</u>.

Sie ist insbesondere erfüllt wenn diese Zufallsvariablen unabhängig und identisch verteilt sind. Einen Test von H_0 können wir daher verwenden, um zu widerlegen, dass die X_i und Y_j identisch verteilt sind.

2) Test auf bedingte Austauschbarkeit

Sei $G = S_m$. Für $\pi \in G$ setzen wir

$$\pi(X,Y) := (X, \pi Y),$$

d.h. es werden nur die Werte von Y permutiert.

Die Nullhypothese lautet dann:

(H_0) $Y_1, \ldots, Y_m$ sind bedingt austauschbar gegeben $X_1, \ldots, X_n$.

Dies ist insbesondere erfüllt, wenn $Y_1, \ldots, Y_m$ austauschbar sind, und unabhängig von $X_1, \ldots, X_n$.

Mit einem Test von H_0 können wir also widerlegen, dass X und Y unabhängig sind. Dies haben wir auch bereits im letzten Abschnitt verwendet.

3) Test auf Vorzeichensymmetrie

Sei $n = 0$ und

$$G := \{+1, -1\}^m = \{ s = (s_i)_{i=1,\dots,m} : s_i \in \{+1, -1\} \}$$

mit der komponentenweisen Multiplikation

$$s \circ \tilde{s} := (s_i \cdot \tilde{s}_i)_{i=1,\dots,m}$$

Die Nullhypothese lautet nun

(H_0) Die Verteilung von $Y = (Y_1, \dots, Y_m)$ ist invariant unter beliebigen Vorzeichenwechseln der Y_i.

Einen Test von H_0 kann man z.B. verwenden, um eine Tendenz für positive oder negative Vorzeichen aufzuzeigen.

Die (allgemeine) Nullhypothese lässt sich auch noch etwas anders formulieren:

LEMMA 5.7 Sei $\Pi \sim \mathrm{Unif}(G)$ ein zufälliges Element aus G, unabhängig von X und Y. Dann ist H_0 genau dann erfüllt, wenn

$$(H_0') \quad \Pi(X,Y) \sim (X,Y).$$

Beispielsweise ist Austauschbarkeit äquivalent zu Invarianz der gemeinsamen Verteilung unter einer unabhängigen Zufallspermutation.

Beweis $H_0 \Rightarrow H_0'$. Ist H_0 erfüllt, dann gilt $g(X,Y) \sim (X,Y)$ für alle $g \in G$. Mit dem Satz von der totalen Wkeit folgt

$$\mathbb{P}[\Pi(X,Y) \in B] = \sum_{g \in G} \frac{1}{|G|} \mathbb{P}[g(X,Y) \in B] = \mathbb{P}[(X,Y) \in B]$$

für alle $B \in \mathcal{B}$.

$H_0' \Rightarrow H_0$. Sei $g \in G$ beliebig. Dann ist die Abbildung $\pi \mapsto g \circ \pi$ eine Bijektion auf G. Daher ist mit Π auch $g \circ \Pi$ gleichverteilt auf G und unabhängig von (X,Y). Somit gilt

$$\mathbb{P}[g(\Pi(X,Y)) \in B] = \mathbb{P}[\Pi(X,Y) \in B] \quad \text{für alle } B \in \mathcal{B}.$$

Ist H_0' erfüllt, dann hat $\Pi(X,Y)$ dieselbe Verteilung wie (X,Y), und somit erhalten wir

$$\mathbb{P}[g(X,Y) \in B] = \mathbb{P}[(X,Y) \in B] \quad \text{für alle } B \in \mathcal{B}$$

$\square$

Konstruktion von Tests via G-Symmetrie

Um Tests für die Nullhypothese H_0 bzw. H_0' zu konstruieren, wählen wir eine beliebige Statistik $T : S^{n+m} \to \mathbb{R}$. Unter der Nullhypothese H_0' hat $T(\Pi(X,Y))$ dieselbe Verteilung wie $T(X,Y)$. Daher können wir erwarten, dass das Ereignis, dass der Wert der Statistik für die permutierten

Daten ~~nur~~ selten ~~kleiner~~ ist als der Wert für die Ausgangsdaten, eine kleine W'keit hat. Um dies zu präzisieren definieren wir einen modifizierten p-Wert.

<u>DEF. 5.8</u> (p-Wert für Test auf G-Symmetrie)

Für einen Datenvektor $(x,y) \in S^{n+m}$ setzen wir

$$p_\ell^G(x,y) := \frac{|\{\pi \in G : T(\pi(x,y)) \le T(x,y)\}|}{|G|} = \mathbb{P}\left[T(\Pi(x,y)) \le T(x,y)\right]$$

$$p_r^G(x,y) := \frac{|\{\pi \in G : T(\pi(x,y)) \ge T(x,y)\}|}{|G|} = \mathbb{P}\left[T(\Pi(x,y)) \ge T(x,y)\right]$$

<u>BEM.</u> 1) Für $(x,y) \in S^{n+m}$ gilt

$$p_\ell^G(x,y) = F_{T(\Pi(x,y))}\big(T(x,y)\big),$$

$$p_r^G(x,y) = 1 - F_{T(\Pi(x,y))}\big(T(x,y)-\big).$$

Hierbei ist nur die Permutation Π zufällig!
Insbesondere hängt die Verteilungsfkt. nicht von der Verteilung von (X,Y) ab, d.h. der entsprechende Test ist verteilungsunabhängig.

2) Im Gegensatz dazu sind die klassischen p-Werte definiert als

$$p_e(x,y) = \mathbb{P}\left[T(X,Y) \leq T(x,y)\right] = F_{T(X,Y)}(T(x,y)),$$

$$p_r(x,y) = \mathbb{P}\left[T(X,Y) \geq T(x,y)\right] = 1 - F_{T(X,Y)}(T(x,y)-).$$

Hier ist (X,Y) zufällig und die Verteilungsfunktion hängt somit von der Verteilung von (X,Y) ab.

