



Universiteit
Leiden
The Netherlands

The tale of Cochran's Rule: my contingency table has so many expected values smaller than five. What am I to do?

Kroonenberg, P.M.; Verbeek, A.

Citation

Kroonenberg, P. M., & Verbeek, A. (2018). The tale of Cochran's Rule: my contingency table has so many expected values smaller than five. What am I to do? *The American Statistician*, 72(2), 175-183. doi:10.1080/00031305.2017.1286260

Version: Not Applicable (or Unknown)

License: [Leiden University Non-exclusive license](#)

Downloaded from: <https://hdl.handle.net/1887/70339>

Note: To cite this publication please use the final published version (if applicable).



The Tale of Cochran's Rule: My Contingency Table has so Many Expected Values Smaller than 5, What Am I to Do?

P. M. Kroonenberg & Albert Verbeek

To cite this article: P. M. Kroonenberg & Albert Verbeek (2018) The Tale of Cochran's Rule: My Contingency Table has so Many Expected Values Smaller than 5, What Am I to Do?, The American Statistician, 72:2, 175-183, DOI: [10.1080/00031305.2017.1286260](https://doi.org/10.1080/00031305.2017.1286260)

To link to this article: <https://doi.org/10.1080/00031305.2017.1286260>



© 2018 The Author(s). Published with license by Taylor and Francis© Pieter M. Kroonenberg and Albert Verbeek



Accepted author version posted online: 28 Feb 2017.
Published online: 15 Mar 2018.



Submit your article to this journal [↗](#)



Article views: 2653



View Crossmark data [↗](#)

The Tale of Cochran's Rule: My Contingency Table has so Many Expected Values Smaller than 5, What Am I to Do?

P. M. Kroonenberg* and Albert Verbeek

Department of Education and Child Studies, Leiden University, Leiden, The Netherlands

ABSTRACT

In an informal way, some dilemmas in connection with hypothesis testing in contingency tables are discussed. The body of the article concerns the numerical evaluation of Cochran's Rule about the minimum expected value in $r \times c$ contingency tables with fixed margins when testing independence with Pearson's χ^2 statistic using the χ^2 distribution.

ARTICLE HISTORY

Received August 2016
Revised January 2017

KEYWORDS

Cochran table; Exact test; Pearson χ^2 test; $r \times c$ contingency table; χ^2 distribution; 2×2 contingency table

1. Introduction

This article uses an example to challenge statisticians, in an informal way, to reflect on their belief in statistics and to ask themselves whether they practice what they preach. This example serves as an introduction to a numerical investigation into the quality of Cochran's well-known rule of thumb about the minimum expected value needed for using the χ^2 distribution as an adequate approximation to that of Pearson's X^2 statistic when testing independence in a contingency table. The article will conclude with some advice on what to do if a contingency table has many expected values smaller than 5.

2. A Dilemma for the Well-Meaning Statistician

Imagine you are at the doctor's and she is breaking some bad news to you, telling you that you suffer from a rare, serious illness. She recommends surgery which will probably cure the disease. However, there are two types of surgery (A and B), and as a good statistician you naturally wonder whether there is any information available that will help you come to a largely rational decision.

"Well," says the doctor, "We don't have much experience of this illness yet and there have only been 12 previous cases. All we know about the results of the treatment is here in Table 1." The doctor adds that, on the basis of the data, her advice is for you to undergo surgery, in particular Method A. As a meticulous statistician, you wonder whether Method A really is the better of the two.

The answer to this question comes down to determining which treatment has the highest chance of success. The implicit

assumption in looking exclusively at this chance is that no other argument will influence your decision other than whether you will survive the surgery (or lack of it). In that case, there is not much point for someone suffering from the illness in selecting a significance level or carrying out a test for independence of *Type of surgery* and *Patient's state*.



After all, it is no great loss if the patient makes a Type I error and wrongly concludes that *Type of surgery* and *Patient's state* are interdependent. In such a case, the surgery will be futile, but aside from the cost this will not influence the patient's survival chances (the null hypothesis is true after all). If the patient still wants to carry out the test of independence, wrongly concluding that there is independence would be a more serious error (a Type II error). The patient would then come to the conclusion that the choice of treatment method is immaterial, whereas it is not. [From the review report to the earlier Dutch version of this paper.]

If you really do not like doctors cutting into your body, there is a different kind of problem. In that case you would only want to undergo surgery if it offered clear advantages over doing nothing, and it then becomes worth finding out if there is a relationship between *Type of surgery* and *Patient's state*, because in case of independence you will choose not to have any surgery.

At least two problems with testing present themselves. First of all, what was the sampling design in this study, and secondly, once we know this, how do we test the null hypothesis of independence between the *Type of surgery* and the *Patient's state*?

2.1. Sampling Design

With respect to the sampling design there are two serious possibilities. The first is that the doctors randomly allocated the

CONTACT P. M. Kroonenberg  kroonenb@fsw.leidenuniv.nl  Department of Education and Child Studies, Leiden University, Wassenaarseweg 52, 2333 AK Leiden, The Netherlands.

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/r/TAS.

*Pieter M. Kroonenberg is an emeritus professor of multivariate analysis in particular of three-mode data. Albert Verbeek was a full professor at the Institute of Sociology and the Institute of Mathematics, both of the University of Utrecht, The Netherlands. He also worked for the Netherlands Central Bureau of Statistics (CBS). He died on 9 September 1990. He was the initiator of the research reported here as well as co-author of the earlier Dutch version of this paper (Kroonenberg & Verbeek, 1993).

© Pieter M. Kroonenberg and Albert Verbeek. Published with license by Taylor and Francis.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Table 1. Results of the treatment of patients with the illness.

Patient's state	Type of surgery			Total
	None	A	B	
Still alive	0	5	1	6
Passed away	3	1	2	6
Total	3	6	3	12

patients to one of the three surgery categories and the uneven distribution across these categories is accidental. In this case, we are dealing with a multinomial experimental design with $N = 12$, so that a multinomial test is appropriate. In the second option, the doctors distributed the patients systematically across categories and we are dealing with a product-multinomial design with the product of three multinomial (here: binomial) distributions with $n_1 = 3$, $n_2 = 6$, and $n_3 = 3$.

