Research Article

Conditional and Unconditional Tests (and Sample Size) Based on Multiple Comparisons for Stratified 2 × 2 Tables

A. Martín Andrés, 1 I. Herranz Tejedor, 2 and M. Álvarez Hernández 3

1 Bioestadística, Facultad de Medicina, University of Granada, 18071 Granada, Spain
2 Bioestadística, Facultad de Medicina, University Complutense of Madrid, 28040 Madrid, Spain
3 Departamento de Estadística e Investigación Operativa, University of Vigo, 36310 Vigo, Spain

Correspondence should be addressed to A. Martín Andrés; amartina@ugr.es

Received 3 March 2015; Accepted 16 April 2015

Academic Editor: Jerzy Tiuryn

Copyright © 2015 A. Martín Andrés et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The Mantel-Haenszel test is the most frequent asymptotic test used for analyzing stratified 2 × 2 tables. Its exact alternative is the test of Birch, which has recently been reconsidered by Jung. Both tests have a conditional origin: Pearson's chi-squared test and Fisher's exact test, respectively. But both tests have the same disadvantage that the result of global test (the stratified test) may not be compatible with the result of individual tests (the test for each stratum). In this paper, we propose to carry out the global test using a multiple comparisons method (MC method) which does not have this disadvantage. By refining the method (MCB method) an alternative to the Mantel-Haenszel and Birch tests may be obtained. The new MC and MCB methods have the advantage that they may be applied from an unconditional view, a methodology which until now has not been applied to this problem. We also propose some sample size calculation methods.

1. Introduction

In statistics it is very usual to have to verify whether association exists between two dichotomic qualities. This is especially frequent in medicine, for example, where the aim is to assess whether the presence or absence of a risk factor conditions the presence or absence of a disease or compare two treatments whose answers are success or failure, and so forth. In all the cases the problem produces data whose frequencies are presented in a 2 × 2 table: the two levels of one of the qualities are set out in the rows, the two levels of the other quality in the columns, and the observed frequencies are set out inside the table.

The exact and the asymptotic analyses of a 2 × 2 table have their roots in the origins of statistics, and hundred of papers have been devoted to the problem [1]. It is traditional to carry out the exact independence test using the Fisher exact test, which is a conditional test (because it assumes that the marginals of the rows and columns are previously fixed). More than thirty years has passed since the situation changed, and it is well known that the unconditional exact test tends to be less conservative and more powerful than the conditional test [2–4], because the loss of information as a result of conditioning may be as high as 26% [5]. The unconditional tests assume that it is only the values that were really previously fixed: the marginal of the rows, the marginal of the columns or the total data in the table. This causes two types of unconditional test: that of the double binominal model (the first two cases) and that of the multinomial model (the third case). The same can be said of the asymptotic tests, generally based on Pearson's chi-squared statistic with different corrections for continuity (cc). However, the unconditional exact tests have the great disadvantage of being very laborious to compute. An overall view of the problem can be seen in Martín Andrés [1, 6].

Frequently the individuals who take part in the study are stratified in groups based on a covariate such as sex or age, which gives rise to several 2 × 2 tables. In this case the aim is to contrast the independence of both the original dichotomic qualities, bearing in mind the heterogeneity of the populations defined by the strata. To this end, the most frequent approach is to suggest a test under the null
hypothesis of Mantel-Haenszel for which the odds ratio (or the risk ratio) for all the strata is equal to unity. For this purpose the most frequent asymptotic tests are those of Cochran [7] and Mantel and Haenszel [8], both of which are very similar; the exact version of the test is due to Birch [9] (and has recently been reconsidered by [10]). In all these cases the proposed tests are conditional and, when there is only one stratum, the test for the case of only one $2 \times 2$ table is obtained (Fisher’s exact test or Pearson’s chi-squared test). Moreover, Jung [10] and Jung et al. [11] propose a sample size calculation method, asymptotic in the first and exact in the second.

The procedures indicated have the drawback of almost all the tests for a global null hypothesis like the one in question that the result of the global (stratified) test may not be compatible with that of the individual tests (the test for each stratum). In this paper, we propose a global test (MC test) which does not have this disadvantage because it is based on a multiple comparisons method: the global test is significant if and only if at least one of the individual tests is significant. In return the MC test will have the drawback of being less powerful, given that it must control both the alpha error of the global test and the alpha errors in the individual tests. Because of this, another procedure is proposed (MCB test) which only controls the alpha error of the global test (just as in the classic stratified tests), although the alpha error in the individual tests will only exceed the nominal value on a few occasions (and generally by very little). The two procedures are applicable from both the conditional and the unconditional point of view and also when carrying out an asymptotic test or an exact test. The advantage of applying them in the form of an unconditional test is that in this way the loss of power mentioned above is reduced with regard to the classic global tests.

In addition this paper shows that the asymptotic tests function well, even for small samples, if they are carried out with the appropriate continuity correction. And finally, the sample size for almost all the cases studied (exact or asymptotic tests, conditional or unconditional tests) is determined.

