# A REVIEW OF SAMPLE SIZE DETERMINATION IN COMPARATIVE STUDIES

by

M.K. Habib[1], K.M. Magruder-Habib[2], and L.L. Kupper[1]

Department of Biostatistics[1]
University of North Carolina at Chapel Hill
and
Division of Biometry[2]
Duke Medical Center

Institute of Statistics Mimeo Series No. 1840
November 1987

# A REVIEW OF SAMPLE SIZE DETERMINATION IN COMPARATIVE STUDIES

by

M.K. Habib,[1] K.M. Magruder-Habib[2], and L.L. Kupper[1]

Department of Biostatistics[1]

The University of North Carolina at Chapel Hill

and

Division of Biometry[2]

Duke University Medical Center

AMS 1980 Subject classification:

KEY WORDS AND PHRASES: Conditional test, Contingency tables Epidemiologic studies, Power, Sample size.

Research of M.K. Habib was partly supported by the Office of Naval Research contract Number N00014-83-K-0387

## SUMMARY

Sample size considerations for small and large sample designs of 2x2 comparative studies are reviewed. A test developed by Boschloo (1970) is recommended for small sample studies. This test is a simple modification of Fisher's unconditional test (FU) which is less conservative and hence more powerful than the FU test. New sample size tables for Boschloo's test are given in this paper. It is also argued that the uncorrected $X^2$-test is quite satisfactory for large sample comparative studies except under extreme situations, such as when compared proportions are near zero or one.

# 1. INTRODUCTION

Sample size determination is a major statistical issue in comparative studies and in particular, epidemiologic studies. There are several goals for sample size determination: to secure a certain power of the statistical test employed in the study, to avoid an inconclusive result because of too few sample units, to minimize the cost of the study (e.g. by adjusting the ratio of the sample sizes), and to minimize experimental risks to the subjects involved in the investigation.

The main goal of this paper is to provide a critical review of the notions and statistical issues associated with the problem of determining sample sizes when comparing two independent binomial distributions. (For a discussion of sample size, power, and smallest detectable effect for multivariate studies see Greenland, 1985, and for a brief review of sample size requirements that account for such issues as patient dropout, stratification of subjects, and time to achieve maximum benefit in the design of large controlled clinical trials see Donner, 1984. See also Lachin and Foulkes, 1986 and the references therein.) In addition, certain more practical issues, such as cost considerations, the determination of "clinically significant" differences or ratios of the parameters of interest, and "optimal" sample sizes in the case of unequal allocations of subjects or study units are discussed.

The determination of sample size depends on the following factors: the statistical test to be employed, the level of significance (or the size) of the test $\alpha$ (also called the nominal level of significance), the power of the test $P = 1-\beta$ (where $\beta$ is the probability of a type II error), and the parameters to be compared. Of these factors, the statistical test is the most crucial, since sample sizes are essentially predetermined to secure a certain power of the test at a certain level of significance. A careful consideration of the statistical tests employed in comparing two binomial distributions is, then, necessary. Upton (1982) reviewed more than twenty-two statistical tests concerning 2 x 2 comparative studies and discussed issues relevant to nominal and actual levels of significance. The author concluded that Fisher's exact test and the Yates-corrected chi-square test are "extremely conservative and inappropriate", and recommended a scaled version of the uncorrected chi-square test. Yates (1984), however, refuted Upton's criticism of both tests.

It is clear that certain considerations, such as the desired level of significance, the required power, and values of the parameters to be

compared, lead to small or large sample sizes. Different statistical tests, then, are appropriate for different considerations. From experience it is known that when one requires a small level of significance $\alpha$, a relatively large power P, and the parameters (under the alternative hypothesis) are "close in value", then relatively large samples are employed to avoid an inconclusive result. On the other hand, when one requires a relatively large $\alpha$, moderate P, and the absolute difference between the parameters is relatively large, then small sample sizes are to be expected.

We will review separately, in Sections 2 and 3, the literature dealing with small and large sample size considerations, the relevant statistical tests, and their actual (as compared to nominal) levels of significance and power. In Section 4, we provide a comparative discussion of the statistical tests reviewed in Sections 2 and 3. Specific suggestions are given as to the appropriate tests to be used in the various situations described. We recommend for small studies, a test introduced by Boschloo (1970) which is a simple modification of Fisher's unconditional test. Boschloo suggested a raised level of significance which results in a conservative actual level of significance that is closer in value to the nominal one than most available alternative tests. Because the critical region of the test recommended by Boschloo frequently contains more points than Fisher's unconditional test, Boschloo's test is frequently more powerful than Fisher's. This of course means that Boschloo's test requires smaller sample sizes than Fisher's unconditional test. New tables for sample sizes based on Boschloo's test for several values of $\alpha$ and P are presented in the appendix.

Finally, in Section 5 we discuss the merits of our recommendations from epidemiologic as well as statistical analytic points of view.

## 2. SMALL SAMPLE SIZE CONSIDERATIONS

In this section we discuss statistical tests which are appropriate for small sample sizes. Their actual and nominal levels of significance as well as power are discussed, and the literature dealing with tables of sample sizes calculated by employing these tests is reviewed.