There is an ongoing debate about how one should test in these sorts of situations, that is, whether a multinomial distribution (fixed N), a product-multinomial distribution (one fixed marginal), or a (generalized) hypergeometric distribution (two fixed marginals) should be used. For a discussion and references about this debate, see Agresti (1992, pp. 146–148) or Verbeek and Kroonenberg (1993, Appendix B.2). From our discussion, it becomes clear that we prefer to work with two fixed marginals, which is what we will do here. However, there are others who are not keen on fixed marginals and for them this article will have little to offer.

2.2. Correct Test

Given fixed margins, the distribution for the observations is a hypergeometric distribution, and a different hypergeometric for each set of fixed margins, which makes it virtually impossible to tabulate the distribution in any sensible way. Not only is the distribution itself discrete, but obviously any statistic defined on the contingency table has a discrete distribution as well. Given the margins, only a limited number of values are possible for any particular statistic S . The number of possible values for the statistic S depends directly on the marginal totals. The standard solution when we test independence is to use asymptotic arguments, and approximate the discrete distribution with a continuous one which can be tabulated, in particular the χ^2 distribution with $(r - 1) \times (c - 1)$ degrees of freedom. Of course, the quality of the approximation of the continuous distribution to the discrete distribution of a particular statistic determines its usefulness, and will be different for each statistic. The literature on evaluating this quality is extensive, if not daunting. The best references are probably Cochran (1952, 1954), Yarnold (1970), Larntz (1978), Fienberg (1979), Koehler and Larntz (1980), Yates (1984); Koehler (1986); several of these authors also treat more complicated situations, such as loglinear models. A more recent thorough review and discussion of the literature can be found in Agresti (2001).

2.3. Tests for Independence

The standard test for independence against an unspecified alternative is Pearson's (1900) X^2 test [$= \sum(\text{observed} - \text{expected})^2/\text{expected}$] which asymptotically approaches a χ^2

distribution. This test is usually, and in our opinion incorrectly, named the " χ^2 -test." Incorrectly, because X^2 is the test statistic and there are many tests that have an asymptotic χ^2 distribution (for example the log-likelihood ratio test, Kruskal–Wallis' test, etc.). Note that it is common practice to use the χ^2 distribution in all three sampling designs mentioned above. Most people are not too worried about the sampling situation, because asymptotically the different sampling situations lead to the same test. In other words, when X^2 is used as test statistic in conjunction with the χ^2 distribution, regardless of the sampling situation, this will always lead to the same result.

2.4. Small Expected Values

Let us return to our starting point about what to do now that it has been established that you have the illness. As X^2 is the standard test for independence, it is rather irksome that after performing this test with a standard software package, one is often warned in the output that a number of cells have an expected value smaller than 5, without any indication as to how to evaluate this and what, if anything, should be done about it. At present, the message in for instance IBM SPSS for our Table 1 is: "6 cells (100.0%) have expected count less than 5. The minimum expected count is 1.50." (Output from the Crosstabs procedure, IBM SPSS Statistics, Version 23).

Suddenly, you remember something from your statistics textbook along the lines of: "The X^2 test should only be used if the expected values are larger than 5" or something to that effect. It is interesting how little is explained about that "should." *Should* in this case means using the test in a way that minimizes the risk of making the wrong decision. It so happens that the p_{χ^2} value [$\Pr_{\chi^2}(X^2 \geq X^2_{\text{obs}})$] of the X^2 test does not necessarily have the same value as the correct exact p_{X^2} value [$\Pr_{X^2}(X^2 \geq X^2_{\text{obs}})$] of the *Exact test*, that is, the p value based on the distribution of X^2 itself. In other words, it is possible that the X^2 test using the χ^2 distribution might lead to the rejection of the null hypothesis, whereas the exact test would come to a different conclusion and vice versa. However, for an exact test the exact distribution is not identical in the three sampling situations mentioned. If one decides to condition on the margins as we have done in this paper, the (generalized) hypergeometric distribution is used in all situations and there is once again one test for all sampling situations. By the way, sometimes the exact test is seen in an incorrect light, as in the remark "Note that Fisher's exact test does not have a 'test statistic,' but computes the p value directly."¹ Thus, the fact is overlooked that X^2 itself is the test statistic being evaluated using its own distribution rather than the χ^2 one.

On the basis of the information in Table 1, should you accept Method A as superior, or could you just as well roll a three-sided die? The X^2 in Table 1 is 6.0, which is larger than the critical χ^2 value ($= 5.991$, if $\alpha = 0.05$ with $df = 2$; $p_{\chi^2} = .0498$), so we can conclude on formal grounds that, given the significance, we can reject the null hypothesis of independence – even with such a small number of observations. If in Table 2 we look at the normalized residuals we see that there are a larger number of survivors of the treatment A (Haberman's (1973) adjusted residuals; see also Verbeek and Kroonenberg,

¹ <http://www.ats.ucla.edu/stat/stata/whatstat/whatstat.htm>

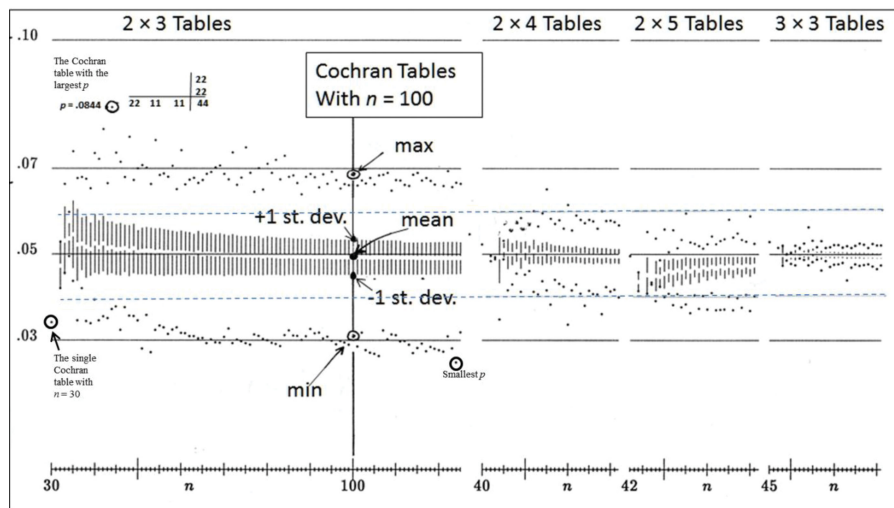


Figure 1. This figure shows the mean, mean ± 1 standard deviation, minimum and maximum for fixed n , r , and c , of $p = p_{\chi^2}(\chi^2 > \chi^2_{df} [0.95])$; vertical axis) for all $r \times c$ Cochran tables for each value of n . The solid horizontal lines indicate the [0.03–0.07] extended Cochran boundaries, and the light dashed horizontal lines the [0.04–0.06] boundaries specified by Cochran.

