Reprinted from the German Journal of Psychiatry · http://www.gjpsy.uni-goettingen.de · ISSN 1433-1055

Exact Tests for Small Sample 3×3 Contingency Tables With Embedded Fourfold Tables: Rationale and Application

Matthias J. Müller
Department of Psychiatry, University of Mainz, Germany

Corresponding author: Dr. M. J. Müller, Department of Psychiatry, University of Mainz, Untere Zahlbacher Straße 8, D-55131 Mainz, Germany. E-mail mjm@mail.psychiatrie.klinik.uni-mainz.de

Abstract

In bio-psycho-social research, one is often confronted with categorical outcome variables obtained in small samples; appropriate non-parametric test statistics are required. The rationale for an exact test procedure for small sample 3×3 contingency tables is given. An algorithm is outlined which can easily be translated into a computer program. A preliminary program based on the recommended algorithm was used for application to three different examples of small sample designs. Thereby it is shown that the algorithm is practical for evaluation of 3×3 and 3×2 cross-tables with low expected cell frequencies as well as for analyzing fourfold tables with low cell frequencies embedded in larger cross-tabulations as derived by bipredictive configural frequency analysis (German J Psychiatry 2001;4:57-62)

Keywords: cross-tables, contingency tables, categorical data, configural frequency analysis

Received: 12.09.2001
Published: 21.12.2001

Introduction

In psychiatric research, like in other biopsychosocial areas, one is often confronted with categorical outcome variables (c.f. Müller et al., 1997; Straube et al., 1998) obtained in small samples (Lienert, 1978; Bortz and Lienert, 1998). One the one hand, clinically meaningful responses can often be categorized (e.g., marked improvement vs. slight improvement vs. no improvement/worsening for response variables, or no vs. slight to moderate vs. severe side-effects). One the other hand, studies on specific psychiatric disorders (e.g. patients with acute katatonia, schizophrenia with persisting negative symptoms, dysmorphophobia or specific personality disorders) are often difficult to conduct, and sample sizes are frequently small. Nevertheless, such studies and their appropriate interpretation can be of paramount clinical interest. In such cases, appropriate non-parametric test statistics are required. The use of χ^2-tests in tables when expected and observed numbers are small is often inaccurate, and parametric tests are in most cases not appropriate.

Early in 1951, a rationale was proposed (Freeman and Halton, 1951) to perform an exact test for contingency tables when it is not certain if the expected numbers are too small or when χ^2-tests are completely unsuitable. The exact method applies whenever the population is infinite or sampling has been done with replacement of sampled members, and sampling was following a random process.

Additionally, finite or exact tests and classical χ^2-tests follow a different rationale: exact tests calculate and cumulate point probabilities based on the binomial function and the actual (finite) distribution of values for the observed and at least equally extreme events. The cumulative probability is then compared with a specified cut-off (e.g. $\alpha = 0.05$). On the other hand, classical χ^2-tests calculate the fit of the observed distribution of values with an expected (infinite) distribution under the assumption of independence. If the deviation of observed from expected values exceeds a specified critical value corresponding to a specified probability (e.g., $\alpha = 0.05$) the null hypothesis of independence is rejected.

To evaluate r×c contingency tables with $r \geq 2$ rows and $c \geq 3$ columns, bipredictive configural frequency analysis (CFA) or interaction structure analysis have been repeatedly proposed (Heilmann et al., 1979; Lienert and Netter, 1987; Stemmler, 1994; Müller and Stemmler, 2000). The main objective of bipredictive CFA is to evaluate the concomitant occurrence of specific patterns of independent variables and particular patterns of dependent variables. According to these evaluation strategies, collapsing of cells leads to a fourfold table embedded in a larger 3×3 or 2×3 table for biprediction analysis (Table 1).

Table 1. Fourfold Table Embedded in a 3×3 Cross-Tabulation

	X_1	X_2	Other X_i	Σ
Y_1	a	b	g	A
Y_2	c	d	h	B
Other Y_i	e	f	k	R
Σ	**C**	D	S	N

Thus, low expectancies for the embedded fourfold tables can occur in large interaction tables (Stemmler, 1994) or in tables with small observed cell frequencies. Asymptotic tests (Kimball, 1954; Castellan, 1965) are available, but in designs with small sample sizes exact or finite tests are strongly recommended.