### 2. Hypothesis Test

#### 2.1. Notation, Models, and Example

In the following (without loss of generality) it will be assumed that each $2 \times 2$ table refers to the successes or failures in two treatments which are applied to $m_j$ and $n_j$ individuals, respectively. Let $J$ be the number of strata, $N_j = m_j + n_j$ ($j = 1, \ldots, J$) the total of individuals in the stratum $j$, $N = \sum_j N_j$ the total sample size, $(x_j, \overline{x}_j) = (m_j - x_j)$ and $(y_j, \overline{y}_j) = (n_j - y_j)$ the number of successes and the number of failures of the treatments 1 and 2, respectively, and $z_j = x_j + y_j$ and $\overline{z}_j = \overline{x}_j + \overline{y}_j$ the total number of successes and failures in the stratum $j$ respectively. These data may be summarized as shown in Table 1. Once the experiment has been performed, the values obtained will be written with an extra subindex “0,” that is, $x_{j0}, y_{j0}, m_{j0}, n_{j0}, \ldots$

Let $p_j$ and $q_j$ ($\overline{p}_j = 1 - p_j$ and $\overline{q}_j = 1 - q_j$) be the probabilities of success (failure) with treatments 1 and 2 in the stratum $j$, respectively. The odds ratio for each stratum is $\theta_j = \frac{p_j \overline{q}_j}{\overline{p}_j q_j}$, and the aim is to contrast the null hypothesis $H: \theta_j = \cdots = \theta_j = 1$ against an alternative hypothesis with one tail ($K: \theta_j > 1$ for some $j$) or with two tails ($K: \theta_j \neq 1$ for some $j$). This paper addresses only the case of one-sided test; for the two-tail test the procedure is similar.

In the previous description it was assumed that the data $(x_j, y_j)$ of each stratum $j$ proceed from a double binomial distribution of sizes $m_j$ and $n_j$ and probabilities $p_j$ and $q_j$ in groups 1 and 2, respectively. Because in each stratum $j$ there are two previously fixed values ($m_j$ and $n_j$) the model will be referred to as Model 2; the model is very frequently used in practice so that it will serve here as a basis for defining and illustrating the procedures MC and MCB. If in each stratum there is conditioning in the observed value $z_j = x_j + y_j$, then one has Model 3; now the three values $m_j$, $z_j$, and $n_j$ are previously fixed in each stratum $j$ and the only variable $x_j$ arises from a hypergeometric distribution. If only the values of $N_j$ are fixed in each stratum $j$, one will get Model 1: $(x_j, y_j, \overline{x}_j)$ proceeding from a multinomial distribution. Finally, if only the global sample size $N$ is fixed (so that now even the values for $N_j$ are obtained at random), one will have Model 0. With conditioning in the appropriate marginal, the model $X$ leads to the model $(X + 1)$. Therefore, whatever the initial model (i.e., whatever the sampling method for the data obtained), by conditioning in all the nonfixed marginals one always obtains Model 3 (which is the one covered by Birch and Mantel and Haenszel).

Each model produces a different sample space, which is formed by the set of all possible values of the set of variables involved in the same. For example, the sample space of stratum $j$ under Model 2 consists of $(m_j + 1) \times (n_j + 1)$ possible values of $(x_j, y_j)$. Each transition from a Model $X$ to Model $(X + 1)$ constitutes a loss of information, because the number of points of the new sample space is very much smaller than that of the previous one. Probably the most dramatic transition is that of Models 2 to 3, a transition in which the loss of information may reach 26% for $J = 1$ [5]. In addition, each transition implies using a conditional rather than an unconditional method of eliminating nuisance parameters, something which is generally never advisable [13].

The data in Table 2, which are given by Li et al. [12], are taken from preliminary analysis of an experiment of three groups to evaluate whether thymosin (treatment 1), compared to a placebo (treatment 2), has any effect on the treatment of bronchogenic carcinoma patients receiving radiotherapy. The one-sided $p$ values are $P_{\text{Birch}} = 0.1563$ by global conditional stratified exact test and $P_1 = 0.80073$, $P_2 = 0.57143$, and $P_3 = 0.14706$ by Fisher’s individual conditional exact test in each stratum. If the global test is carried out to an error

| Treatment | Response | Total |
|-----------|----------|-------|
| 1         | $x_j$    | $\overline{x}_j$ | $m_j$ |
| 2         | $y_j$    | $\overline{y}_j$ | $n_j$ |
| Total     | $z_j$    | $\overline{z}_j$ | $N_j$ |

Table 1: Frequency data of $2 \times 2$ table for stratum $j$. 


Table 2: Response to thymosin in cancer patients (yes = success, no = failure).

|        | Stratum 1 |        | Stratum 2 |        | Stratum 3 |        |
|--------|-----------|--------|-----------|--------|-----------|--------|
|        | Yes | No | Total | Yes | No | Total | Yes | No | Total |
| Thymosin | 10  | 1 | 11  | 9  | 0 | 9  | 8  | 0 | 8  |
| Placebo | 12 | 1 | 13 | 11 | 1 | 12 | 7  | 3 | 10 |
| Total   | 22 | 2 | 24 | 20 | 1 | 21 | 15 | 3 | 18 |

$\bar{\alpha} = 0.1563$ we conclude $K$, so that now $\theta_j > 1$ at least once. However no individual test has significance if these are carried out to an alpha error that respects the former global error; for example, by using Bonferroni's method, the smaller of the three $p$ values $P_j = 0.14706 > 0.1563/3$. The same thing occurs if asymptotic tests are used. Our aim is to define procedures in which these incompatibilities will not occur.