Table 2. Normalized Haberman Residuals for Table 1.

Patient's condition	Operation method		
	None	A	B
Still alive	–2.0	2.3	–0.7
Passed away	2.0	–2.3	0.7

NOTE: The normalized residuals, r_n , are defined as $r_n = (\text{observed value} - \text{expected value}) / \text{estimated standard deviation}$; asymptotically they are normally distributed, so that they can be compared (roughly, for small numbers) to the critical values of the standard normal distribution.

1990, 1993, section 9.1). So the doctor's choice for surgery, in particular Method A, seems justified.

2.5. Exact Test and Algorithms

But can the p_{χ^2} be trusted in this situation? After all, all the expected values are smaller than 5. The solution is clear: we will have to use the Exact test. The essence of this test is that, given the marginals, we determine all possible tables, calculate both the X^2 and the hypergeometric probability for the table and then add up the probabilities for all the tables that have an X^2 value that is larger than or equal to the observed X^2 value. This problem, simple in concept, has led many to try and design an algorithm to solve it (for an overview, see Verbeek and Kroonenberg, 1985). The trick is, of course, to do this as efficiently as possible, because when one is dealing with large tables and many observations the calculations can easily get out of hand. As far as we know, the champion exact p -value calculator is Mehta and Patel's (1983, 1986a, 1986b, 1991) network algorithm. In 1993, Clarkson et al., stated "The network algorithm of Mehta and Patel [1986] is currently the best general algorithm for computing exact probabilities in $r \times c$ contingency tables with fixed marginals," but in their article they suggested several improvements to Mehta and Patel's algorithm. The FISHER enumeration algorithm (Verbeek and Kroonenberg, 1985, 1990, 1993) does not do too badly either.

Agresti (1992, p. 144) remarked in this respect that "Among the most popular and versatile programs in the past decade have been ones using the network algorithm [...] applied to several

problems in a series of papers by Cyrus Mehta, Nitin Patel and some co-workers." No explicit comparative studies have been published as far as we are aware. However, for specific cases such as stratified 2×2 tables, mostly light improvements to Mehta and Patel's algorithms have been proposed. In a similar vein, Shan notes in his 2016 book (p. 47) that "The existing network-based algorithm is a general approach, and it may not be the best algorithm for a particular problem. The improved algorithm developed by Engels can potentially be used to motivate the research in this area for the problems, such as reliability testing, and homogeneity testing among strata." Some alternatives have been explored by simulations using bootstrap analyses on asymptotic χ^2 and exact X^2 tests; see Lin, Chang and Pal (2015).

2.6. Programs for Exact Tests

Mehta and Patel's algorithm has been implemented in their StatXact–Cytel Software Corporation². It can handle many situations which benefit from exact tests and has a nice user interface. At the moment, many large statistical program suites such as IBM SPSS, SAS, and STATA have exact tests as an option for many statistics, complemented by a Monte Carlo version for too time-consuming exact calculations, in case of large numbers of possible tables with the same marginals as the one observed. Each of these program suites uses the network algorithm of Mehta and Patel; the R-package `fisher.test`³ is also based on this algorithm; another R-package is `aylmer.test`⁴, which is specifically geared to handling tables with structural zeroes and uses enumeration of "boards" of possible tables (West and Hankin 2008). There are several other related R packages which can be found via the R search site.⁵ Our FORTRAN program FISHER was initially designed within a DOS environment and is now being converted to a Windows version; information can be obtained from The Three-Mode Company⁶. Agresti

² <http://www.cytel.com/software/statxact>

³ <https://stat.ethz.ch/R-manual/R-devel/library/stats/html/fisher.test.html>

⁴ <http://finzi.psych.upenn.edu/R/library/aylmer/html/aylmer-package.html>

⁵ <http://finzi.psych.upenn.edu/search.html>

⁶ <http://three-mode.leidenuniv.nl>

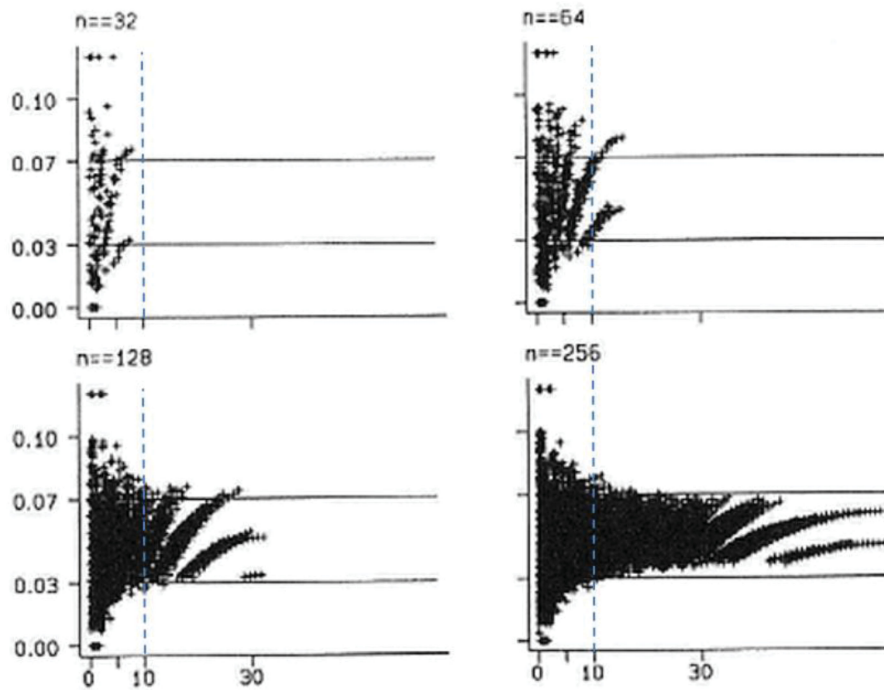


Figure 2. All 2×2 tables for which $n = 32, 64, 128$ or 256 . $\Pr(\chi^2 > \chi^2_{1, [0.95]} (= 3.8461))$ against smallest expected value; p value truncated at 0.10; maximum value was 0.207. The vertical dashed lines indicate the minimum expected value of 10. The data used to create this figure are available as a supplementary file.