3) Ist $|G|$ klein, dann kann man die p-Werte durch Abzählen ermitteln. Normalerweise ist dies aber nicht praktikabel, z. B. gilt $|S_n| = n!$.

Stattdessen betrachtet man dann die Monte-Carlo p-Werte

$$p_e^{MC}(x,y) = \frac{\left|\{i \in \{1,\dots,B\} : T(\pi_i(x,y)) \leq T(x,y)\}\right| + 1}{B+1}$$

und entsprechend $p_r^{MC}(x,y)$, wobei $\pi_1, \dots, \pi_B \sim \mathrm{Unif}(G)$ unabhängige Stichproben von der Gleichverteilung auf G sind

SATZ 5.9 Unter H_0 gilt

$$\mathbb{P}\left[p_e^G(X,Y) \le \alpha\right] \le \alpha \text{ und } \mathbb{P}\left[p_r^G(X,Y) \le \alpha\right] \le \alpha$$

für alle $\alpha \in [0,1]$.

Beweis Nach Lemma 5.7 folgt aus H_0:

$$(*) \quad \mathbb{P}\left[p_e^G(X,Y) \le \alpha\right] = \mathbb{P}\left[p_e^G(\Pi(X,Y)) \le \alpha\right]$$

Wir wollen zeigen, dass die rechte Seite kleiner oder gleich α ist. Nach dem Satz von der totalen W'keit ist dies der Fall, wenn

$$(**) \quad \mathbb{P}\left[p_e^G(\Pi(x,y)) \le \alpha\right] \le \alpha \quad \text{für alle } (x,y) \in S^{n+m}$$

gilt. Wegen

$$p_e^G(\Pi(x,y)) = F_{T(\Pi(x,y))}(\Pi(x,y))$$

folgt $(**)$ aus Lemma 1.5. Der Beweis für den rechtsseitigen p-Wert verläuft analog. □

Nach Satz 5.9 erhalten wir einen Hypothesentest für H_0, der die Niveaubedingung exakt einhält:

<u>Exakter Symmetrietest</u>

Verwerfe H_0 falls $p_\ell^G(X,Y) \le \alpha$ (bzw. falls $p_r(X,Y) \le \alpha$).

In der Praxis ersetzen wir den p-Wert durch den Monte-Carlo p-Wert.

<u>Monte Carlo Symmetrietest</u>

Verwerfe H_0 falls $p_\ell^{MC}(X,Y) \le \alpha$ (bzw. falls $p_r^{MC}(X,Y) \le \alpha$)

Auch der Monte Carlo Test erfüllt die Niveaubedingung exakt. Dies zeigt man auf ähnliche Weise wie im Beweis von Lemma 4.4, siehe [Dümbgen, Lemma 8.3]

Beispiele

1) Test auf identische Verteilung der X_i und Y_j

Hier wählen wir $G = \mathcal{S}_{n+m}$, siehe oben, und betrachten eine Statistik der Form

$$T(X,Y) = |M(X) - M(Y)| .$$

Dabei ist M eine Statistik bzgl. der wir den Unterschied zwischen den Verteilungen erkennen möchten, z. B. der Mittelwert, Median, oder eine beliebige andere Kenngröße. Welches M wir betrachten hängt von der Form der Alternative ab.

2) Test auf Unabhängigkeit bzw. bedingte Austauschbarkeit

Wir betrachten beispielsweise gepaarte Daten (X_i, Y_i), $i = 1, \dots, n$, mit Werten in $\{0,1\} \times \mathbb{R}$.
Um zu testen ob X und Y unabhängig sind, wählen

wir $G = \mathcal{S}_n$ und setzen $\pi(X,Y) := (X, \pi Y)$.

Wieder kann man viele Arten von Teststatistiken betrachten, beispielsweise

$$T(X,Y) = \sum_{i=1}^{n} \mathbb{1}_{\{X_i = 1\}} R_i(Y),$$

wobei $R_i(Y)$ der Rang von Y_i unter $Y_1, \ldots, Y_n$ ist. Hier würde man testen, ob die Y_i tendenziell größere oder kleinere Ränge haben, falls $X_i = 1$ gilt.

3) Zeitreihen Sei $n = 0$ und $Y_1, \ldots, Y_m : \Omega \to \{0,1\}$.

Um zu testen, ob die Y_i austauschbar sind, oder ob es eine Abhängigkeit der Wert vom (Zeit-) Index i gibt, wählt man $G = \mathcal{S}_m$. Teststatistiken für verschiedene Alternativen sind zum Beispiel

$$T(Y) = \sum_{i=1}^{m} Y_i \cdot i \quad \text{(Test auf monotonen Trend)},$$

$$T(Y) = \sum_{i=1}^{m-1} \mathbb{1}_{Y_i \neq Y_{i+1}} \quad \text{(Test auf Clustering)}.$$

4) <u>Vorzeichentests</u> Sei $n=0$ und $Y_1, \dots, Y_m : \Omega \to \mathbb{R}$

Um Vorzeichensymmetrie zu testen, betrachten wir $G = \{+1, -1\}^m$. Teststatistiken sind etwa

$$T(Y) = \sum_{i=1}^{m} \operatorname{sign}(Y_i) \quad \text{(Pearsons Vorzeichentest)}$$

$$T(Y) = \sum_{i=1}^{m} \operatorname{sign}(Y_i)\, \tilde{R}_i \quad \text{(Wilcoxon signed rank)}$$

Diese Tests haben wir schon oben betrachtet.