EDV in Medizin und Biologie **10** (1), S. 19–21, ISSN 0300-8282

Zum Problem der Bestimmung zweiseitiger Überschreitungswahrscheinlichkeiten beim exakten Vierfeldertest

Hans-Peter Krüger

Zusammenfassung

Drei Methoden zur Bestimmung zweiseitiger Überschreitungswahrscheinlichkeiten bei der exakten Prüfung von Vierfeldertafeln nach FISHER-YATES werden diskutiert. An einem Beispiel wird aufgezeigt, daß diese Methoden zu verschiedenen Signifikanzentscheidungen führen können. Es werden Möglichkeiten aufgezeigt, wie diese Schwierigkeiten in der Praxis umgangen werden können.

Summary

Three methods to determine two-sided exact probabilities in fourfold tables following FISHER-YATES are discussed. An example illustrates that these methods may lead to different decisions. Some hints are given how these difficulties could be avoided in practical work.

Wird zur Überprüfung der Signifikanz der exakte Vierfeldertest nach FISHER (1956, 98) und YATES (1934) herangezogen, stellen sich in der Praxis zwei Probleme:
a) wie können zweiseitige Überschreitungswahrscheinlichkeiten festgelegt werden?
b) wie ist zu entscheiden, wenn diese exakt ermittelten Überschreitungswahrscheinlichkeiten das festgelegte Signifikanzniveau nur »um weniges« übertreffen?

Beide Fragestellungen treten bereits beim exakten Binomialtest auf (z.B. Vorzeichentest, Anpassungs- und Randomisierungstests). Fragestellung b) ist bei jedem exakten Test vorhanden. Das hier vorgeschlagene Procedere gilt für diese Fälle analog.

Verfahren zur Bestimmung zweiseitiger Überschreitungswahrscheinlichkeiten

Die exakte Wahrscheinlichkeit einer beobachteten Vierfeldertafel unter der Bedingung festgehaltener Randsummen (= Punktwahrscheinlichkeit) ergibt sich aus der Hypergeometrischen Verteilung zu

$$p_0 = \frac{(a+b)!\,(c+d)!\,(a+c)!\,(b+d)!}{N!\,a!\,b!\,c!\,d!}$$

$$= \frac{N_1!\,N_2!\,M_1!\,M_2!}{N!\,a!\,b!\,c!\,d!}$$

Dabei sind a, b, c und d die Zellenfrequenzen, N_1 und N_2 die Zeilensummen, M_1 und M_2 die Spaltensummen, N der Stichprobenumfang. Zur Bestimmung der exakten Überschreitungswahrscheinlichkeit P werden die Punktwahrscheinlichkeiten p_i aller extremeren Konfigurationen a_i, b_i, c_i und d_i zu diesen Randsummen aufsummiert. Dieses Verfahren bereitet für die einseitigen Überschreitungswahrscheinlichkeiten P_1 keine Schwierigkeiten, da hier das Problem der Definition der »extremeren« Tafeln nicht auftritt. Anders dagegen bei den zweiseitigen Überschreitungswahrscheinlichkeiten P_2.