To fix ideas, consider two independent samples of sizes $n_1$ and $n_2$ from two binomial distributions with parameters $p_1$ and $p_2$, respectively. Assume that it is desired to test the null hypothesis $H_0 : p_1 = p_2$ against the alternative hypotheses $H_1 : p_1 > p_2$ (one-sided test). Let X and Y be random variables representing the number of successes in the two samples respectively, and let

$$\psi = \frac{p_1 (1-p_2)}{p_2 (1-p_1)} ,$$

| | Success | Failure | Totals |
|---|---|---|---|
| Sample I | x | $n_1 - x$ | $n_1$ |
| Sample II | y | $n_2 - y$ | $n_2$ |
| Totals | t | $N - t$ | N |

where x,y denote the observed number of successes in both samples. The conditional distribution of X given t and $\psi$ is given by the non-central hypergeometric distribution

$$\Pr(X=x \mid t,\psi) = \frac{\binom{n_1}{x}\binom{n_2}{t-x}\psi^x}{\sum_{k=L}^{U}\binom{n_1}{k}\binom{n_2}{t-k}\psi^k} , \qquad (2.1)$$

where $L = \max(0, t - n_2)$ and $U = \min(n_1, t)$ (Fisher, 1935). Notice that under $H_0$ ($\psi = 1$), (2.1) simplifies to the central hypergeometric distribution

$$\Pr(X=s|t) = \frac{\binom{n_1}{x}\binom{n_2}{t-x}}{\binom{N}{t}} . \qquad (2.2)$$

With fixed marginal totals t, $n_1$, and $n_2$, the conditional probabilities in (2.2) may be used to establish evidence for or against the null hypothesis. When a point in the sample space is observed, the probability of observing this point or a more extreme one is calculated. The user then reports this probability, called the p-value, and may reject the null hypothesis if the p-value is, from his or her point of view, low enough. This approach to statistical testing was advocated by R. A. Fisher (see e.g. Fisher, 1973). A more common approach to statistical inference is the Neyman-Pearson hypothesis testing procedure. This approach suggests that a tolerable probability of a type I error be prespecified, thus determining the corresponding critical region of the test.

For a given nominal level of significance $\alpha$, let $x_c$ be the value of x (c.f. Table 1) such that for each t = 0, 1, . . . , N,

$$\sum_{k=x_c}^{U} \Pr(X=k/t) = \alpha_t \leq \alpha , \qquad (2.3)$$

and

$$\sum_{k=x_c-1}^{U} \Pr(X=k/t) > \alpha \qquad (2.4)$$

where U = min $(n_1, t)$. $\alpha_t$ is called the actual conditional level of significance and $x_c$ is called the critical value. The corresponding conditional critical region $C(x,t) = \{(x,t-x) = x_c \leq x \leq U\}$, t = 0,1, . . . , N, thus satisfies

$$\sum_{(x,t-x)\in C(\alpha,t)}^{U} \Pr(X=x/t) \leq \alpha .$$

The statistical test corresponding to this region is called Fisher's exact conditional (FEC) test.

Finney (1948) published tables of the critical values of the FEC test for $n_2 \leq n_1 \leq 15$ with the corresponding actual conditional levels of significance $\alpha_t$. Because of the discrete nature of the test, the values of $\alpha_t$ are occasionally much less than the nominal level of significance $\alpha$. In this sense, the FEC test is a conservative test. In other words, the test does not reject the null hypothesis as often as expected at the desired nominal level of significance $\alpha$. Finney's tables were extended by Finney et al. (1963) to $n_2 \leq n_1 \leq 30$ for nominal levels of significance $\alpha = 0.05, 0.01$ for single-tailed tests and $\alpha = 0.025, 0.005$ for two-tailed tests. Tables for $31 \leq n_2 \leq n_1 \leq 40$ were also given, however, without the corresponding actual levels of significance.

For the purpose of determining sample sizes, the FEC test is inadequate since the value of t is not known in advance. Furthermore, in comparing two binomial distributions, the sum of the number of successes in the two samples is a random variable. In this case, it is clear that an unconditional test is needed. Consider the unconditional level of significance

$$\alpha_F = \sum_{t=0}^{N} \alpha_t \Pr(T=t), \tag{2.5}$$

where $\Pr(T=t) = \binom{N}{t} p^t (1-p)^{N-t}$ under $H_0$, and p is the unknown common value of $p_1$ and $p_2$ (under $H_0$). The unconditional critical region corresponding to $\alpha_F$ is given by $c(\alpha) = \bigcup_{t=0}^{N} C(\alpha,t)$. The statistical test based on $C(\alpha)$ is called Fisher's unconditional (FU) test. From (2.5) it is clear that $\alpha_F$ is a function of the nuisance parameter p. Futhermore, notice that

$$\alpha_F = \sum_{t=0}^{N} \alpha_t \Pr(T=t) \leq \max_t \alpha_t \leq \alpha; \tag{2.6}$$

that is, the FU test is a conservative test. This point will be futher elaborated in Section 4.

The conditional power of the FEC test is given by

$$P(\psi/t) = \sum_{k=x_c}^{U} \Pr(x=k/t;\psi),$$

and the unconditional power of the FU test is defined by

$$P(\psi) = \sum_{t=0}^{N} P(\psi/t)\ \Pr(T=t), \qquad (2.7)$$

where

$$\Pr(T=t) = \sum_{k=L}^{U} \binom{n_1}{k} p_1^k (1-p_1)^{n_1-k} \binom{n_2}{t-k} p_2^{t-k} (1-p_2)^{n_2-t+k}.$$

Mainland and Sutcliffe (1953) discussed at length the factors involved in the determination of sample size. They confined themselves to the case of equal sample sizes $n_1 = n_2 = n$, and published several tables of the actual power of the FU test (c.f. 2.6) for selected sample sizes n and values of the parameters $p_1$ and $p_2$ at a nominal level of significance $\alpha = 0.05$. Bennett and Hsu (1960) presented power contours, based on the critical regions defined by the FU test for $n_2 \leq n_1$, 5(5)20, at nominal significance levels of 0.01 and0.05 using one-sided tests.