(2001) provided a review of the then current state of affairs on exact tests in contingency tables, and many technical issues are treated in the books by Hirji (2006) and Shan (2016).

2.7. Practical Value of Testing

In your precarious state of health you discover that the exact test produces a p_{χ^2} of 0.12, in other words, the test for independence cannot be rejected. What do you do now? Have a go at rolling that three-sided die, or just not have the surgery? Or perhaps become a Bayesian? This last option will probably not make much of a difference, because the necessary a priori information is missing which would mean putting yourself in the hands of the Dirichlet distribution instead of the exact or χ^2 distributions. Good advice is not just hard to come by, it is simply not available. As a reviewer remarked, just one new observation could throw a completely new light on the results, as could one misclassified case (see also Figure 3). But the problem is that you yourself are this new observation. Other statistical problems spring up as well, such as: How was the research that produced Table 1 carried out? Was randomization applied? In addition, there are some more mundane problems, such as: Is information available about the risks of surgery quite aside from the illness, about the competence and experience of your doctor, and about the quality of life with and without surgery etc.?

To make the decision even more difficult, you might consider in any case doing *something*, because letting the illness run its course and waiting until other patients decide to have treatment or not seems a bit too scary. The table about the choice between two types of surgery (supposing that we could test independently of the previous test) has an X^2 of 2.25, with $p_{\chi^2} = 0.46$ and $p_{\chi^2} = 0.13$. In this case, you technically do not know which type of surgery you would choose, although we think that most people would prefer Method A.

We will leave you to wrestle with your problem. We have given all the statistical information we could offer. Quite aside from this specific context, this example confronts us with the important question: “How strong is your belief in hypothesis testing when it really matters?” while at the same time showing how many choices need to be made, and how much depends on the circumstances in which the data have been collected.

3. Cochran’s Rule

Above, we loosely cited a random textbook about the applicability of the X^2 test in the case of small expected values. Aside from a few precursors, it was Cochran (1952, 1954) who gave a precise formulation about the minimum expected values in a contingency table for which the asymptotic χ^2 test would still be accurate enough. In these two papers, Cochran described his famous ‘working rule’ for contingency tables with fixed margins as follows:

Contingency tables with more than 1 d.f. If relatively few expectations are less than 5 (say in 1 cell out of 5 or more, or 2 cells out of 10 or more), a minimum expectation of 1 is allowable in computing X^2 . (Cochran 1954, p. 420).

In the earlier article, Cochran indicated what he meant by ‘allowable’:

A disturbance [i.e. a difference between the exact and tabulated P] is regarded as unimportant if when the P is 0.05 in the χ^2 table, the exact P lies between 0.04 and 0.06, and if when the tabular P is 0.01, the exact P lies between 0.007 and 0.015. (Cochran, 1952, pp. 328, 329).

If we define a *Cochran table* as a contingency table or a pair of margins with $df \geq 2$, all expected values ≥ 1 , and 80% or more of all cells ≥ 5 , then *Cochran’s Rule* can be formulated as

For Cochran tables $0.04 < p_{\chi^2_{0.05}} < 0.06$ and $0.007 < p_{\chi^2_{0.01}} < 0.015$.

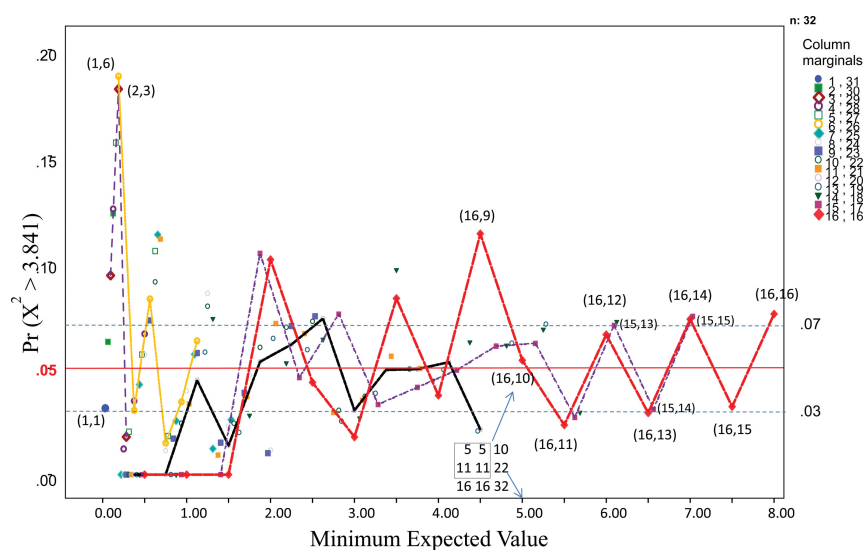


Figure 3. All 2×2 tables for which $n = 32$. $\Pr(\chi^2 > \chi^2_{1, [0.95]} (= 3.8461))$ against minimum expected value. Some families of tables with the same column marginals are connected. The horizontal lines indicate the extended Cochran boundaries of 0.03 and 0.07.

Cochran's Rule is often quoted in a washed-down version, such as "For tables with more than a single degree of freedom, a minimum expected frequency of 5 can be regarded as adequate, although when there is only a single degree of freedom a minimum expected frequency of 10 is much safer" (Hays, 1973, p. 736). On the other hand, other authors, such as Bradley et al. (1979) gave a precise and correct formulation of Cochran's rule. Note how precise Cochran's rule is, and how remarkable this precision is in those pre-computer days. Regrettably, in his papers Cochran does not give any indication of the basis for his rule and its precise formulation.