Drei verschiedene Methoden zur Bestimmung von P_2 werden angewendet:
a) das Verfahren nach FREEMAN und HALTON (1951). Die Autoren empfehlen, alle Punktwahrscheinlichkeiten p_i aufzusummieren, die kleiner oder gleich dem p_o der beobachteten Tafel sind. Bei der Prüfung auf Homogenität werden damit alle Tafeln in die Ablehnungsregion von H_0 einbezogen, deren Anteilsdifferenzen $D = |a/N_1 - c/N_2|$ größer oder gleich der beobachteten Anteilsdifferenz sind. Im Fall der Kontingenzprüfung werden alle Tafeln einbezogen, deren Kontingenzkoeffizient $\varphi = (ad-bc)/\sqrt{N_1N_2M_1M_2}$ größer oder gleich dem φ der beobachteten Tafel ist.
b) die »Niveau-Verfahren«. Bei stetigen Prüfverteilungen ist diese Methode das Verfahren der Wahl. An beiden Enden der Verteilung wird eine Ablehnungsregion der Größe $\alpha/2$ definiert. Das kann bei den diskreten und in der Regel schiefen Nullverteilungen des exakten Vierfeldertests auf zwei Arten nachgebildet werden:
b_1) das $2P_1$-Verfahren. Man bestimmt die einseitige Überschreitungswahrscheinlichkeit P_1 einer Tafel und prüft diese gegen ein Niveau von $\alpha/2$. Rechentechnisch äquivalent ist eine Verdoppelung von P_1 und eine Prüfung gegen α. Dieses Verfahren wird häufig in der Literatur angegeben (z.B. bei PEARSON u. HARTLEY, 1970, 73 und bei OWEN, 1962, 480).
b_2) das $\alpha/2$-Verfahren. An beiden Enden der Prüfverteilung werden von den Extremen her die Punktwahrscheinlichkeiten p_i solange aufsummiert, bis jeweils das $\alpha/2$ eines vorgegebenen Signifikanzniveaus möglichst nah erreicht, nicht aber überschritten wird. Die unterhalb dieser Grenzen liegenden Tafeln bilden die Ablehnungsregion. Überschreitet die kleinste Punktwahrscheinlichkeit auf einer Seite der Verteilung bereits dieses $\alpha/2$, muß H_0 beibehalten werden. In aller Regel sind die beiden Verfahren b_1 und b_2 identisch. Sie unterscheiden sich lediglich in dem Fall, daß die kleinste Punktwahrscheinlichkeit bereits größer als $\alpha/2$ ist. Dann ist nach b_2 kein zweiseitiger Test möglich. Als Konsequenz muß H_0 beibehalten werden.

Das Problem der Rundung von Wahrscheinlichkeiten

Vor der Demonstration an einem Beispiel muß die Frage der Rundung von Wahrscheinlichkeiten diskutiert werden.

Die einfache »kaufmännische« Rundung (Aufrundung ab 5) läßt sich bei unkonventionellen Signifikanzgrenzen nicht durchführen. Werden etwa 3 simultane Tests zu einem $\alpha = 0.01$ durchgeführt, resultiert daraus ein adjustiertes $\alpha^\circ = \alpha/3 = 0.0033$ (siehe dazu KRAUTH & LIENERT, 1973). Bislang existiert keine einheitliche Regelung für ein Vorgehen in diesen Fällen. Folgendes Procedere wird vorgeschlagen:

Das Signifikanzniveau wird jeweils auf zwei Stellen nach den führenden Nullen ausgedrückt, also z.B. $\alpha = 0.050$, $\alpha = 0.01/6 = 0.0017$. Die exakten Wahrscheinlichkeiten werden auf die gleiche Stellenzahl berechnet, wobei kaufmännisch gerundet wird. Zum gegebenen Niveau α wird H_0 solange abgelehnt, solange die gefundene Wahrscheinlichkeit P das α um weniger als $q = 10\,\%$ übersteigt. Der absolute Fehler ist damit umso kleiner, je extremer das α angesetzt ist. Für das übliche 5%Niveau entspricht das der gewohnten kaufmännischen Rundung. Für ein $\alpha = 0.0017$ wird H_0 dann beibehalten, wenn $P > \alpha + \alpha/10 > 0.0017 + 0.0002 > 0.0019$ ist.

Die Verfahren an einem Beispiel

Die drei Verfahren der Bestimmung von zweiseitigen Überschreitungswahrscheinlichkeiten können für dieselben Vierfeldertafeln einer gegebenen Randverteilung zu verschiedenen Entscheidungen führen. Das zeigt das Beispiel der Tabelle 1 mit $N = 39$, $N_1 = 11$, $M_1 = 25$. Für die zwei Niveaus $\alpha = 5\,\%$ und $\alpha = 0.2\,\%$ (etwa 5 simultane Tests zu $\alpha = 1\,\%$) sind in der letzten Spalte die Tafeln angekreuzt, die zur Ablehnung von H_0 zu den beiden Niveaus führen würden. Es zeigen sich folgende Unterschiede:

a) bei $\alpha = 5\,\%$ wird Tafel H mit $a = 4$ nur nach dem FREEMAN-HALTON-Verfahren signifikant.[1]

b) bei $\alpha = 0.2\,\%$ werden die Tafeln J, K und L nach dem FREEMAN-HALTON- und nach dem $2P_1$-Verfahren signifikant,

[1] Nach der angegebenen Rundungsregel würde für ein zweiseitiges $\alpha = 0.0050$ Tafel A nach allen 3 Verfahren signifikant: $\alpha + \alpha/10 = 0.0050 + 0.0005 = 0.0055$. Dieser Wert wird von $2\,P_1 = 0.00532$ unterschritten.