An alternative expression for $P(\psi)$ (c.f.2.7), invoking the assumption of independence of the samples, is given by

$$P(\psi) = \sum_{(x,y)\in C(\alpha)} \binom{n_1}{x} p_1^x (1-p_1)^{n_1-x} \binom{n_2}{y} p_2^y (1-p_2)^{n_2-y}. \qquad (2.8)$$

Gail and Gart (1973) computed the unconditional power of the FU test, using (2.8) in the case of equal sample sizes $n_1 = n_2 = n$, for several values of n at $p_1 = 0.8$ and $p_2 = 0.2$, where a one-tailed test was employed at $\alpha = 0.05$. They compared the power of the FU test to that of the $X^2$-test based on the arc-sine transformation (see e.g. Sillitto, 1949) and noted that the former is less powerful than the latter. Their main contribution, though, was to invert the power contours of Bennett and Hsu (1960) to produce tables of minimal sample sizes, n, required to obtain powers of at least 0.50, 0.80, and 0.90 at 0.01 and 0.05 nominal levels of significance for several values of $p_1$ and $p_2$. These tables were supplemented by employing the well known sample size formula using the arc-sine transformation (Sillitto, 1949) whenever n exceeded 35.

Remark 2.1

It should be noted here that the word "exact" in Fisher's exact conditional test refers to the fact that the conditional level of significance does not depend on any unknown parameters, and also that no appeal was made to the asymptotic methods of large sample theory. On the other hand, the unconditional level of significance of the FU test depends on the nuisance parameter p (c.f. 2.5.), and hence the FU test is not an exact test in the above sense. In addition, the title of the paper by Gail and Gart (1973) "The determination of sample sizes for use with the exact conditional test in 2x2 comparative trials" is misleading, since their tables were based on the FU test which is neither exact nor conditional. It should also be noted that in the summary of this paper, the authors erroneously state that "Mainland and Sutcliffe (1953) and Bennett and Hsu (1960) have calculated the power of the exact conditional (Fisher-Irwin) test for differences between proportions." This, of course, is inaccurate since the papers they referred to calculate the powers of the FU test and not of the FEC test as they claimed.

Haseman (1978) provided sample size tables based on the FU test in the case of equal sample sizes ($n_1 = n_2 = n$). Some of the sample sizes reported in Haseman's tables are as large as n = 503. It should be noted, though, that tables of the critical regions of the FU test for such large sample sizes are not available in the literature. Haseman addressed the fact that the FU test is a conservative test and warned that for small values of the nuisance parameter p (e.g. $p < .01$), the unconditional level of significance could be much smaller than the desired nominal level of significance so that the test in this case is virtually useless. Casagrande, Pike, and Smith (1978a) independently published several tables of sample sizes similar to those of Haseman's. Remark 2.1 applies also to the papers by Haseman (1978) and Casagrande et al. (1978a).

## 3. LARGE SAMPLE SIZE CONSIDERATIONS

For a relatively small nominal level of significance $\alpha$ and a relatively large power P, and when comparing two binomial parameters $p_1$ and $p_2$ which are close in value, large sample sizes should be expected to establish significance. In this case, the determination of the exact levels of significance and power of the FEC test or of the FU test involves extensive calculations. Several asymptotic approximations, therefore, have been suggested. The first approximation is the $X^2$-test. As in Section 2, consider two independent samples of sizes $n_1$ and $n_2$ from two binomial distributions with parameters $p_1$ and $p_2$, respectively. Assume that one is interested in testing the null hypothesis $H_0 : p_1 = p_2$ vs. the alternative hypothesis $H_1 : p_1 > p_2$. For specific $\alpha$, $\beta$, and for the case of equal sample sizes $n_1 = n_2 = n$, the formula for determining the common sample size, n, is given by (see e.g. Fleiss, 1981)

$$n = \frac{[z_\alpha \sqrt{(2\,\bar{p}\,\bar{q})} + z_\beta \sqrt{(p_1 q_2 + p_2 q_2)}\,]^2}{(p_1 - p_2)^2}, \tag{3.1}$$

where $\bar{p} = (p_1 + p_2)/2$, $\bar{q} = 1-p$, $q_i = 1-p_i$, $i = 1,2$, $z_\gamma$ is such that $1 - \phi(z_\gamma) = \gamma$, and $\phi$ is the distribution function of a standard normal random variable. Kramer and Greenhouse (1959) modified formula (3.1) for sample sizes when the Yates-continuity correction is incorporated in the test statistic. Their formula is given by

$$n_c = (n/4)\,(1 + [1+8/n(p_1-p_2)])^2. \tag{3.2}$$

For a fixed nominal level of significance $\alpha$, it is well known that the $X^2$-test produces actual levels of significance which occasionally exceed $\alpha$. On the other hand, the corrected $X^2$-test frequently results in actual levels of significance which are much less than $\alpha$. For this reason, the power of the $X^2$-test is sometimes inflated, and the power of the corrected $X^2$-test is frequently much less than expected. As a consequence, formula (3.2) produces, in general, larger sample sizes than (3.1). To remedy this situation, Casagrande, Pike, and Smith (1978b) derived a formula that produces sample sizes which lie roughly in the middle of n and $n_c$. This formula is given

by

$$n' = (n/4)\{1+\sqrt{[1+4/n(p_1-p_2)]}\}^2 . \qquad (3.3)$$

The arguments used in deriving formula (3.3) are unconvincing, and furthermore, the authors failed to indicate which statistical test is to be used with their formula. From their Table 2, it also is noticed that the sample sizes often exceed the sample sizes calculated by employing the FU test which are already known to be inflated. Based on formula (3.3), Aleong and Bartlett (1979) published graphs for calculating sample sizes for selected values of $\alpha,\beta$ and several values of $p_1$ and $p_2$ for one- and two-sided tests.