It should be pointed out that Cochran's Rule was not meant to apply to 2×2 tables, and that his recommendation only applies in case of conditioning on the margins, even though this is not mentioned in the quote above. Actually, Cochran also made a specific recommendation for 2×2 tables:

The 2×2 table. Use Fisher's exact test (i) if the total N of the table < 20 , (ii) if $20 < N < 40$ and the smallest expectation is less than 5. [...]. If $N > 40$ use X^2 , corrected for continuity. (Cochran 1952, p. 334; Cochran 1954, p. 420).

That this is not always appreciated is evident for instance from a paper by Rayson et al. (2004) who seem to imply that Cochran's Rule also applies to 2×2 tables.

Many authors have looked into the accuracy of Cochran's rule, and tried to improve on it, either by proposing corrections to X^2 or to the asymptotic χ^2 distribution. Note that the first approach is a bit awkward because the statistic itself is modified (turned into a different statistic) to bring its critical region into agreement with that of the asymptotic distribution, rather than vice versa. One such attempt was made by Yates (1934) who proposed a correction especially for the case of 2×2 tables. The performances of this and other corrections have always been either unsuccessful or uncertain.

4. Quality of Cochran's Rule

It is remarkable, but not surprising, that little research has gone into an evaluation of how sharp the Cochran boundaries are and whether they could be improved. The difficulty with such

an evaluation is that for all possible margins given n , one needs to establish whether a table is a Cochran table and subsequently establish the exact p value for that table by generating the exact distribution of the test statistic. In the literature most effort has gone into comparing asymptotic p values with exact values, and adjusting the asymptotic p values so that they approximate the exact value more closely; see the paper by Agresti (2001) and the books by Hirji (2006) and Shan (2006) for overviews and detailed technical treatments.

In order to find the exact distribution of a test statistic such as X^2 given a set of margins, it is necessary to enumerate the *iso-marginal family*, that is, all possible contingency tables given a specific pair of margins. In 1934, Fisher (Section 21.02) was the first to describe such an exact test for 2×2 tables, and he must have realized that exact testing is also possible in tables with more degrees of freedom. Yates (1934) was probably the first to publish an application of this.

4.1. Quality of Cochran's Rule for Tables with $df \geq 2$

The main purpose in this part of our paper is to give an indication of how well Cochran's rule of thumb works in four sets of Cochran Tables. To give an overall impression of this kind of table: 2×3 , 2×4 , and 2×5 Cochran tables never have more than one expected value smaller than 5. To be a Cochran table a 2×3 table must have at least 30 observations. In fact, for $n = 30$ there is only one; all cells are equal to 5 and it is indicated in Figure 1. For a 2×3 table one needs at least 66 observations to get a Cochran table with an expected value of 1.

Figure 1 gives an overview of the quality of Cochran's rule of thumb for all 2×3 Cochran tables with an n smaller than or equal to 125, and all 2×4 , 2×5 , and 3×3 Cochran tables with $n \leq 72$. In total, this adds up to 1.3 million tables. The vertical axis in the figure represents the probability in the exact distribution of X^2 being larger than or equal to the χ^2 value associated with $p = 0.05$. Thus, the vertical deviations in the plot indicate the size and direction of the differences between the exact and the asymptotic p values. Rather than showing the exact p value for each Cochran table, we have marked by dots the maximum and minimum exact p value given n , the averaged exact p value,

and the intervals of -1 to $+1$ standard deviation: see the example of all 2×3 Cochran tables for which $n = 100$.

In Figure 1, a wider interval $[0.03-0.07]$ is shown, as well as the $[0.04-0.06]$ proposed by Cochran, in order to properly evaluate the exact p values. For 2×3 tables, we can see that the wider boundaries are the “correct” ones, as there are only 73 violations in around 1 million tables when we use the wider interval, where a table is a violation if its exact p value lies outside the designated boundaries. The figure shows that the p values for most Cochran tables fall between 0.03 and 0.07. The table with the largest violation, that is, whose p -value even falls outside the 0.07 boundary, can be seen in the top left-hand corner; it has an n of 44. Note that all its marginal frequencies are proportional and multiples of 11 (see further remarks below). The smallest p -value falls below the 0.03 boundary and is found for a 2×3 table with $n = 124$. Problems with tables with p values < 0.03 only start at $n = 100$.

For the 2×4 and 2×5 tables we see that Cochran boundaries are the “correct” ones and for the 3×3 tables the boundaries are even tighter. It is interesting to note that for each size of table n bears very little relation to the size of the boundaries, the standard deviation, and only slightly more to the average. All in all, the quality of Cochran’s Rule (except for 2×3 tables) is impressive, especially considering the fact that Cochran had virtually no equipment to help him with the calculations. We could, however, say that 2×3 Cochran tables do not behave according to Cochran’s rule and, unless we are prepared to use the wider boundaries, the asymptotic results cannot be trusted in this case and only exact tests are acceptable.

A final remark about the tables that cause the violations. It turns out that especially tables that have equal and/or proportional marginals, as our example in the first section of this article, lead to many larger exact p -values than the 0.05 using the χ^2 distribution. The primary reason for this seems to be that in the case of equal or proportional marginals there are many more tables with equal X^2 values, in other words, the exact distribution of X^2 becomes more discrete so that there are larger jumps in its probability distribution (see also Figure 3 and the Appendix). Therefore, in such situations it is very important whether the observed X^2 value is larger or smaller than the critical value for the chosen α . For 2×3 tables, the asymptotic critical value is $\chi^2_{\alpha=0.05} = 5.991$, a little under a whole number. With proportional marginals such whole numbers occur quite frequently; in our opening example $X^2_{\text{obs}} = 6.0$ and the margins are $(6,3,3)$ and $(6,6)$. This makes the situation unstable around the critical value. Having equal marginals is obviously not dependent on n . Agresti (1992, p. 132) already remarked that “the sample size n often has less relevance than the discreteness of the sampling distribution.” So, even when n is large there can be tables with large discrepancies between the exact and asymptotic critical values. This finding suggests that in studies with discrete response variables it may be better to have a slightly unbalanced design with an unequal number of people in the experimental and the control group.