2.2. Conditional Tests Obtained by Using Classic Methods (Model 3). The $p$ value of exact test is $P_{\text{Birch}} = 0.1563$. Table 3 shows this value and the remaining $p$ values in this paper. This result is based on determining the probability of all the configurations $(x_j | N_j, m_j, z_j), j = 1, 2, \ldots, J$, such as $S = \sum_j x_j \geq S_0 = \sum_j x_{j0} = 27$. Here $S$ is a test statistic determining the order in which the points of the sample space $(x_1, x_2, x_3)$ enter the region $R$, a region whose probability under $H$ yields the value of $P_{\text{Birch}}$. Note that as the sample spaces in each stratum are $9 \leq x_1 \leq 11, 8 \leq x_2 \leq 9$, and $5 \leq x_3 \leq 8$, the possible values of $(x_1, x_2, x_3)$ will be $3 \times 2 \times 4 = 24$, which is the total number of points in the global sample space; of these, four belong to $R$ (three with $S = 27$ and one with $S = 28$), so that $4/24 = 0.1667$. Moreover note that, under the original Model 2, the number of points in the sample space of strata 1,2, and 3 are $(m_j + 1) \times (n_j + 1) = (11 + 1) \times (13 + 1), (9 + 1) \times (12 + 1), (8 + 1) \times (10 + 1)$, respectively. The total points for the global sample space will be $168 \times 130 \times 99$: more than two million, compared to only 24 in Model 3. To determine the value $P_{\text{Birch}}$ have developed various programs (see references in [14]); an easy way to get it is through http://www.openepi.com/Menu/OE_Menu.htm (option “Two by Two Table”).

The asymptotic test of Mantel-Haenszel based on $\sum x_j$ is asymptotically normal with mean $\sum E_j = \sum_i m_i \bar{x}_i / N_j$ and variance $\sum V_j = \sum_i m_i \bar{x}_i \bar{z}_i / N_j^2 (N_j - 1)$. Therefore the contrast statistic is $X_{\text{MH}} = (\sum x_j - \sum E_j) / (\sum V_j)^{0.5}$, whose $p$ value $P_{\text{MH}} = 0.0760$ patently does not agree with $P_{\text{Birch}} = 0.1563$. However because the variable $S$ is discrete, it is convenient to carry out a continuity correction [15]. As $S$ jumps one space at a time, the cc should be 0.5 and so the statistic with cc will be $X_{\text{MHC}} = (\sum x_j - \sum E_j - 0.5) / (\sum V_j)^{0.5}$ [8]. The new $p$ value $P_{\text{MHC}} = 0.1573$ itself is already compatible with the exact value.

2.3. MC and MCB Tests Based on the Criterion of the Multiple Comparisons: General Observations. Let us suppose that in each stratum the hypotheses $H_j; \theta_j = 1$ versus $K_j; \theta_j > 1$ to error $\alpha_j$ are contrasted. Thereby $H = \cap H_j$ and $K = \cup K_j$. If the global null hypothesis $H$ is rejected when there exists at least one $j$ in which the individual test rejected $H_j$, then the alpha error $\bar{\alpha}$ of the global test ($H$ versus $K$) will be [16]

$$\bar{\alpha} = 1 - \prod (1 - \alpha_j).$$

(1)

In particular, if $\alpha_j = \alpha (\forall j)$ method MC is obtained (the “method of the multiple comparisons”), and its global alpha error will be

$$\bar{\alpha} = 1 - (1 - \alpha)^J.$$  

(2)

Method MC guarantees the compatibility of the results of the global test and of the individual tests, because the global test is significant if and only if at least one of the individual tests is so. When $J = 1$, the global test is the same as the individual test.

On the basis of the above, in general the test can be defined as follows. In each stratum $j$ an order statistic $S_j$ will have been defined which allows the $p$ value for each one of its points to be determined. If the points from all strata are mixed, they are ordered from the lowest value of their $p$ value to the highest and will be introduced one by one into the global critical region $R$ until a given condition (stopping rule) has been verified; then $R = \cup R_j$, with $R_j$ the critical region formed by the points in the stratum $j$ which belong to $R$. Let $\alpha_j$ be the largest of the $p$ values of the points in $R_j$. The real global alpha error $\bar{\alpha}_{\text{MC}}$ of the test constructed thus will be given by expression (1).