In many practical situations, such as studies which compare two medical treatments with different levels of risk to the human subjects involved, or when the cost per unit of one study sample is markedly different from that of the other sample, it is prudent to consider different sample sizes. Consider, then, two samples of sizes $n_1$ and $n_2$ such that $n_1 = m$ and $n_2 = rm$ (say), where r is some fixed positive number. In this case, formulae (3.1) and (3.2) take the forms

$$m = \frac{[z_\alpha\sqrt{\{(r+1)\,\bar{p}\,\bar{q}\}} + z_\beta\sqrt{(r\,p_1q_1 + p_2q_2)}]^2}{r(p_1 - p_2)^2} , \qquad (3.4)$$

where $\bar{p} = (p_1 + rp_2)/(r+1)$ and $\bar{q} = 1 - \bar{p}$, and

$$m' = (m/4)\{1 + \sqrt{[1 + 2(r+1)/rm(p_1-p_2)]}\}^2 . \qquad (3.5)$$

Because of fixed cost considerations (e.g. only enough money to interview a certain number of subjects) or the limited availability of sample units (e.g. as in case-control studies involving rare diseases), it is frequently the case that the sample sizes are of necessity prespecified. In this case, the inverse problem of estimating the actual power of the test for predetermined sample sizes is of interest. For this purpose Fleiss, Tytun, and Ury (1980) have presented the following simple approximation of $m'$ (c.f.(3.5))

$$m^* = M + (r+1)/r(p_1-p_2) . \qquad (3.6)$$

Using (3.4) and (3.6), the percentile corresponding to the actual power

is given by

$$z_\beta = \frac{\sqrt{(rd^2m^* - (r+1)\,d)} - z_\alpha \sqrt{((r+1)\,\bar{p}\,\bar{q})}}{\sqrt{(rp_1q_1 + p_2q_2)}}, \qquad (3.7)$$

where $d = p_1 - p_2$. Utilizing (3.7) they observed that, other factors being fixed, the actual power of the test is increased when relatively more sample units are taken from the population whose underlying probability (under $H_1$) is further from 0.50.

Ury and Fleiss (1980) derived still another approximation of formula (3.5). They argued that since the quantity $p(1-p)$, $0 < p < 1$, varies little over a wide range of values of p, then one may replace $p_1q_1$ and $p_2q_2$ in (3.4) by $\bar{p}\,\bar{q}$, where $\bar{p} = (p_1 + rp_2)/(r+1)$ and $\bar{q} = 1 - \bar{p}$, to obtain from (3.5) the approximation

$$m^{**} = \frac{(r+1)}{4rWd}\left\{1 + \sqrt{(1 + 2W)}\right\}^2, \qquad (3.8)$$

where $W = d/(z_\alpha + z_\beta)^2\, p\, q$. In their paper, they compared sample sizes obtained from formulae (3.4), (3.5), (3.6), and (3.8), for the case $r = 1$, to sample sizes obtained by employing the FU test. See Gail et al. (1976) for futher discussion of issues concerning the number of controls needed in clinical and epidemiologic studies (See also Lee, 1984). Furthermore, Brittain and Schlesselman (1982) discussed strategies of optimal allocation of study units for the comparision of proportions in two groups which maximize the precision in the estimation of the difference between the proportions, maximize the precision in the estimation of their ratio, maximize the power to detect a group difference, and minimize the cost of a study.

An alternative sample size formula is obtained by employing the familiar arc sine (or the angular) transformation of the square root of a binomially distributed random variable (Eisenhart, 1947). It is well known that if X is a binomially distributed random variable with parameters n and p, then the random variable

$$F = \arcsin\sqrt{\left(\frac{X}{n}\right)} \qquad (3.9)$$

is asymptotically normally distributed about a mean of arc sine $\sqrt{p}$ and a variance of $1 / \sqrt{(4n)}$, where the angle is measured in radians (see e.g. Curtiss, 1943). In the case of small or moderate sample sizes, the variance of the transformation in (3.9) is a function of the parameter p for a range of values of n. Several modifications of this transformation have been suggested in order to improve the stabilization of the variance. Chanter (1975) showed that a modification due to Anscombe (1948) is the most successful at stabilizing the variance. The arc sine transformation also improves the closeness of the binomial distribution to normality (see e.g. Bartlett, 1947). Thus, if $x_1$ and $x_2$ are the number of successes in two independent binomial samples for parameters $(n_1,p_1)$ and $(n_2,p_2)$, respectively, then the statistic

$$\frac{2(f_1 - f_2)}{\sqrt{\{(n_1)^{-1} + (n_2)^{-1}\}}} , \tag{3.10}$$

where $f_i$ = arc sine $\sqrt{x_i/n}$, has asymptotically a normal distribution with mean zero and variance one under $H_0$: $(p_1 = p_2)$, and with mean $\mu$ = 2(arc sine $\sqrt{p_1}$ - arc sine $\sqrt{p_2}$ / $\{(n_1)^{-1} + (n_2)^{-1}\}$ under $H_1$: $(p_1 > p_2)$. In this case, Sillitto (1949) gave the following sample size formula, for the case of equal sample sizes $n_1 = n_2 = n$

$$n = \frac{(z_\alpha + z_\beta)^2}{2(\text{arc sine } \sqrt{p_1} - \text{arc sin } \sqrt{p_2})^2} \tag{3.11}$$

It has been noted by several authors, among them Aleong and Bartlett (1979), that formulae (3.1) and (3.11) give similar sample sizes. They also noted that (3.1) and (3.11) give smaller sample sizes than those derived by using the FU test (Haseman, 1978). This, of course, is expected since the FU test is more conservative and hence less powerful than both the uncorrected $X^2$-test and the $X^2$-test based on the arc sine transformation (c.f.(3.10)) which tend to overestimate the power and hence underestimate the sample size in particular for extreme values of $p_1$ and $p_2$. For this reason, Walters (1979) proposed an ad hoc continuity-corrected version of the arc sin transformation of the form

$$(z_{1-\alpha} - z_{\beta})^2 = 2n \{\text{arc sin} \sqrt{(P_1 - 1/2n)} - \text{arc sin} \sqrt{(P_2 + 1/2n)}\}^2 . \qquad (3.12)$$