Rationale and algorithm for an exact 3×3 contingency test

The rationale for an exact bipredictive test concerning the embedded fourfold table of Table 1 is rather simple. With fixed marginals, according to Freeman & Halton (1951), all point probabilities p_i of the possible *r*×*c* tables should be computed. The point probabilities p_i with the same or smaller values than the point probability p_0 of the originally observed table are cumulated to obtain the probability p_{cum} which has to be compared with the pre-specified level α for H_0. The value of p_{cum} represents implicitly all possibly occurring tables which are at least as extreme as the observed table under H_0 and thus can be interpreted as a two-tailed transition probability (formula 1).

$$p_{cum} = \sum p_i \quad \text{with } p_i \leq p_0 \qquad (1)$$

Following the binomial distribution, p_i is given by

$$p_i = \frac{A!B!C!D!R!S!}{N!a!b!c!d!e!f!g!h!k!} \qquad (2)$$

A computer program should systematically compute all point probabilities p_i of the possibly occurring tables and select those of them with at least extreme cell distributions compared to p_0. For the computation of p_i of the embedded fourfold tables in a *r*×*c* contingency tables (c.f. Table 1), all possible tables have to be generated. For that purpose, four cells (*a*, *b*, *c*, and *d*) are systematically varied under pre-specified restrictions. One possible algorithm which can easily be implemented in an computer program consists of the following four nested loops (c.f. notations of Table 1):

```
$a$ -> $Min (A, C)$;                                          (3)
    for all $a$:      $b$ -> $Min (A, D)$;
        for all $b$:      $c$ -> $Min (B, C)$;
            for all $c$:      $d$ -> $Min (B, D)$;
```

- compute p_0 for the observed combination (a_0, b_0, c_0, d_0) with marginals A_0, B_0, C_0, D_0, R_0, S_0
- compute p_i for each combination (*a*, *b*, *c*, *d*) with fixed marginals *A*, *B*, *C*, *D*, *R*, *S*
- cumulate p_i for all $p_i \leq p_0$

For each combination of (*a*, *b*, *c*, *d*), the point probability p_i according to formula (2) is computed. Apparently, the denominator in formula (2) is constant for all computations. Thus, only the nominator in formula (2) has to be repeatedly calculated. Instead of computing the faculties *A!*, *B!*, *C!*, *D!*, *R!*, and *S!*, several alternatives can be proposed. It has been suggested to compute χ^2-values for the observed table (χ_o^2) and for each possible fourfold table (χ_i^2) and then to cumulate the point probabilities of all tables with $\chi_i^2 \leq \chi_0^2$ (Krauth, 1973).

G.A. Lienert (personal communication) has proposed to calculate partial association coefficients ϕ_i for the observed and all possible fourfold tables embedded in the 3x3 table of interest. Accordingly,

$\phi_i = (ad - bc) / \sqrt{ABCD(A+B)/N}$ or, more straightforward, the difference of cross-products $CP_i(ad - bc)$ can be computed and only the point probabilities for the tables with $CP_i \leq CP_0$ are then cumulated. In the program used for data in the present paper, the original alternative from formula (2) has been chosen. All mentioned alternatives differ solely in the way of calculation and lead to identical results with regard to p_i and p_{cum}.

Examples and application

In the following examples, the aforementioned algorithm was used for a preliminary program (QuickBasic). Additional restrictions, e.g. $a \geq a_0$, with a_0 = observed cell frequency *a*, can easily be implemented in a program. Such restrictions depend solely on the research hypothesis and are analogous to the issue of global vs. specific testing (e.g. MANOVA as global test, and ANOVA or t–test as specific tests) and one-sided vs. two-sided tests in parametric statistics. The program was applied to two already published examples and one additional unpublished example. All data examples had originally been provided by G.A. Lienert (personal communication) and are used in this paper exclusively to illustrate the testing rationale. Accordingly, transfer and application of the testing principles to psychiatric research should be facilitated.

Example 1

The first example is taken from Lienert (1978) and represents a 3×3 cross-tabulation (df = 4) with very low cell frequencies and a total N = 9 (Table 2a). A new antibiotic was tested in three bacterial species (Staphylococcus, Streptococcus, Pyocyaneus) with regard to its effect on reduction of colony growth (++, strong effect; +, moderate effect; 0, no effect). The null hypotheses was that the antibiotic treatment was equally effective in all tested bacteria species. Thus, in this example, the program was used for global testing of independence of an antibiotic treatment and different pathogenic agents (bacteria).