When the stopping rule is "stop introducing points into $R$ when the maximum of the $\alpha_j$ is as close as possible to $\alpha$ (but less than or equal to $\alpha$)," with $\alpha$ given by

$$\alpha = 1 - \sqrt{1 - \bar{\alpha}},$$

(3)

Table 3: $p$ values obtained by various methods for the data in the example of Li et al. [12]. Each asymptotic method is placed directly below the exact method from which it proceeds.

| Model | Test | Procedure | Statistic used | $p$ value |
|-------|------|-----------|----------------|-----------|
| Exact | Birch | Sum of successes (treated group) | $X_{\text{MH}}$ of Mantel-Haenszel  
(without cc) | 0.1563 |
| Asymptotic | MH | $X_{\text{MH}}$ of Mantel-Haenszel  
(with cc) | 0.0760 |
| 3 | Exact | $p$ value Fisher | 0.1573 |
| Asymptotic | MC | $p$ value Fisher | 0.3795 |
| Asymptotic | MCB | $\chi_3$ of Yates | 0.3887 |
| 2 | Exact | $p$ value Barnard | 0.1471 |
| Asymptotic | MC | $\chi_2$ of Martin et al. | 0.1513 |
| Asymptotic | MCB | $\chi_2$ of Martin et al. | 0.3533 |
| 1 | Exact | $p$ value Barnard | 0.1282 |
| Asymptotic | MC | $\chi_1$ of Pirie and Hamdan | 0.1352 |

Note: MH = Mantel-Haenszel test; MC = multiple comparisons method; MCB = method based on the multiple comparisons.
then method MC is obtained, and this method simultaneously controls global error \( \alpha \) and the individual error \( \alpha \). Now, the critical region \( R_j = R_{jMC} \) of each stratum consists of all the points whose \( p \) value is smaller or equal to \( \alpha_j = \alpha_{jMC} \leq \alpha \), \( R = R_{MC} = \cup R_{jMC} \) and the real global error will be \( \alpha_{MC} = 1 - \prod (1 - \alpha_{jMC}) \leq 1 - (1 - \alpha)^J = \bar{\alpha} \).

It is a simpler process to obtain the \( p \) value \( P_{MC} \) of some observed data. Let \( P_j \) be the \( p \)-value of the individual test in stratum \( j \). The first individual alpha error for which \( \alpha \) is concluded will be \( \alpha = P_0 = \min P_j \), so that for expression (2) the \( p \) value of the global test will be

\[
P_{MC} = 1 - (1 - P_0)^J. \tag{4}
\]

When the stopping rule is "stop introducing points into \( R \) when \( 1 - \prod (1 - \alpha_j) \) is the closest possible to \( \bar{\alpha} \) (but smaller than or equal to \( \bar{\alpha} \))," method MCB is obtained (the method "based on the multiple comparisons"). Because now only the global error \( \bar{\alpha} \) is controlled, its goal is similar to that of Jung's method [10]. The method MCB causes that \( R_j = R_{jMC} \), \( \alpha_j = \alpha_{jMC} \), \( R = R_{MC} = \cup R_{jMC} \) and the real global error is \( \bar{\alpha}_{MC} = 1 - \prod (1 - \alpha_{jMC}) \leq \bar{\alpha} \). Note that \( R_{MC} \subseteq R_{MC} \) since \( \bar{\alpha}_{MC} \leq \bar{\alpha} \), something to be expected given that method MC controls two errors and the MCB method controls only one of these.

Let us see how we can obtain the \( p \) value \( P_{MCB} \) of some observed data in which \( P_0 = P_j \) for example. The region \( R_{MCB} \) which yields the first significance of the global test is obtained when the observed point in stratum 1 is the last introduced into \( R_{MCB} \), that is, when \( \alpha_{jMCB} = P_j \); in the other strata it should be \( \alpha_{jMCB} \leq P_j \), but as close as possible to \( P_j \). Thus the \( p \) value will be \( P_{MCB} = 1 - \prod (1 - \alpha_{jMCB}) \). It can now be seen that \( \alpha_{jMCB} = \alpha_{jMCB} \) where \( \alpha_{jMCB} \) are the values of the MC test when this is carried out to the error \( \alpha = P_j \). Therefore \( P_{MCB} \leq P_{MC} \) and, for effects of calculating the \( p \) value \( P_{MCB} \), the \( p \) values \( \alpha_{jMCB} = \alpha_{jMC} \) and the regions \( R_{jMC} = R_{jMCB} \) will be written just as \( \alpha_j^* \) and \( R_j^* \), respectively. Thus, if \( \alpha_j^* \) is the largest \( p \) value in stratum \( j \) which is smaller or equal to \( P_j \),

\[
P_{MCB} = 1 - \prod (1 - \alpha_j^*). \tag{5}
\]

Methods MC and MCB may be applied with exact methods or with asymptotic methods and to any of the three models, as illustrated in the following sections.

### 2.4. MC and MCB Tests under Model 3

The \( p \) values of the Fisher exact test in each stratum are \( P_1 = 0.80073 \), \( P_2 = 0.57143 \), and \( P_3 = 0.14706 \). So, \( P_0 = P_3 = 0.14706 \) and \( P_{MC} = 0.3795 \) by expression (4). In order to apply method MCB the critical regions \( R_j^* (j = 1 \text{ and } 2) \) must be determined to the objective error \( \alpha = 0.14706 = P_0 = \alpha_j^* \). For \( j = 1 \), \( 9 \leq x_1 \leq 11 \) with \( \text{Pr}(x_1 = 11 | H_1) = 0.2862 > P_1 \); thus \( R_1^* = \phi \) and \( \alpha_1^* = 0 \). This same occurs for \( j = 2 \) (\( \alpha_2^* = 0 \)). For expression (5), \( P_{MCB} = 0.1471 \) (smaller than \( P_{jMCB} \)). Generally speaking the critical region of Birch [9] and Jung [10] has the form \( \sum x_j \geq S_0 = \sum x_j x_{j0} \), while that of method MCB is in the form \( \cup \{x_j \geq x_j^* \} \), with \( x_j^* \geq x_{j0} \). It can be proved that this generally implies that the Birch method will yield a \( p \) value smaller than or equal to that of method MCB when the \( p \) values \( P_j \) are similar or when the observed values \( x_{j0} \) are the highest possible.