An iterative procedure is used to determine the sample size n from (3.12). Ury (1981;82) studied the power of the corrected arc sin test under different study designs. Dobson and Gebski (1986) introduced a modification of the corrected arc sin test which yields a closed-form expression for the sample size for the equal sample size design. This modification was accomplished by substituting the first two terms of the Taylor expansion of arc sin $\sqrt{p+h}$ around h=0 in equation (3.12). This modification yields the equation

$$Z^2 = 2n [\Delta - C/4n]^2 , \qquad (3.13)$$

where $Z = (z_{1-\alpha} - z_{\beta} = \Delta \quad (\text{arc sin} \sqrt{p_1} - \text{arc sin} \sqrt{p_2})$, and $c = \{(p_1 q_1)^{-1/2} + (p_2 q_2)^{-1/2}\}$. From (3.13), the sample size n is given by

$$n = \{Z + \sqrt{Z^2 + 2c\Delta}\}^2/8\Delta^2.$$

## 4. DISCUSSION

As indicated in Section 1, a primary reason for predetermining sample sizes when conducting a statistical study is to secure a certain power for the test employed at a certain nominal level of significance. Most often, however, the actual level of significance of the appropriate test is different from the nominal one. This is the case either because of the discrete nature of the test or else because of the use of asymptotic tests. A critical investigation of the actual levels of significance and power of the statistical tests discussed in Section 2 and 3 are, then, necessary. It was pointed out in Section 2 that the actual level of significance $\alpha_F$ of Fisher's unconditional test is often much less than the nominal one, $\alpha$. The following is a modification of an example given by McDonald, Davis, and Milliken (1977), which gives a clear picture of the extent of conservatism of the FU test. Assume that two binomial distributions with parameters $p_1$ and $p_2$ are to be compared. Two independent samples of sizes $n_1=5$ and $n_2=4$, respectively, are obtained, and the hypothesis of interest is $H_0$: $p_1=p_2$ (=p: unknown) vs. $H_1$ : $p_1 > p_2$ at a nominal level of significance $\alpha= 0.05$. Let x and y be the number of successes in the two samples, respectively, and put (x+y) = t (c.f. Table 1). Notice that $\alpha_t = 0$ for t =0,1,2,3,7,8,9, and $\alpha_4 = 0.0396$, $\alpha_5 = 0.0079$, $\alpha_6 = 0.0476$. The actual (unconditional) level of significance of the FU test is given by (c.f. 2.5.)

$$\alpha_F(p) = \sum_{t=0}^{N} \alpha_t\, g_0(t) = \alpha_4 g_0(4) + \alpha_5 g_0(5) + \alpha_6 g_0(6)$$

$$= \Pr(X=4,\ Y=0) + \Pr(X=5,\ Y=0) + \Pr(X=5,\ Y=1)$$

$$= \binom{5}{4} p^4 (1-p) \binom{4}{0} (1-p)^4 + \binom{5}{5} p^5 \binom{4}{0} (1-p)^4 + \binom{5}{5} p^5 \binom{4}{1} p(1-p)^3$$

$$= 5p^4 (1-p)^5 + p^5 (1-p)^4 + 4p^6 (1-p)^3 , \qquad (4.1)$$

where $g_0(t) = \Pr(T=t) = \binom{N}{t} p^t (1-p)^{N-t}$. It is clear that $\alpha_F$ is a function of the nuisance parameter p. The least upper bound on $\alpha_F$ is given by

$$\alpha_F^* = \sup_{0<p<1} \alpha_F(p) = 0.027 \ , \qquad (4.2)$$

and it should be noted that for any $\alpha$ between 0.04 and 0.119, the same critical region is obtained with the same $\alpha_F^*$ as in (4.2). If the value of the nuisance parameter p is other than the one which corresponds to $\alpha_F^*$, then the actual level of significance will be even less than 0.0207 while $\alpha = 0.05$. It is clear then that the FU test is quite conservative for small sample sizes. It will be shown later in this section that this conservatism persists even for moderate and large sample sizes.

Bernary (1945a) advocated the use of $\alpha_F^*$ as the actual level of significance which obviously results in a conservative test (see also Bernard, 1945b and Fisher, 1945). Boschloo (1970) proposed another modification of the FU test in order to produce a much closer maximum unconditional level of significance to $\alpha$ than $\alpha_F^*$. This modification will be referred to here as the Fisher-Boschloo modified (FBM) test. The FBM test is performed in a manner similar to that of the FU test. However, instead of determining the conditional critical regions (CCR) of the FU test such that $\alpha_t \leq \alpha$, t = 0,1, . . . , N (c.f. (2.3) and (2.4)), a raised level of significance $\gamma_\alpha$ is selected and the CCR are determined such that $\alpha_t \leq \gamma_\alpha$, t = 0,1, . . . , N (instead of $\alpha_t \leq \alpha$). $\gamma_\alpha$ is chosen in such a way that the unconditional level of significance $\gamma_\alpha(p) = \sum_{t=0}^{N} \alpha_t \Pr(T=t)$ (C.F. 2.5.) for the FBM test does not exceed the nominal level of significance $\alpha$ for any value of the nuisance parameter p. For the example considered in the beginning of this section, $\gamma_\alpha$ could be chosen as high as $\gamma_{0.05} = 0.165$, and, as a result, the point (3,0) is added to the critical region {(4,0), (5,0), (5,1)} of the FU test. In this case, $\gamma^* = \sup \gamma_\alpha(p) = 0.45$ which is much closer to $\alpha = 0.05$ than $\alpha_F^*$ (= 0.0207). Furthermore, since the unconditional critical region of the FBM test, in general, contains additional points to those of the unconditional critical region of the FU test, the FBM test is more powerful and hence requires smaller sample sizes than the FU test to secure a certain required level of power. Boschloo (1970) published tables of raised levels of significance $\gamma_\alpha$ for selected values of $\alpha$, $\beta$ and $n_2 \leq n_1 \leq 50$. It should be noted though that he did not publish the critical regions of the FBM test. However, Suissa and Shuster (1985) developed an analytical method for the maximization necessary to obtain Boschloo's unconditional actual levels of

significance.