Table 2a. Example 1: Cross-Tabulation

	Efficacy of Antibiotic Treatment			
Bacteria	**++**	**+**	**0**	Σ
Staphylococcus	3	1	0	4
Streptococcus	2	0	0	2
Pyocyaneus	0	0	3	3
Σ	5	1	3	9

The low cell frequencies with expected frequencies between 0.22 and 2.22 make asymptotic χ^2-testing inaccurate. When a conventional χ^2-value is nevertheless computed, χ^2_4 = 9.9 is obtained (P = 0.042). In a Monte Carlo simulation, it has been shown (Craddock, 1966) that for values of *n* between 10 and 25, the χ^2_4 statistic is becoming inaccurate. Derived from 10,000 random samples with N = 10, critical χ^2_4 values of 8.69 and 11.47 have been estimated for the 95th and 99th percentile of the P(χ^2) distribution, respectively, (Craddock, 1966), compared to χ^2_4 values of 9.49 (α = 0.05) and 13.28 (α = 0.01) when asymptotically $n \rightarrow \infty$. An exact permutation test as outlined above revealed the following results (Table 2b).

The obtained results show that the exact value of P = 0.012 is below an arbitrarily chosen level of statistical significance, α = 0.05, and corresponds well with the results of Craddock (1966).

Table 2b. Example 1: Results of Exact Testing Procedure

	Freeman-Halton Test (Finite Test, 2-Tailed Probabilities)			
Summary:	T1	T2	other T	Sum
pattern 1	a=3	b=1	g=0	A=4
pattern 2	c=2	d=0	h=0	B=2
other pattern	e=0	f=0	k=3	R=3
Sum	C=5	D=1	S=3	N=9
p (tot)=1.0000000000	[not all tables are shown]			
p (cum)=**0.0119047621**				
p (obs)=0.0079365079	[Table No.2]			
Detailed results (only tables with p <= of the observed table):				
Results	a=0, b=1, c=2, d=0, e=3, f=0, g=3, h=0, k=0:			
Table No. 1	p=0.0079365079			
Results	a=3, b=1, c=2, d=0, e=0, f=0, g=0, h=0, k=3:			
Table No. 2	p=0.0079365079			
Results	a=4, b=0, c=1, d=1, e=0, f=0, g=0, h=0, k=3:			
Table No. 3	p=0.0039682540			

Example 2

The second example was chosen from Freeman and Halton (1951) and is also published in Bortz and Lienert (1998). The contingency of a diagnostic group membership (obsessive-compulsive disorder or generalized anxiety disorder) and test outcome (clinical outcome after 8 weeks behavioral therapy) was evaluated in a 3×2 cross-tabulation (df = 2) (Table 3a) using an exact test yielding the following results.

Table 3a. Example 2: Cross-Tabulation

	Clinical Improvement After Behavioral Therapy			
Diagnostic Group	Marked improvement	Moderate improvement	Slight or no improvement	Σ
OCD	0	3	2	5
GAD	6	5	1	12
Σ	6	8	3	17

OCD, obsessive-compulsive disorder; GAD, generalized anxiety disorder

Table 3b. Example 1: Results of Exact Testing Procedure

	Freeman-Halton Test (Finite Test, 2-Tailed Probabilities)			
Summary:	T1	T2	other T	Sum
pattern 1	a=0	b=3	g=2	A=5
pattern 2	c=6	d=5	h=1	B=12
other pattern	e=0	f=0	k=0	R=0
Sum	C=6	D=8	S=3	N=17

p (tot)=1.0000000379	[not all tables are shown]
p (cum)=**0.0882352963**	
p (obs)=0.0271493223	[Table No.2]

Detailed results (Tables with p <= of the observed table):

Results	a=0, b=2, c=6, d=6, e=0, f=0, g=3, h=0, k=0:
Table No. 1	p=0.0045248870
Results	a=0, b=3, c=6, d=5, e=0, f=0, g=2, h=1, k=0:
Table No. 2	p=0.0271493223
Results	a=0, b=5, c=6, d=3, e=0, f=0, g=0, h=3, k=0:
Table No. 3	p=0.0090497741
Results	a=1, b=1, c=5, d=7, e=0, f=0, g=3, h=0, k=0:
Table No. 4	p=0.0077569492
Results	a=2, b=0, c=4, d=8, e=0, f=0, g=3, h=0, k=0:
Table No. 5	p=0.0024240466
Results	a=3, b=0, c=3, d=8, e=0, f=0, g=2, h=1, k=0:
Table No. 6	p=0.0096961865
Results	a=4, b=0, c=2, d=8, e=0, f=0, g=1, h=2, k=0:
Table No. 7	p=0.0072721399
Results	a=4, b=1, c=2, d=7, e=0, f=0, g=0, h=3, k=0:
Table No. 8	p=0.0193923731
Results	a=5, b=0, c=1, d=8, e=0, f=0, g=0, h=3, k=0:
Table No. 9	p=0.0009696187