Tab. 1. Zusammenstellung der phi-Koeffizienten, der ein- und zweiseitigen Überschreitungswahrscheinlichkeiten nach drei Methoden für ein Beispiel mit festgehaltenen Randsummen

	Zellenfrequenzen					Punktwahrsch.	einseitige Überschreitungswahr.	zweiseitige Überschreitungswahrscheinl.		Testentscheidung für $\alpha = 5\,\%$			$\alpha = 0.2\,\%$		
Nr.	a	b	c	d	φ	p°	P_1	P_1	nach FH	nach $2P_1$	FH	$2P_1$	α/2	FH	$2P_1$	α/2
A	11	0	14	14	+.469	.00266 (Σ aus	–	.00266 A	.00303 A, J–L	.00532	×	×	×	–	–	–
B	10	1	15	13	+.350	.02730 (Σ aus	–	.02996 A–B	.06035 A–B, H–L	.05992	–	–	–	–	–	–
C	9	2	16	12	+.231	.11092 (Σ aus	–	.14088 A–C	.23607 A–C, G–L	.28176	–	–	–	–	–	–
D	8	3	17	11	+.113	.23489 (Σ aus	–	.37577 A–D	.71289 A–D, F–L	.75134	–	–	–	–	–	–
E	7	4	18	10	−.006	.28709 (Σ aus	.62421 E–L	–	.99999 A–L	–	–	–	–	–	–	–
F	6	5	19	9	−.125	.21154 (Σ aus	.33712 F–L	–	.47800 A–C, F–L	.67424	–	–	–	–	–	–
G	5	6	20	8	−.244	.09519 (Σ aus	.12558 G–L	–	.15554 A–B, G–L	.25116	–	–	–	–	–	–
H	4	7	21	7	−.362	.02590 (Σ aus	.03039 H–L	–	.03305 A, H–L	.06078	×	–	–	–	–	–
I	3	8	22	6	−.481	.00412 (Σ aus	.00449 I–L	–	.00715 A, I–L	.00898	×	×	×	–	–	–
J	2	9	23	5	−.600	.00036 (Σ aus	.00037 J–L	–	.00037 J–L	.00074	×	×	×	×	×	–
K	1	10	24	4	−.719	.00001 (Σ aus	.00001 K–L	–	.00001 K–L	.00002	×	×	×	×	×	–
L	0	11	25	3	−.838	.00000 (Σ aus	.00000 L	–	.00000 L	.00000	×	×	×	×	×	–

Unter den Überschreitungswahrscheinlichkeiten sind die Nummern der Tafeln angegeben, die zur Summation herangezogen wurden (z.B. J–L: Summation der Überschreitungswahrscheinlichkeiten der Tafeln von J bis L). Die einseitigen Überschreitungswahrscheinlichkeiten wurden nur in Richtung des Vorzeichens von φ berechnet. Abkürzungen:
FH: Bestimmung nach FREEMAN und HALTON
$2P_1$: Bestimmung nach der doppelten einseitigen Überschreitungswahrscheinlichkeit
α/2: Bestimmung nach dem α/2-Verfahren.
Ein »×« in der letzten Spalte bedeutet, daß die Tafel zum vorgegebenen Signifikanzniveau α als signifikant betrachtet werden kann. Nach unserer Rundungsregel hätten die Wahrscheinlichkeiten für $\alpha = 0.050$ dreistellig, für $\alpha = 0.0020$ vierstellig ausgedruckt werden müssen. Um dem Leser die Möglichkeit zu geben, auch noch an anderen Beispielen die Verschiedenheit der Verfahren überprüfen zu können, wurde hier eine fünfstellige Darstellungsform gewählt.

während nach dem α/2-Verfahren überhaupt keine Tafel die Ablehnung von H_0 rechtfertigt, denn die Punktwahrscheinlichkeit der Tafel A ist mit .0027 bereits größer als das geforderte $\alpha/2 = 0.0010 + q = 0.0011$.