Garside and Mack (1976) compared the actual levels of significance of several statistical tests, among them the FU test, the FBM test, the uncorrected $X^2$-test ($X_c^2$), and the Yates-corrected $X^2$-test ($X_c^2$). Several sample sizes and values of the nuisance parameter p were considered. Three of their tables are reproduced below:

Table 2

Actual levels of significance: $n_1 = n_2 = 40$, $\alpha = 0.05$

| p | FU | FBU | $X_c^2$ | $X^2$ |
|---|---|---|---|---|
| 0.1 | 0.0194 | 0.0405 | 0.0193 | 0.0544 |
| 0.2 | 0.0296 | 0.0444 | 0.0258 | 0.0509 |
| 0.3 | 0.0306 | 0.0486 | 0.0247 | 0.0529 |
| 0.4 | 0.0278 | 0.0451 | 0.0278 | 0.0474 |
| 0.5 | 0.0284 | 0.0465 | 0.0284 | 0.0466 |

Table 3

Actual levels of significance, $n_1 = n_2 = 500$, $\alpha = 0.05$

| p | Test FU | FBU | $X_c^2$ | $X^2$ |
|---|---|---|---|---|
| 0.1 | 0.0400 | - | 0.0400 | 0.0500 |
| 0.2 | 0.0424 | - | 0.0420 | 0.0504 |
| 0.3 | 0.0436 | - | 0.0436 | 0.0502 |
| 0.4 | 0.0436 | - | 0.0436 | 0.0499 |
| 0.5 | 0.0446 | - | 0.0436 | 0.0500 |

Table 4

Actual levels of significance $n_1 = 100$, $n_2 = 10$, $\alpha = 0.001$

| p | FU | FBU | $X_c^2$ | $X^2$ |
|---|---|---|---|---|
| 0.4 | 0.0000 | - | 0.0000 | 0.0000 |
| 0.5 | 0.0003 | - | 0.0001 | 0.0006 |
| 0.6 | 0.0003 | - | 0.0003 | 0.0011 |
| 0.7 | 0.0004 | - | 0.0006 | 0.0017 |
| 0.8 | 0.0004 | - | 0.0010 | 0.0030 |
| 0.9 | 0.0003 | - | 0.0016 | 0.0064 |

From Table 2, it can be seen that, for a nominal level of significance $\alpha = 0.05$, the largest actual level of significance of the FU test is 0.0306 at $p = 0.3$. This shows, as noted before, that the FU test is conservative even for such moderate sample sizes ($n_1 = n_1 = 40$), resulting in an unnecessary loss of power (see the comment by Starmer, et al in the paper by Conover, 1974). It follows then that the sample sizes based on the FU test (the so called exact sample sizes) are inflated. The same remarks hold for the $X_c^2$-test which is at least as conservative as the FU test. Sample sizes based on the $X_c^2$-test (c.f. 2.3.) are thus, as or more inflated than the "exact" sample sizes. The $X_c^2$-test has been criticized by several authors, among them Plackett (1964), Grizzle (1967), Conover (1974), and Upton (1982). The use of the $X_c^2$-test has, nevertheless, been defended by several others, including Mantel and Greenhouse (1968), and Yates (1984). (See also the discussion of several continuity corrections by Haber, 1980.)

On the other hand, the actual levels of significance of the FBU test are much closer to the nominal levels (Table 2); and, as noted by Boschloo, this results in a considerable increase in power. For large sample sizes, the uncorrected $X^2$-test performs reasonably well. Its actual levels of significance are reasonably close to the nominal ones (Tables 2 and 3). It should be warned, though, that the actual levels of significance of the $X^2$-test occasionally exceed $\alpha$. This problem is particularly alarming when the sample sizes are very different and $\alpha$ is chosen to be very small. This

phenomenon is apparent in Table 4 where $n_1 = 100$, $n_2 = 10$, and $\alpha = 0.001$. For example at $p = 0.9$, the actual level of significance is 0.0064, more than six times the nominal level. It should be noted, though, that most of the available statistical tests suffer from the same problem, in particular, for extreme values of p.

Finally, a test similar to the FBM was developped by McDonald et al. (1977), in which Boschloo's ideas of raised levels of significance were employed. However, the critical regions of their test, which we will call Fisher's unconditional raised level (FURL) test, are determined along the same lines as Bernard (1947). McDonald et al. (1977) published the critical regions of the FURL test for $n_1 \leq n_2 \leq 15$, and tables for $n_1 \leq n_2 \leq 20$ are available in McDonald et al. (1975). The actual levels of significance of the FURL test are very similar to those of the FBM test.

## 5. RECOMMENDATIONS

For the small sample case, there are two chief advantages to using the FBM test instead of the more traditional FU test. The first is in the design phase of a study where it is possible to save on the numbers of subjects required to reach a valid conclusion. We can crudely estimate savings in sample size by subtracting the values of our Table A.2 from the corresponding values of Haseman (1978), and expressing the difference as a percent decrease from the Haseman values. With an $\alpha$ level of .05, the average decrease in sample size is 13.1% for power of .9, 16.4% for power of .8, and 26.1% for power of .5. The corresponding percentages at $\alpha = .01$ are 9.1% 11.7%, and 16.6%.