Example 3

The third example is an application of the exact test procedure for biprediction analysis of an fourfold table embedded in a larger 3×3 cross-tabulation. The relationship of extreme personality traits (extraversion [E], and neuroticism [N]) and results (reaction time [RT], and detection errors [DE]) in a vigilance task was studied in 24 subjects. The complete original design comprised two discrete independent variables (high or low personality traits according to norm values, E+/- and N+/-) and two dichotomized outcome variables (high or low values according to the group median, RT+/- and ED+/-). The specific ("local") hypothesis was that subjects with E-N+ ("dysthymics" in the terminology of Eysenck, 1957) act faster (RT-) compared to subjects with E+N+ ("hysterics"), whereas both groups (E-N+ and E+N+) have a relatively high rate of detection errors (DE+). Thus, the 2×2 interaction of E-N+ vs. E+N+ and RT+DE+ vs. RT-DE+ was specifically tested using an embedded fourfold table (Table 4).

Table 4a. Example 3: Cross-Tabulation; Fourfold Table Embedded in A 3×3 Table

Personality trait pattern	RT+ DE+	RT- DE+	Other response patterns	Σ
E- N+	$a_0 = 0$	$b_0 = 3$	$g_0 = 3$	$A = 6$
E+N+	$c_0 = 3$	$d_0 = 0$	$h_0 = 3$	$B = 6$
Other trait patterns	$e_0 = 4$	$f_0 = 2$	$k_0 = 6$	$R = 12$
Σ	$C = 7$	$D = 5$	$S = 5$	$N = 24$

RT, reaction time; DE, detection errors; E, extraversion; N, neuroticism; -, low; + high, dichotomized response variables RT and DE; dichotomous trait variables E and N

For computing the point probability of all tables with at least as extreme cell frequencies as the observed table under H_0, the following restrictions have been imposed:

Fixed marginals A, B, C, D, E, R, S

$a_i \leq a_0 = 0$; $b_i \geq b_0 = 3$; $c_i \geq c_0 = 3$; $d_i \leq d_0 = 0$

$e_i \leq e_0$; $f_i \leq f_0$; $g_i \leq g_0$; $h_i \leq h_0$

Hence, only the parameter k could freely vary within its given upper and lower limits. The results of the exact test are given in Table 4b, the cumulated probability for all tables under the restrictions mentioned above is $P = 0.0047 < 0.05$.

Discussion

The use of exact tests in contingency tables with categorical data is appropriate when expected and observed numbers are small. Based on a long-known rationale (Freeman and Halton, 1951), an algorithm is presented which can easily be implemented in any available program language. However, the necessity of high precision of computation should be taken into account. The nowadays available high computer performance with very fast processors should make the calculation of all interesting point probabilities feasible. The computer program should be adjusted according to the research requirements and the output has to be strictly interpreted with respect to specified hypotheses. Accordingly, several questions have to be answered a priori. The researcher has to determine the restrictions, e.g. the most extreme marginals, which should be considered. Furthermore, the term "at least extreme as observed" has to be strictly defined. On the one hand, it could mean that all possible combinations of assessed features with a point probability lower than the expected one are considered "more extreme". On the other hand, "extremity" could be defined in relation to the observed cell frequency focussing on one or more particular cells.

Table 4b. Example 3: Results of Exact Testing Procedure

	FREEMAN-HALTON-Test (Finite Test, 2-tailed probabilities)			
Summary:	T1	T2	other T	Sum
pattern 1	a=0	b=3	g=3	A=6
pattern 2	c=3	d=0	h=3	B=6
other pattern	e=4	f=2	k=6	R=12
Sum	C=7	D=5	S=12	N=24