Die Diskrepanzen zwischen den verschiedenen Verfahren treten völlig unregelmäßig auf. Dabei ist die Entscheidung nach dem α/2-Prinzip in aller Regel extrem konservativ. Vor allem bei schiefen Nullverteilungen wird das α/2 am »steilen« Ende der Verteilung rasch überschritten. Die Lösung nach FREEMAN-HALTON ist meist am schärfsten, während das $2P_1$-Verfahren zwischen beiden liegt. Wie aus der Vertafelung der exakten Überschreitungswahrscheinlichkeiten durch KRÜGER, LEHMACHER und WALL (1979, Band I) hervorgeht, treten diese Unterschiede so häufig auf, daß eine Diskussion notwendig ist. Die Diskrepanzen werden mit wachsendem N häufiger und treten vor allem im empirisch kritischen Bereich mittlerer Zusammenhänge (φ-Koeffizienten zwischen .3 und .5) auf.

Diskussion

Wie ist nun zu entscheiden? Das FREEMAN-HALTON-Verfahren ist konsistent in dem Sinn, daß es alle Tafeln in den Ablehnbereich einbezieht, deren Anteilsdifferenzen oder φ-Koeffizienten größer sind als die beobachtete. Diese Regel führt zu einem Omnibustest, der unspezifiziert alles Extremere in den Ablehnbereich nimmt. Das bedeutet nicht zwingend, daß der Test dann auch echt zweiseitig ist. So sind in unserem Beispiel der Tabelle 1 die Omnibus- und einseitigen Überschreitungswahrscheinlichkeiten für $a = 0$, $a = 1$ und $a = 2$ gleich. Der Omnibustest ist hier kein zweiseitiger Test mehr.

Ebenso kann das $2P_1$-Verfahren nicht garantieren, daß der zweiseitig angelegte Test unter den Bedingungen der gegebenen Randsummen nicht zum einseitigen Test entartet. Für das Beispiel-α von 0.2 % in der Tabelle 1 kann nach diesem Verfahren Tafel J als signifikant betrachtet werden, obwohl auf der anderen Seite der Verteilung die Punktwahrscheinlichkeit der Tafel A das α/2 übersteigt. Lediglich das α/2-Verfahren definiert den Ablehnbereich klar zwei-seitig. Für den Untersucher kann das einen erheblichen Nachteil bedeuten, wenn selbst extreme Tafeln nicht zur Signifikanz führen.

Der Verschiedenheit der Ergebnisse liegt eine Verschiedenheit der zu prüfenden Nullhypothese zugrunde. Für den Omnibustest lautet H_0, daß kein Unterschied in den Anteilen (bzw. kein von Null unterschiedener φ-Koeffizient) besteht, während H_1 die Ungleichheit postuliert. Demgegenüber kann der zweiseitige Test nach dem $2P_1$-Verfahren begriffen werden als zwei simultane einseitige Tests mit entsprechend adjustiertem α-Risiko. Die Alternativhypothese ist gesplittet in eine Hypothese, die eine positive und in eine zweite Hypothese, die eine negative Anteilsdifferenz (bzw. Kontingenz) postuliert. Entsprechend ist ein signifikantes Ergebnis auch anders zu interpretieren. Beim FREEMAN-HALTON-Verfahren kann zum gegebenen Niveau dann davon ausgegangen werden, daß ein Anteilsunterschied (bzw. keine Nullkontingenz) vorliegt. Beim $2P_1$-Verfahren kann weitergehend von einem positiven oder negativen Unterschied (bzw. Zusammenhang) gesprochen werden. Diese erhöhte Aussagekraft findet ihr numerisches Pendant in der erhöhten Konservativität des Tests.