Let us now apply an asymptotic test. In general, whatever the model is, the appropriate statistic is the chi-squared statistic [6]:

\[
\chi_j = \frac{x_j y_j - y_j x_j - c_j}{\sqrt{m_j n_j z_j (N_j - 1)}}. \tag{6}
\]

The appropriate value for the continuity correction \( c_j \) depends on the assumed model, and that value is what causes the results of the three models to be different. When \( c_j = 0 \) (\( V_j \) Pearson's classic chi-squared statistic is obtained. In the case here of Model 3, by making \( c_j = N_j / 2 \) the classic statistic \( \chi_{3j} \) (or the Yates chi-squared statistic) is obtained. Its maximum value is reached in stratum 3 (\( \chi_{33} = 1.0308 \)), which yields the \( p \) values \( P_0 = 0.15132 \) and \( P_{MC} = 0.3887 \). In order to apply method MCB, one must obtain in the other two strata the first value \( \chi_{3j}^* \) of \( \chi_{3j} \) which is larger than or equal to \( \chi_{33} \). As there is none, \( \alpha_1^* = \alpha_2^* = 0 \), \( \alpha_3^* = 0.15132 \) and \( P_{MCB} = 0.1513 \). Note that the asymptotic \( p \) values are similar to the exact ones, both with method MC and with method...
MCB. Despite the small size of the samples, the asymptotic methods function well (something which also occurs with the rest of the methods, as will be seen).

2.5. MC and MCB Tests under Model 2. The data in the example in reality proceeds from Model 2. In determining the \( p \) value \( P_j \) of an observed table \((x_{j1}, y_{j2} | m_j, n_j)\), the same steps are followed as in Model 3 (except the last, which is special): (1) define an order statistic \( S_j(x_j, y_j | m_j, n_j) \), which does not need to be the same one in each stratum; (2) determine the set of points \( R_j = \{(x_j, y_j | m_j, n_j) | S_j(x_j, y_j | m_j, n_j) \geq S_{j0}(x_{j0}, y_{j0} | m_{j0}, n_{j0})\} \); (3) calculate the probability of \( R_j \) under \( H_j \); \( P_j = \alpha_j = \pi_j \) given by \( \alpha_j(\pi_j) = \sum_{R_j} C_{m_j} C_{n_j} \pi_j^x (1 - \pi_j)^y \); and (4) determine the \( p \) value as \( P_j = \max_{\pi_j} \alpha_j(\pi_j) \), where \( \pi_j \) is the nuisance parameter that is eliminated by maximization (the most complicated step). Note that \( \pi_j \) is the marginal probability of columns under \( H_j \). In the case of Model 3 there is only one order statistic \( S_j \) possible [17], because the convexity of the region \( R_j \) must be verified and the points ordered “from the largest to the smallest value of \( x_j \).” In the case of Model 2 there are many possible test statistics. One of these is the order \( F_j \) of Boschloo [18]: order the points from the smaller to larger value of its one-tailed \( p \) value obtained using the Fisher exact test. It is already known [19] that the unconditional test based on the order \( F_j \) is uniformly more powerful (UMP) than Fisher’s own exact test. Although no unconditional order is UMP compared to the rest, the generally most powerful order is [3] the complex statistic \( B_j \) of Barnard [20].

As far as we know, the only program that carries out the above calculations for the statistic \( B_j \) is SMP.EXE, which may be obtained free of charge at http://www.ugr.es/local/bioest/software.htm. The program also gives the solution for other simpler test statistics. Using this program, because the minimum \( p \) value is \( P_j = 0.05653 \) then \( P_{MCB} = 0.1602 \). In order to obtain \( P_{MCB} \) one has to proceed as in the previous section, although now the process is now somewhat more difficult. In stratum 1, the table \((x_1, y_1) = (11, 10)\) is the one that gives a larger \( p \) value \( \alpha_1^* = 0.05462, \) but smaller than or equal to \( \alpha_1^* = 0.05653 \). In stratum 2 the results are \((x_2, y_2) = (4, 1)\) and \( \alpha_2^* = 0.05609 \). So, \( P_{MCB} = 0.1533 \), which is similar to that of \( P_{Birch} \) (the results are alike if other order statistics of the program SMP.EXE are used). It can be seen that the use of the unconditional method allows the inherent conservatism in the definitions of methods MC and MCB to be reduced.