4.2. Quality of “Cochran’s Rule” for Tables with $df = 1$

In Figure 2, the vertical axis is again represents the probability in the exact distribution that X^2 larger than the χ^2 value for $p = 0.05$ (here: 3.8461). The points in the figure represent the exact

p -value of a table with fixed margins for the n indicated. Due to the single degree of freedom, switching the row and column margins produces the same table with the same distribution; therefore, in order to avoid counting the same table several times, we have set the first column marginal c_1 at $\leq n/2$ and the first row margin, $r_1 \leq c_1$.

In the upper left-hand corner we see that many of the tables for which $n = 32$ have exact p values, not only way beyond the $[0.04-0.06]$ interval, but beyond the $[0.03-0.07]$ interval as well. The points at the far right are those with the highest minimum expected values. For the n shown they occur if all marginals are equal to $n/2$. These highest minimum expected values are 8 for $n = 32$, 16 for $n = 64$, 32 for $n = 128$, and 64 for $n = 256$, respectively. Note that given a fixed row margin, say r_i , the expected values $e_{j|i}$ are exactly linear with the column margin, as $e_{j|i} = c_j \times [r_i/n]$ for $c_j = 1, \dots, n/2$, and vice versa.

If we keep to the wider boundaries for 2×2 tables, the minimum expected value would have to be something in the order of 10–15 (Figure 2; dashed vertical lines), but referring to the boundaries proposed by Cochran it is only if $n = 256$ and the minimum expected values are larger than 45–50 that the exact p values stay more or less within the Cochran boundaries. On the upside, the calculations for the exact test for 2×2 tables can very easily be done, even on a programmable pocket calculator. Luckily most large statistical packages, such as IBM SPSS, SAS, and STATA now provide the required exact p values for contingency tables in general.

It is instructive to explore the results presented in Figure 2 in more detail and we will use the graph in the upper left-hand corner with $n = 32$ to do this. In Figure 3, we show an enlarged and embellished version of this graph. The tables represented by the points are characterized by their 1st column (c_1) and 1st row margin (r_1), thus $(16,15)$ is a table with column margins $[16,16]$ and row margins $[15,17]$. The legend on the top right indicates the two column marginals. In the figure, all tables with the same column margins have the same marker and some of the sets of such tables have been connected. This reveals clearly how sensitive the p -value is for small misclassifications. In fact, for tables with the same column margin a shift of one in the row margin may change the p -value from above 0.05 to below 0.05 and vice versa, for example, as shown in Figure 3, the p -value of $(16,15) = 0.032$, while that of $(16,14) = 0.073$ and that of $(16,13)$ is 0.029. This zig-zag pattern can be seen all through the graph. A more detailed analysis of the shapes in Figure 2 is not directly central for the main thrust of our article but some further insight is provided in the Appendix.

Campbell (2007) discussed 2×2 tables in more detail and suggested that a modification of the Karl Pearson X^2 statistics proposed by his son Egon Pearson (1947), that is, using $(N - 1)/N * X^2$, is more accurate in case of conditioning on only one of the margins, but this involves changing the test statistic. However, he also states that in the case of two fixed margins “there is no dispute that the Fisher–Irwin test (or Yates’s approximation to it) should be used. This last research design is rarely used and will not be discussed in detail.” (p. 3662). Unfortunately, the discussion for 2×2 tables is somewhat muddled in the sense that some authors discuss and evaluate the X^2 test conditional on the margins, while others condition on only one margin or not at all, i.e., they are using different sampling designs; see, for instance, Bradley et al. (1979, p. 1291), who indicated that in their Monte

Carlo studies to “evaluate the Type I error rate of the chi-square test of independence in $R \times C$ contingency tables” [...] “neither the row nor column marginal frequencies were fixed.” Our results as displayed in Figure 2 only apply to the case of conditional testing on both margins, in line with our discussion of the tables with more degrees of freedom.

5. Conclusion

Besides our example, what can we now say in general about the question in the title of this article? Below we list a number of recommendations, not all of them from the present text, but largely based on research we did in the past. Note that, as indicated above, we look only at conditional tests with fixed margins. Rules of thumb are per definition always wrong and should never be followed blindly. A general point is that programs with exact tests are not too time-consuming. In the cases where they might be, usually a Monte Carlo estimate of the exact p -value is given first, so that on the basis of that outcome the user can decide to do a more time-consuming exact test instead.

First, with respect to 2×2 tables with fixed margins, the smallest expected value must be at least 15 in order to stay within the extended Cochran limits (see Figure 2). However, the time involved in an exact test is so minimal that we might as well carry out the exact test. The Yates correction is unnecessarily conservative (see, for an example, the Appendix).

Second, Figure 1 shows that in 2×3 Cochran tables with fixed margins the deviation from the Cochran limits can be rather large, in fact larger than the Cochran bounds; which means it is advisable to do the exact test for all 2×3 tables.

Third, for Cochran tables with more than two degrees of freedom, the asymptotic critical value is very reliable and it can generally be used with impunity. However, it is necessary to take care when dealing with equal or proportional marginals. Moreover, for non-Cochran tables with more than two degrees of freedom an exact test is advisable.

Finally, given the quality of the χ^2 approach for the distribution of X^2 in the case of two fixed margins, it is rarely necessary to use exact tests when either one or neither margin is fixed (see Verbeek and Kroonenberg, 1993, Appendix B.2).⁷

Appendix: p -Value Patterns in 2×2 Tables

In all panels of Figure 2, we see patterns akin to peacock feathers. This Appendix sheds some light on these patterns. To do this, we present a simplified and shortened version (Figure A1) of the right-hand top graph with $n = 64$.