Such considerations become very real in the study of rare diseases, such as megakaiyoblastic leukemia, promyelocytic leukemia, male breast cancer, and osteogenic sarcoma, to mention just a few. For such diseases, even at major tertiary care medical centers, only 2 or 3 cases might be diagnosed in a 12 month period. Thus, in addition to expense, several years might be saved by making use of a test statistic with a smaller sample size requirement. Accumulation of cases within shortened time frames is also critical as treatments tend to vary markedly over time. In the case of male breast cancer, for example, where the few case series that are reported have been accumulated over time periods in excess of 10 years, even though sample sizes may be adequate, treatment changes markedly so that the results may not be valid as the cases are not necessarily homogeneous.

Although parsimonious use of human subjects is important in all epidemiologic studies, it can be especially critical in randomized clinical trials, in contrast to cohort or case control studies in which it is possible to increase the ratio of unexposed to exposed or of controls to cases. For example, in a case-control study of a rare cancer where only very few cases are available for study within a reasonable time, we can increase the ratio of controls to cases. However, if these same cases were in a randomized clinical trial to compare two different therapies, there is no way around the fact that only half of the cases would be available to each treatment group. Thus, the FBM test may improve the study of rare diseases and their treatments. It may also make possible the study of multiple treatments versus a control treatment. Futhermore, when patients are subjected to a treatment that carries especially deleterious side effects (as many oncological

treatments do) or is especially invasive (e.g. surgical versus non-surgical treatment), it is important to determine sample sizes parsimoniously.

The second major advantage to the FBM test is at the analytic stage. Because it is less conservative than the FEC test (the actual alpha level being closer to, but still less than, .05), hypotheses that may have been accepted by the FEC test with alpha levels up to .07, will be rejected by the FBM test at the .05 level.

We conclude, then, by recommending the FBM test for reasonably small sample studies, and the $X^2$-test for moderate to large sample studies. New tables of sample sizes based on the FBM test are given in the appendix.

REFERENCES


Aleong, J. and Bartlett, D.E. (1979). Improved graphs for calculating sample sizes when comparing two independent binomial distributions. Biometrics 35 , 875-881.

Anscombe, F.J. (1948). The transformation of Poisson, Binomial, and Negative-Binomial data.

Barnard, G.A. (1945a). A new test for 2x2 tables. Nature , 156 , 177.

Barnard, G.A. (1945b). A new test for 2x2 tables. Nature , 156 , 783-784.

Bartlett, M.S. (1947). The use of transformation. Biometrics 3 , 39-52.

Bennett, B.M. and Hsu, P. (1960). On the power function of the exact test for the 2x2 contingency table. Biometrics 47 , 393-398.

Boschloo, R.D. (1970). Raised conditional level of significance for the 2x2 table when testing for the equality of two probabilities. Statistica Neerlandica , 21 , 1-35.

Brittain, E. and Schlesselman, J.J. (1982). Optimal allocation for the comparision of proportion. Biometrics. 38, 1003-1009.

Casagrande, J.T., Pike, M.C., and Smith, P.G. (1978a). The power function of the "exact" test for comparing two binomial distributions. Applied Statistics 27 , 176-180.

Casagrande, J.T., Pike, M.C., and Smith, P.G. (1978b). An imporoved approximate formula for calculating sample sizes for comparing two binomial distributions. Biometrics 34 , 483-486.

Chanter, D.O. (1975). Modifications of the angular transformation. Applied Statistics 24 , 354-359.

Conover, W.J. (1974). Some reasons for not using the Yates continuity correction on 2x2 contingency tables. Journal of the American Statistical Association , 69 , 374-382

Curtiss, J.H. (1943). On transformation used in the analysis of variance. Annals of Mathematical Statistics 14, 107-122.

Dobson, A. J. and Gebski, V. J. (1986). Sample sizes for comparing two independent proportions using the continuity-corrected arc sine transformation. The Statistician 35, 51-53.

Donner, A. (1984). Approaches to sample size estimation in the design of clinical trials - a review. Statis. Medic. 3, 199-214.

Eisenhart, G. (1947). Inverse sine transformation of proportions. In Selected Techniques of Statistical Analysis, 14, eds. G. Eisenhart et al. Chapter 16. New York: McGraw-Hill.

Feigl, P. (1978). A graphical aid for determining sample size when comparing two independent proportions. Biometrika 34 , 111-122.

Finney, D.J. (1948). The Fisher-Yates test in 2x2 contingency tables. Biometrika 35 , 145-156.

Finney, D.J., Latscha, R., Bennett, B.M., Hsu, P. and Person, E.S.V. (1963). Tables for significance testing in a 2x2 contingency table. Cambridge University Press, London.

Fisher, R.A. (1935). The logic of inductive inference. Journal of the Royal Statistical Society Series A, 98, 39-54.

Fisher, R.A. (1945). A new test for 2x2 tables. Nature, 156 , 388.

Fisher, R.A. (1973). Statistical Methods and Scientific Inference, 3rd ed. Hafner Press, London.

Fleiss, J.L. (1981). Statistical Methods for Rates and Proportions, Second Edition, John Wiley and Sons, New York.

Fleiss, J., Tytun, A., and Ury, H.K. (1980). A simple approximation for calculating sample sizes for comparing independent proportions. Biometrics 36 , 343-346.

Gail, M., Williams, R., Byar, D. and Brown, C. (1976). How many controls? J. Chron. Disease, 29 723-731.

Gail, M. Gart, J.J. (1973). The determination of sample sizes for use with the exact conditional test in 2x2 comparative trials. Biometrics 29 , 441-448.