In order to carry out the asymptotic test we shall use the optimal version of expression (6) for Model 2: \( \chi^2_{2j} \) is the value of expression (6) when \( c_j = 1 \) (or 2) if \( m_j \neq n_j \) (or \( m_j = n_j \)) [6]. Now the maximum \( \chi^2_{2j} \) is \( \chi^2_{2j} = 1.5805 \), whereby \( P_2 = 0.05700 \) and \( P_{MCB} = 0.1614 \) (a value, i.e., very near the 0.1602 of the exact method). Proceeding as above, the first values \( \chi^2_{2j} \) of \( \chi^2_{2j} \) (\( j = 1 \) or 2) which are larger than or equal to \( \chi^2_{23} \) are \( \chi^2_{21} = 1.5822 \) for \((x_1, y_1) = (10, 8)\) and \( \chi^2_{22} = 1.6056 \) for \((x_2, y_2) = (2, 0)\). This makes \( \alpha_1^* = 0.05680, \alpha_2^* = 0.05418, \) and \( P_{MCB} = 0.1588 \) (which is also a value, i.e., very close to the 0.1533 of the exact method).

2.6. MC and MCB Tests under Models 1 and 0. Let us suppose now that the data contained in the example in Table 2 proceed from Model 1. The determining of the \( p \) value \( P_j \) of an observed table \((x_{j1}, y_{j2} | m_j, n_j)\) is the same as in Model 2, but now the calculations are more complicated because the nuisance parameters must be eliminated (the marginal probabilities of rows and columns under \( H_j \)). Again there are many possible test statistics [1, 21], although none of them is UMP compared to the others. The generally more powerful statistic is again Barnard’s \( B_j \) statistic [22] and, as far as we know, the only program to apply it is TMP.EXE which may be obtained free of charge at http://www.ugr.es/local/bioest/software.htm. The program also gives the solution using other simpler test statistics. Using this program, the minimum \( p \) value is \( P_j = 0.04472 \) and from this \( P_{MCB} = 0.1282 \) (substantially smaller than \( P_{Birch} \)).

In order to carry out the asymptotic test we shall use the optimal version of expression (6) for Model 1: \( \chi^2_{1j} \) is the value of expression (6) when \( c_j = 0.5 \forall j \). The statistic is given by Pirie and Hamdan [23]. Now the maximum \( \chi^2_{1j} = 1.6149 \), with the result that \( P_j = 0.05317 \) and \( P_{MCB} = 0.1512 \).

Method MCB (which is very laborious to calculate) is omitted here, because the large number of points in the sample space will make \( \alpha_1^* = \alpha_2^* = \alpha_3^* = P_3 \) and so \( P_{MCB} \approx P_{MCB} \). Note that stratum 1 under Model 2 consists of \((m_1 + 1)(n_1 + 1) = (11 + 1) \times (13 + 1) = 168 \) points, but under Model 1 it consists of \((N_1 + 1)(N_2 + 1)(N_3 + 3)/6 = 25 \times 26 \times 27/6 = 2,925 \) points. For similar reasons, Model 0 can be treated as if it were Model 1 (by conditioning in the real obtained values \( N_j \)).

3. Sample Size under Model 2

3.1. Example and Conditional Solutions Obtained by Classic Methods. Jung [10] proposes a sample size calculation for its stratified exact test. For the example described in Section 2.1, he accepts Model 2 and sets out a case study with \( N_j = N/3 \) and \( m_j = N/6 \). The aim is to determine the value of \( N \) for the alternative hypotheses \((\theta_1, \theta_2, \theta_3) = (1, 30, 30)\), a type I error of \( \alpha = 0.1 \) and a power of \( 1 - \beta = 0.8 \). Jung also assumes that \((q_1, q_2, q_3) = (0.9, 0.75, 0.6)\), so that under the alternative hypothesis \( p_j = \theta_j q_j / (\theta_1 q_1 + \theta_2 q_2 + \theta_3 q_3) \). His solution is \( N_{Jung} = 62 \). From what can be deducted from other parts of his paper, the detailed solution is \( n_1 = n_2 = 20, n_3 = 22, m_1 = m_2 = 10, \) and \( m_3 = 11 \). These values are included in Table 4 (as well as the most relevant ones obtained in all the following). This sample size provides a real error of \( \alpha_{Jung} = 0.0563 \) and a real power of \( 1 - \beta_{Jung} = 0.8105 \).