In Figure A1, we see that the peacock feathers arise from families of tables with the same row margin and with the other margin running from 3 to $n/2$. In the present case, there are four subfamilies. Given a subfamily, as the minimum expected value increases the X^2 decreases so that the exact p value, $[\Pr_{X^2}(X^2 \geq \chi^2_{1, [0.95]})]$ becomes larger. Up to the point that X^2 becomes larger than 3.846, which leads the p -value to drop to a very low value. This is the same zig-zag patterns mentioned in the body of the paper. The last points of 0.000 indicate that the null hypothesis cannot be rejected given these margins. Thus, the exact p -value is only piecewise highly correlated with the smallest minimum expected value and at certain points the p -value changes drastically. It only requires a single point or misclassification to do this.

From the columns “Points not in critical region” we see that the Yates correction is more conservative than the Pearson X^2 test as it always has an equal or larger number of tables not in the critical region.

Acknowledgement

Many thanks go to the two reviewers of the original Dutch article, who pointed out some errors in our reasoning. The cooperative attitude of the TAS editor and her comments are also gratefully acknowledged.

⁷ Available at https://www.researchgate.net/publication/28648596_FISHER_310_Testing_independence_in_rxc_tables_Manual_2nd_revised_edition_software

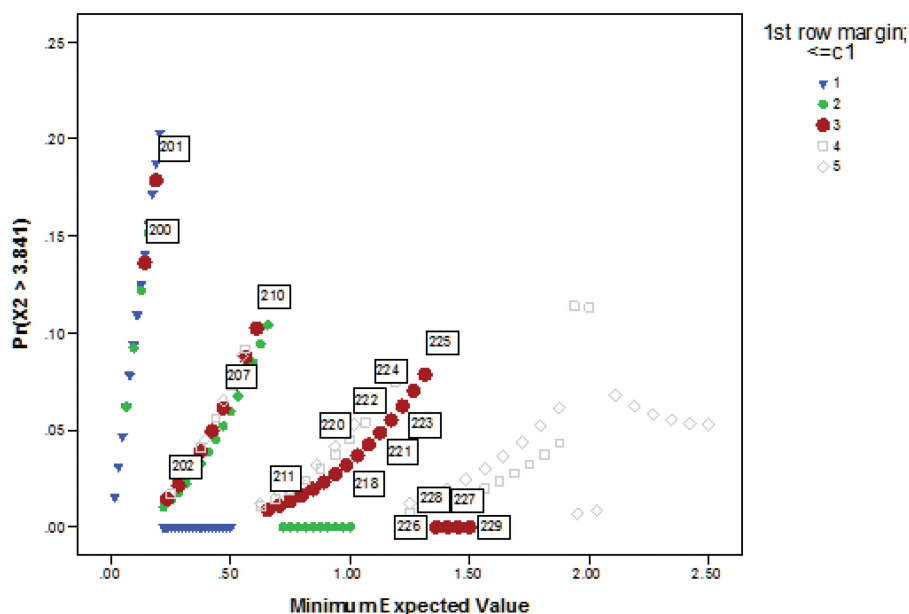


Figure A1. Detail of the top right-hand graph of Figure 2 ($n = 64$). The (red) filled circles are tables with row margins of [3, 61]. The coordinates are given in Table A.1. Boxed numbers correspond to those in Table A.1.

Table A.1. Results for the family of tables with $r_1 = 3$ and $r_2 = 61$ and $3 \leq c_1 \leq 32$.

Plot number	Row margin		Column margin		Smallest $\chi^2 \geq 3.85$	Points not in critical region		Exact p value χ^2	Minimum expected value
	First r_1	Second r_2	First c_1	Second c_2		Statistic = χ^2	Statistic = Yates		
200	3	61	3	61	5.78	1	2	0.136	0.14
201	3	61	4	60	3.94	1	2	0.179	0.19
202	3	61	5	59	15.14	2	2	0.014	0.23
203	3	61	6	58	12.16	2	2	0.021	0.28
204	3	61	7	57	10.04	2	2	0.030	0.33
205	3	61	8	56	8.44	2	2	0.039	0.38
206	3	61	9	55	7.21	2	3	0.050	0.42
207	3	61	10	54	6.22	2	3	0.061	0.47
208	3	61	11	53	5.41	2	3	0.074	0.52
209	3	61	12	52	4.74	2	3	0.088	0.56
210	3	61	13	51	4.17	2	3	0.102	0.61
211	3	61	14	50	11.24	3	3	0.009	0.66
212	3	61	15	49	10.28	3	3	0.011	0.70
213	3	61	16	48	9.44	3	3	0.013	0.75
214	3	61	17	47	8.70	3	3	0.016	0.80
215	3	61	18	46	8.04	3	3	0.020	0.84
216	3	61	19	45	7.45	3	3	0.023	0.89
217	3	61	20	44	6.92	3	3	0.027	0.94
218	3	61	21	43	6.45	3	4	0.032	0.98
219	3	61	22	42	6.01	3	4	0.037	1.03
220	3	61	23	41	5.61	3	4	0.043	1.08
221	3	61	24	40	5.24	3	4	0.049	1.13
222	3	61	25	39	4.91	3	4	0.055	1.17
223	3	61	26	38	4.60	3	4	0.062	1.22
224	3	61	27	37	4.31	3	4	0.070	1.27
225	3	61	28	36	4.05	3	4	0.079	1.31
226	3	61	29	35	3.80#	4	4	0.088#	1.36
227	3	61	30	34	3.57#	4	4	0.097#	1.41
228	3	61	31	33	3.35#	4	4	0.107#	1.45
229	3	61	32	32	3.15#	4	4	0.238#	1.50

NOTE: The plot numbers correspond with those in Figure A1. In the column labeled " $\chi^2 \geq 3.85$ " the first χ^2 larger than 3.85 is given. The # in that column indicated that no $\chi^2 \geq 3.85$ and the closest value smaller than 3.85 is listed. In all cases, the isomarginal family has a size of 4. The column "Points not in critical region" shows how many members of the isomarginal family do not have an $\chi^2 \geq 3.85$, both for χ^2 and for the Yates corrected version. A "4" indicates that the null hypothesis of independence can never be rejected given the marginals.