Praktisch kann dieser Unterschied werden, wenn z.B. durch eine Behandlung ein U-förmiger (bitoner) Trend in der Wirkungsvariablen erwartet werden kann. Nach der Theorie der reaktiven Anspannungssteigerung von DÜKER (1963) führen Beeinträchtigungen der Leistungsfähigkeit zuerst zu einer Leistungsverminderung, dann kompensatorisch zu einer Leistungsverbesserung. Soll eine Dosis eines leistungshemmenden Medikaments auf seine generelle Wirkung untersucht werden (im Unterschied zur Nicht-Medikation), ist das FREEMAN-HALTON-Verfahren indiziert. Ein signifikantes Ergebnis ist dann zu interpretieren als unspezifizierte Medikationswirkung. Soll weitergehend unterschieden werden, ob die Medikation einen leistungsmindernden oder leistungssteigernden Effekt hat, ist die $2P_1$-Lösung angebracht. Die Richtung des Unterschieds ist dann ebenfalls zu interpretieren, da im Grund zwei simultane Tests durchgeführt wurden.

Die Entscheidung für eines der beiden Verfahren ist so nur inhaltlich zu treffen. Wird Wert auf die Richtung des Zusammenhangs gelegt, muß der konservativere $2P_1$-Test herangezogen werden. Lassen sich so von der Interpretationsseite her das FREEMAN-HALTON- und das $2P_1$-Verfahren in ihrem Anwendungsbereich unterscheiden, ist die noch weiter erhöhte Konservativität des Tests beim α/2-Verfahren interpretativ nicht mehr aufzufangen. Für dieses Verfahren spricht lediglich seine logische Stringenz, wenn es auf einer echten Zweiseitigkeit besteht. Praktisch wichtige Anwendungsfälle, bei denen auf eine klare Zweiseitigkeit Wert gelegt werden müßte, sind nicht zu sehen.

Die Schwierigkeiten sind oft zu umgehen, wenn bereits bei der Versuchsplanung darauf geachtet wird, daß die Randverteilungen der Vierfeldertafel nicht extrem asymmetrisch werden. Erreicht wird das durch eine möglichst gleich große Besetzung der Stichproben und ein Kriterium mit mittlerer Auftretenswahrscheinlichkeit in der Gesamtstichprobe. Können diese Desiderate nicht eingehalten werden, treten Unterschiede zwischen FREEMAN-HALTON- und $2P_1$-Verfahren häufig auf. Ein Rekurs auf das Verfahren, das das signifikante Ergebnis erbringt, ist nicht zulässig. Wurde eine zweiseitige Hypothese aufgestellt, überschreitet aber $2P_1$ das gesetzte α, muß die Untersuchung wiederholt werden, wobei die gefundene Richtung des Unterschieds jetzt zur Begründung eines einseitigen Tests herangezogen werden kann.

Literaturverzeichnis

DÜKER, H.: Über reaktive Anspannungssteigerung. Z. exp. ang. Psych., 10, 1963, 46–72

FISHER, R. A.: Statistische Methoden für die Wissenschaft. London: Oliver-Boyd, 1956

FREEMAN, G. H. und HALTON, J. H.: Note on an exact treatment of contingency, goodness of fit, and other problems of significance. Biometrika, 1951, 38, 141–149

KRAUTH, J. und LIENERT, G. A.: Nichtparametrischer Nachweis von Syndromen durch simultane Binomialtests. Biometrische Zeitschrift, 15, 1973, 13–20

KRÜGER, H.-P., LEHMACHER, W. und WALL, K.-D.: Statistische Tafeln für Sozial- und Biowissenschaftler. Band I: Die Vierfeldertafel. Weinheim: Beltz, 1979

OWEN, D. B.: Handbook of statistical tables. Reading, Mass.: Addison-Wesley, 1962

PEARSON, E. S. und HARTLEY, H. O.: Biometrika tables for statisticians. Vol. I. 3rd ed. Cambridge: Univ. Cambridge Press, 1970

YATES, F.: Contingency tables involving small numbers and the χ^2-test. J. Roy. Stat. Soc. (B) 1, 1934, 217–235

Anschrift des Verfassers: PD Dr. Hans-Peter Krüger, Universität Erlangen–Nürnberg–Fachbereich 11–, Regensburger Str. 160, 85 Nürnberg