Let us suppose that generally \( n_j = k_j m_j \), with \( k_j \) known values, and that the aim is to determine the values \( m_j \) which guarantee the desired power, which implies using Model 2. The reasoning that follows is the same as that with which Casagrande et al. [24] and Fleiss et al. [25] obtained the classic formula for sample size in the comparison of two independent proportions. The solutions without cc that
follow are a special case of those of Jung et al. [11]. The test \(\chi_{MHc}\) in Section 2.2 is based on the statistic \(\sum (x_j - E_j) - 0.5 = \sum D_j - 0.5\), where \(D_j = (k_j x_j - y_j)/(k_j + 1)\). Because \(D_j\) is distributed asymptotically as a normal distribution with the mean \(D_j = k_j m_j (p_j - q_j)/(k_j + 1)\) and the variance \(S_j^2 = k_j m_j (k_j p_j \bar{p}_j + q_j \bar{q}_j)/(k_j + 1)^2\), \(\bar{D} = \sum D_j\) will be asymptotically normal with the mean \(D = \sum D_j\) and the variance \(S^2 = \sum S_j^2\). Under \(H\), \(p_j = q_j = \pi_j (Vj)\), with the result that the mean and variance of \(D\) will be \(D_H = 0\) and of \(S_H^2 = \sum k_j m_j \pi_j \bar{\pi}_j/(k_j + 1)\), respectively, with \(\bar{\pi}_j = 1 - \pi_j\). Because under \(H\) the nuisance parameter \(\pi_j\) is estimated by \(z_j/N_j\), it is usual to substitute it by its average value under \(K\), that is, by \(\bar{\pi}_j = (p_j + k_j q_j)/(k_j + 1)\); hence \(\bar{\pi}_j = (\bar{p}_j + k \bar{q}_j)/(k_j + 1)\). Consequently the statistic \(\bar{D}\) will reach significance in the critical value \(D^*\) which verifies \(\bar{\pi} = \Pr(\bar{D} > D^* \mid H) = \Pr(z > (D^* - 0.5)/S_H)\), in which the number 0.5 corresponds to the cc indicated above and \(z\) refers to a normal standard variable. Therefore \(D^* = z_{1-\bar{\pi}} S_H + 0.5\), with \(z_{1-\bar{\pi}}\) the 100 \((1 - \bar{\pi})\)-percentile of the normal standard distribution. Under \(K\) the parameters \(D = D_K\) and \(S^2 = S_K^2\) are obtained in the values \(p_j\) and \(q_j\) which specify \(K\): \(D_K = \sum k_j m_j (p_j - q_j)/(k_j + 1)\) and \(S_K^2 = \sum k_j m_j (k_j p_j \bar{p}_j + q_j \bar{q}_j)/(k_j + 1)^2\). Given the above, the error beta will be

\[
m = m_0 \left[ 1 + \sqrt{1 + \frac{4}{m_0 \sum (p_j - q_j)}} \right] \]

with \(m_0 = \frac{z_{1-\bar{\pi}} \Sigma (p_j + q_j) (\bar{p}_j + \bar{q}_j)/2 + z_{1-\bar{\pi}} \Sigma (p_j \bar{p}_j + q_j \bar{q}_j)}{\Sigma (p_j - q_j)}\).

For the example at the beginning of this section (in which \(k_j = 1\)), if at first we restrict the solution to \(m_1 = m_2 = m_3 = m\), expression (9) indicates that \(m_0 = 8.27\) and \(m = 11.3\). Assuming that in this example the values of \(m_1\) are allowed to differ at most by 1, then the solution that is sought must be \(8 \leq m_j \leq 9 (Vj)\) without cc or \(11 \leq m_j \leq 12 (Vj)\) with cc. In the second phase, expression (7) indicates that in \(m_1 = m_2 = 11\) and \(m_3 = 12\) is the first time that \(\bar{\beta}_{MHc} = (0.183) \leq 0.2\), so that this is the solution with cc that was being sought \((N = 68)\). The solution without cc is obtained in the same way \((m_1 = m_2 = 8, m_3 = 9, \text{and } N = 50)\), but it is too liberal.

### 3.2. Solution Using the Exact Method MC

For fixed values of the global error \(\bar{\pi}\) and the sample sizes \((m_j, n_j)\), the method MC described in Section 2.3 allows one to obtain the critical region \(\bar{\beta}_{MC}\) and the real type I error \(\alpha_{MC} \leq \bar{\pi}\). Moreover, let \(\beta_{jMC}\) be the error beta for each individual test, with \(1 - \beta_{jMC}\) equal to the probability of the region \(\beta_{jMC}\) under \(K_j\). Because of the way method MC was defined, the real global error beta will be

\[
\bar{\beta}_{MC} = \prod_j \beta_{jMC}
\]

\[
= \prod_j \left[ 1 - \sum_{k_{MC}} \left( m_j \begin{pmatrix} n_j \\ y_j \end{pmatrix} \right) \right]
\]

If \(\bar{\beta}_{MC} \leq \bar{\pi}\), these values \((m_j, n_j)\) guarantee the desired power. If \(\bar{\beta}_{MC} > \bar{\pi}\), it is necessary to increase some values of \(m_j\) and/or \(n_j\) to repeat the previous procedure.

Let us initially assume that \(m_j = n_j\). The process for determining the sample sizes \(m_j\) may be shortened if it begins with a value \(m_j = m (Vj)\) like that of expression (8). With the method MC, one obtains that \(m = 12\) is not a solution because \(\bar{\beta}_{MC} = 0.2262 > 0.2\), but \(m = 13\) is a solution because \(\bar{\beta}_{MC} = 0.1723 \leq 0.2\). The solution can now be refined allowing values \(m_j\) to differ by a maximum of one. The final
solution is $m_1 = m_2 = 12, m_3 = 13$ $(N = 74)$, $\overline{\alpha}_{MC} = 0.0881$, and $\overline{\beta}_{MC} = 0.1880$.

Unconditional tests are more powerful when the sample sizes are slightly different [3], since the number of ties that produces any statistic $S_i$ that is used is reduced. By planning $n_j = m_j + 1$ and making the values of $m_i$ consecutive, the solution $m_1 = 10, m_2 = 11,$ and $m_3 = 12$ $(N = 69)$ is obtained, with $\overline{\alpha}_{MC} = 0.0924$ and $\overline{\beta}_{MC} = 0.1821$ (the solution based on $n_j = m_j - 1$ is worse). Actually, stratum 1 is of virtually no interest since in it $H_1 = K_1$. Despite everything, if it is introduced, the configuration $n_j = m_j + 1, m_1 = 1, m_2 = 11$ and $m_3 = 12$ $(N = 51)$ is correct because $\overline{\alpha}_{MC} = 0.0602$ and $\overline{\beta}_{MC} = 0.1833$.