References

Agresti, A. (1992), "A Survey of Exact Inference for Contingency Tables," *Statistical Science*, 7, 131–177. [176,177,180]

— (2001), "Exact Inference for Categorical Data: Recent Advances and Continuing Controversies," *Statistics in Medicine*, 20, 2709–2722. [176,178,179]

Bradley, D. R., Bradley, T. D., McGrath, S. G., and Cutcomb, S. D. (1979), "Type I Error Rate of the Chi-Square Test in Independence in $R \times C$ Tables that have Small Expected Frequencies," *Psychological Bulletin*, 86, 1290–1297. [179,180]

Campbell, I. (2007), "Chi-Squared and Fisher-Irwin Tests of Two-by-Two Tables with Small Sample Recommendations," *Statistics in Medicine*, 26, 3661–3675. [180]

Clarkson, D. B., Fan, Y., and Joe, H. (1993), "A Remark on Algorithm 643: FEXACT: An Algorithm for Performing Fisher's Exact Test in $r \times c$ Contingency Tables," *ACM Transactions on Mathematical Software*, 19, 484–488. [177]

Cochran, W. G. (1952), "The χ^2 Test of Goodness of Fit," *Annals of Mathematical Statistics*, 23, 315–345. [176,178,179]

— (1954), "Some Methods for Strengthening the Common χ^2 Tests," *Biometrics*, 10, 417–451. [176,178,179]

Feinberg, S. E. (1979), "The Use of Chi-Squared Statistics for Categorical Data Problems," *Journal of the Royal Statistical Society, Series B*, 41, 54–64. [176]

Fisher, R. A. (1934), *Statistical Methods for Research Workers* (originally published 1925, 14th ed. 1970), Edinburgh: Oliver and Boyd. [179]

Haberman, S. J. (1973), "The Analysis of Residuals in Cross-Classified Tables," *Biometrics*, 29, 205–220. [176]

Hays, W. L. (1973), *Statistics for the Social Sciences*, New York: Holt, Rinehart and Winston. [179]

Hirji, K. F. (2006), *Exact Analysis of Discrete Data*. Boca Raton, FL: Chapman & Hall. [178,179]

Koehler, K. J. (1986), "Goodness-of-Fit Tests for Log-Linear Models in Sparse Contingency Tables," *Journal of the American Statistical Association*, 81, 483–493. [176]

Koehler, K. J., and Larntz, K. (1980), "An Empirical Investigation of Goodness-of-Fit Statistics for Sparse Multinomials," *Journal of the American Statistical Association*, 75, 336–344. [176]

Kroonenberg, P. M., and Verbeek, A. (1993), "Mijn Kruistabel heeft zo veel Verwachte Waarden Kleiner dan 5. Wat Moet Ik Nu Doen?," *Tijdschrift voor Onderwijsresearch*, 18, 3–10. [176,181]

Larntz, K. J. (1978), "Small-Sample Comparisons of Exact Levels for Chi-Squared Goodness-of-Fit Statistics," *Journal of the American Statistical Society*, 73, 253–263. [176]

Lin, J.-J., Chang, C.-H., and Pal, N. (2015), "A Revisit to Contingency Table and Tests of Independence: Bootstrap is Preferred to Chi-Square Approximations as well as Fisher's Exact Test," *Journal of Biopharmaceutical Statistics*, 25, 438–458. [177]

Mehta, C. R. (1991), "StatXact: A Statistical Package for Exact Nonparametric Inference." *The American Statistician*, 45, 74–75. [177]

Mehta, C. R., and Patel, N. R. (1983), "A Network Algorithm for Performing Fisher's Exact Test in $r \times c$ Contingency Tables," *Journal American Statistical Association*, 78, 427–434. [177]

— (1986a), "Algorithm 643. FEXACT: A FORTRAN subroutine for Fisher's exact test on unordered $r \times c$ contingency tables," *ACM Transactions on Mathematical Software*, 12, 154–161. [177]

— (1986b), "A hybrid algorithm for Fisher's exact test in unordered $r \times c$ contingency tables," *Communications in Statistics. Series A*, 15, 387–404. [177]

Pearson, K. (1900), "On a Criterion that a Given System of Deviations from the Probable in the Case of a Correlated System of Variables is such that

- it can be Reasonably Supposed to have Arisen from Random Sampling,” in *Philosophical Magazine - Series 5*, 50, 157-175. (Reprinted 1948 in *Karl Pearson's Early Papers*, ed. E. S. Pearson, Cambridge, UK: Cambridge University Press). [176]
- Pearson, E. S. (1947), “The Choice of Statistical Tests illustrated on the Interpretation of Data Classed in a 2×2 Table,” *Biometrika*; 34, 139-167. [180]
- Rayson, P., Berridge, D., and Francis, B. (2004), “Extending the Cochran Rule for the Comparison of Word Frequencies between Corpora,” in *Le Poids des Mots. Actes des 7es Journées Internationales d'Analyse Statistique des Données Textuelles (JADT 2004)*, pp. 926-936, eds. G. Purnelle, C. Fairon and A. Dister, Louvain, Belgium: Presses Universitaires de Louvain. [179]
- Shan, G. (2016), *Exact Statistical Inference for Categorical Data*, London, UK: Academic Press. [177,179]
- Verbeek, A., and Kroonenberg, P. M. (1985), “A Survey of Algorithms for Exact Distributions of Test Statistics in $r \times c$ Contingency Tables with Fixed Margins,” *Computational Statistics and Data Analysis*, 3, 159-185. [177]
- (1990, 1993), *FISHER 3.0: Testing Independence in $r \times c$ Tables*. (2nd revised edition, FISHER 3.10). Leiden: The Three-Mode Company. [177]
- West, L.J., and Hankin, R.K.S. (2008), Exact Tests for Two-Way Contingency Tables with Structural Zeros, *Journal of Statistical Software*, 28, 11, 1-19. [177]
- Yarnold, J. K. (1970), “The Minimum Expectation in X^2 Goodness of Fit Tests and the Accuracy of Approximations for the Null Distribution,” *Journal of the American Statistical Association*, 65, 864-886. [176]
- Yates, F. (1934), “Contingency Tables involving Small Numbers and the χ^2 Test,” *Journal of the Royal Statistical Society, Supplement*, 1, 217-235. [179]
- (1984), “Tests of Significance for 2×2 Contingency Tables (with Discussion),” *Journal of the Royal Statistical Society, Series A*, 147, 426-463. [176]