p (tot)=1.0000000031	[not all tables are shown]
p (cum)=**0.0047000833**	
p (obs)=0.0025886081	[Table No.1]
Detailed results (Tables with p <= of the observed table):	
Results	a=0, b=3, c=3, d=0, e=4, f=2, g=3, h=3, k=6:
Table No. 1	p=0.0025886081
Results	a=0, b=3, c=4, d=0, e=3, f=2, g=3, h=2, k=7:
Table No. 2	p=0.0011094035
Results	a=0, b=3, c=5, d=0, e=2, f=2, g=3, h=1, k=8:
Table No. 3	p=0.0001664105
Results	a=0, b=3, c=6, d=0, e=1, f=2, g=3, h=0, k=9:
Table No. 4	p=0.0000061634
Results	a=0, b=4, c=3, d=0, e=4, f=1, g=2, h=3, k=7:
Table No. 5	p=0.0005547017
Results	a=0, b=4, c=4, d=0, e=3, f=1, g=2, h=2, k=8:
Table No. 6	p=0.0002080132
Results	a=0, b=4, c=5, d=0, e=2, f=1, g=2, h=1, k=9:
Table No. 7	p=0.0000277351
Results	a=0, b=4, c=6, d=0, e=1, f=1, g=2, h=0, k=10:
Table No. 8	p=0.0000092450
Results	a=0, b=5, c=3, d=0, e=4, f=0, g=1, h=3, k=8:
Table No. 9	p=0.0000277351
Results	a=0, b=5, c=4, d=0, e=3, f=0, g=1, h=2, k=9:
Table No. 10	p=0.0000009245
Results	a=0, b=5, c=5, d=0, e=2, f=0, g=1, h=1, k=10:
Table No. 11	p=0.0000011094
Results	a=0, b=5, c=6, d=0, e=1, f=0, g=1, h=0, k=11:
Table No. 12	p=0.0000000336

In this context, the question has to be answered whether a one-sided or a two-sided research question should be investigated. One-sided tests are appropriate in such cases when there is an explicit and directed a priori hypothesis, e.g. in Example 3 of the present paper, when the a priori hypothesis assumed a certain pattern of response associated with a specified personality trait pattern. A two-sided hypotheses would have left open the direction of the assumed association, i.e. an inverse pattern would also have been accepted.

Additionally, an adjustment for multiple testing (e.g. the Bonferroni procedure) has often to be discussed because the local or regional testing of subtables embedded in larger tables (e.g. Example 3) raises the question, how many of such subtables could be chosen. Again, it strongly depends on the research question and the availability of a priori specified hypotheses whether or not an adjustment for multiple testing has to be taken into account. The present approach can only serve as an algorithm which could be implemented in the appropriate research context.

The present approach is recommended in situations when an asymptotic test does not seem to be appropriate and no commercially available exact test procedure is applicable. Particularly, when bipredictive testing of fourfold tables with low expected frequencies embedded in larger cross-tabulations is required, there has been no exact test yet.

A standard statistical software module or a SAS macro does not exist so far for the outlined finite tests. A Quick-Basic program code used for the present analyses can be obtained from the author on request (email: mjm@mail.psychiatrie.klinik.uni-mainz.de); however, by using the present algorithm programming should be quite easy and, additionally, can provide further insight for interested researchers when data analysis follows explicitly stated a priori research hypotheses.

Acknowledgement

In memoriam Professor em. Dr. Dr. Dr. h.c. mult. G.A. Lienert

References

Bortz J, Lienert GA. Kurzgefaßte Statistik für die klinische Forschung. Berlin: Springer, 1998:83ff.

Castellan NJ. On the partitioning of contingency tables. Psychological Bulletin 1965;64:330–338.

Craddock JM. Testing the significance of a 3×3 contingency table. The Statistician 1966;16:87–94.

Eysenck HJ. The dynamics of anxiety and hysteria. London: Routledge and Kegan Paul.

Freeman GH, Halton JH. Note on an exact treatment of contingency goodness-of-fit and other problems of significance. Biometrika 1951;38:141–149.

Heilmann T, Lienert GA, Maly V. Prediction models in configural frequency analysis. Biometrical Journal 1979;21:79-86.

Kimball AW. Short-cut formulae for the exact partition of χ^2 in contingency tables. Biometrics 1954;20:452–458.

Krauth J. Nichtparametrische Ansätze zur Auswertung von Verlaufskurven. Biometrische Zeitschrift 1973; 15:557–566.

Lienert GA, Netter P. Nonparametric analysis of treatment-response tables by bipredictive configural frequency analysis. Methods of Information in Medicine 1987;26:89–92.

Lienert GA. Verteilungsfreie Methoden in der Biostatistik (2. Aufl., Band II). Meisenheim am Glan: Verlag Anton Hain, 1978:409ff.

Müller MJ, Netter P, von Eye A. Catecholamine response curves of male hypertensives identified by Lehmacher's two sample Configural Frequency Analysis. Biometr J, 1997;39: 29 – 38.

Müller MJ, Stemmler M. Bipredictive Configural Frequency Analysis of small sample MANOVA designs. Psychologische Beiträge, 2000;42;327 – 336.

Stemmler M. A nonparametrical evaluation of ANOVA and MANOVA designs using interaction structure analysis. Biometrical Journal 1994;36:911–925

Straube ER, von Eye A, Müller MJ. The symmetry of symptom patterns in pre-post treatment designs. Pharmacopsychiatry 1998;31:83-88