3.3. Solution Using the Asymptotic Method MC Based on the Chi-Square Test with $cc$. In the following the procedure is the same as in Section 2.1, assuming for the moment that $m_j$ and $n_j$ can be any values. The numerator of $\chi_2$, may be written as $\hat{d}_j - c_j$, where $c_j$ is the $cc$ of Model 2 ($c_j = 2$ or 1 depending on whether $m_j$ and $n_j$ are equal or different, resp.) and $\hat{d}_j = n_jx_j - m_jy_j$ (the base statistic for the test) is asymptotically normal with mean $\hat{d}_j = m_jn_j(p_j - q_j)$ and variance $s^2_j = m_jn_j(n_jp_j + m_jq_j)$.

Under $H_j$, $p_j = q_j = \pi_j$ and $\hat{d}_j$ is asymptotically normal with mean 0 and variance $s^2_j = Nm_jn_j\pi_j(1 - \pi_j)$, with $\hat{d}_j = m_jn_j - m_jq_j$ (the base statistic for the test) is asymptotically normal with mean $\hat{d}_j = m_jn_j(p_j - q_j)$ and variance $s^2_j = m_jn_j(n_jp_j + m_jq_j)$. Under $H_j$, $\pi_j$ is estimated by $\hat{\pi}_j/N_j$, it is usual to substitute it by its average value under $K_j$, that is, by $\pi_j = (m_jp_j + n_jq_j)/N_j$; hence $\pi_j = (m_j\hat{\pi}_j + n_j\hat{\pi}_j)/N_j$. If each individual test is realized to the error $\alpha$ of expression (3), the critical value $d^*_j$ for $\hat{d}_j$ will verify $\alpha = P r (\hat{d}_j < d^*_j \mid H_j) = P r (\hat{d}_j < d^*_j \mid K_j) = P r (\hat{d}_j < d^*_j \mid K_j)$ and applying the first equality in expression (10)

$$\overline{\beta}_{MC} = \prod_j \overline{\beta}_{MC} = \prod_j P r \left\{ z \leq \frac{z_{1-\alpha/2\hat{H}_j} + c_j - d_{Kj}}{s_{Kj}} \right\},$$

in particular, if $m_j = n_j = m$ (i.e. $j = 2$, and

$$\overline{\beta}_{MC} = \prod_j P r \left\{ z \leq \frac{z_{1-\alpha/2\hat{H}_j} + c_j - d_{Kj}}{s_{Kj}} \right\}.$$
[13] A. Agresti, "Dealing with discreteness: making ‘exact’ confidence intervals for proportions, differences of proportions, and odds ratios more exact," *Statistical Methods in Medical Research*, vol. 12, no. 1, pp. 3–21, 2003.

[14] A. Agresti, "A survey of exact inference for contingency tables," *Statistical Science*, vol. 7, no. 1, pp. 131–177, 1992.

[15] D. R. Cox, "The continuity correction," *Biometrika*, vol. 57, pp. 217–219, 1970.

[16] Z. Šidák, "Rectangular confidence region for the means of multivariate normal distributions," *Journal of the American Statistical Association*, vol. 62, pp. 626–633, 1967.

[17] L. J. Davis, "Exact tests for $2 \times 2$ contingency tables," *The American Statistician*, vol. 40, no. 2, pp. 139–141, 1986.

[18] R. D. Boschloo, "Raised conditional level of significance for the $2 \times 2$ table when testing the equality of two probabilities," *Statistica Neerlandica*, vol. 24, no. 1, pp. 1–35, 1970.

[19] E. S. Pearson, "The choice of statistical tests illustrated on the interpretation of data classed in a $2 \times 2$ table," *Biometrika*, vol. 34, pp. 139–167, 1947.

[20] G. A. Barnard, "Significance tests for $2 \times 2$ tables," *Biometrika*, vol. 34, pp. 123–138, 1947.

[21] G. Shan and G. Wilding, "Unconditional tests for association in $2 \times 2$ contingency tables in the total sum fixed design," *Statistica Neerlandica*, vol. 69, no. 1, pp. 67–83, 2015.

[22] M. Andrés and T. García, "Optimal unconditional test in $2 \times 2$ multinomial trials," *Computational Statistics and Data Analysis*, vol. 31, no. 3, pp. 311–321, 1999.

[23] W. R. Pirie and M. A. Hamdan, "Some revised continuity corrections for discrete distributions," *Biometrics*, vol. 28, no. 3, pp. 693–701, 1972.

[24] J. T. Casagrande, M. C. Pike, and P. G. Smith, "An improved approximate formula for calculating sample sizes for comparing two binomial distributions," *Biometrics*, vol. 34, no. 3, pp. 483–486, 1978.

[25] J. L. Fleiss, A. Tytun, and H. K. Ury, "A simple approximation for calculating sample sizes for comparing independent proportions," *Biometrics*, vol. 36, pp. 343–346, 1980.