Garside, G.R. and Mack, C. (1976). Actual type 1 error probabilities for various tests in the homogeneity case of the 2x2 contingency table. The American Statistician, 30, 18-21.

Greenland, S. (1985). Power, sample size and smallest detectable effect determination for multivariate studies. Statistics in Medicine, 4, 117-127.

Grizzle, J. (1967). Continuity correction in the $X^2$-test for 2x2 tables. The American Statistician, 21, 28-32.

Haber, M. (1980). A comparison of some continuity corrections for the chi-squared test on 2x2 tables. Journal of the American Statistical Association, 75, 510-515.

Haseman, J.K. (1978). Exact sample sizes for use with the Fisher-Irwin test for 2x2 tables. Biometrics 34, 106-109.

Irwin, J.O. (1935). Tests of significance for differences between percentages based on small numbers. Metron 12, 83-94.

Lachin, J.M. and Foulkes, M.A. (1986). Evaluation of Sample Size and Power for Analyses of Survival with Allowance for Nonuniform Patient Entry, Losses to Follow-up, Noncompliance, and Stratification. Biometrics 42, 507-519.

Lee, Y.J. (1984). Quick and Simple Approximation of Sample Sizes for Comparing Two Independent Binomial Distributions: Different-Sample-Size Case. Biometrics, 40, 239-241.

Kramer, M. and Greenhouse, S.W. (1959). Determination of sample size and selection of cases. In Psychopharmacology: Problem Evaluation. J.W. Cole and R.W. Grevard (eds.), National Academy of Sciences, National Research Council. Washington, D.C., Publication 583,356-371.

Mainland, D. and Sutcliffe, M.I. (1960). Statistical methods in medical research. II. Sample sizes in experiments involving all-or-none responses. Canadian Journal of Medical Science 31, 406-416.

Mantel, N. and Greenhouse, S.W. (1968). What is continuity correction? The American Statistician, 22, 27-30.

McDonald, L.L., Davis, B.M. and Milliken, G.A. (1975). A nonrandomized unconditional test comparing two proportions. College of Commerce and Industry. Research paper 94, Univ. of Wyoming at Laramie.

McDonald, L.L., Davis, B.M. and Milliken, G.A. (1977). A nonrandomized unconditional test for comparing two proportions in 2x2 contingency tables. Technometrics, 19, 145-157.

Person, E.S. (1947). The choice of statistical tests illustrated in the interpretation of data classed in 2x2 table. Biometrika 34, 130-167.

Plackett, R.L. (1964). The continuity correction in 2x2 tables. Biometrika 51 , 327-337.

Sillitto, G.P. (1949). Note on approximations to the power function of the 2x2 comparative trial. Biometrika 38 , 347-352.

Suissa, S. and Shuster, J.J. (1985). Exact unconditional sample sizes for the 2x2 binomial trial. Journal of the Royal Statistical Society, A 148, 317-327.

Upton, G.J.G. (1982). A comparison of alternative tests for the 2x2 comparative trial. Journal of the Royal Statistical Society A 145, 86-105.

Ury, H.K. (1981). Continuity-Corrected Approximations to Sample Size or Power When Comparing Two Proportions: Chi-Squared or Arc Sine? The Statistician, 30, 199-203.

Ury, H.K. (1982). Comparing Two Proportions: Finding $p_2$ when $p_1$, n, $\alpha$, and $\beta$ are Specified. The Statistician, 31, 245-250.

Ury, H.K. and Fleiss, J.L. (1980). On approximate sample sizes for comparing two independent proportions with the use of Yate's correction. Biometrics 36 , 347-351.

Walters, D. E. (1979). In Defence of the Arc Sine Approximation. The Statistician, 28, 219-222.

Yates, F. (1934). Contingency tables involving small numbers and the $X^2$-test. Journal of the Royal Statistical Society Suppl. 1, 217-235.

Yates, F. (1984). Tests of significance for 2x2 contingency tables. Journal of the Royal Statistical Society, A 147, 526-463.

TABLE A.1

SAMPLE SIZE $n=n_1=n_2$ REQUIRED TO OBTAIN A SPECIFIED POWER WHEN

$\alpha = 0.10$

| $P_2$ | .95 | .9 | .85 | .8 | .75 | .7 | .65 | .6 | .55 | .5 | .45 | .4 | .35 | .3 | .25 | .2 | .15 | .1 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| .85 | | | | | | | | | | | | | | | | | | |
| | 41 | | | | | | | | | | | | | | | | | |
| .8 | 45 | | | | | | | | | | | | | | | | | |
| | 22 | 47 | | | | | | | | | | | | | | | | |
| .75 | 41 | | | | | | | | | | | | | | | | | |
| | 30 | | | | | | | | | | | | | | | | | |
| | 16 | 24 | | | | | | | | | | | | | | | | |
| .7 | 30 | | | | | | | | | | | | | | | | | |
| | 22 | 39 | | | | | | | | | | | | | | | | |
| | 11 | 17 | 30 | | | | | | | | | | | | | | | |
| .65 | 22 | 38 | | | | | | | | | | | | | | | | |
| | 18 | 27 | 44 | | | | | | | | | | | | | | | |
| | 6 | 11 | 18 | 37 | | | | | | | | | | | | | | |
| .6 | 18 | 28 | 42 | | | | | | | | | | | | | | | |
| | 13 | 19 | 29 | 47 | | | | | | | | | | | | | | |
| | 5 | 9 | 13 | 19 | 35 | | | | | | | | | | | | | |
| .55 | 14 | 20 | 30 | 46 | | | | | | | | | | | | | | |
| | 10 | 16 | 21 | 31 | | | | | | | | | | | | | | |
| | 4 | 6 | 9 | 14 | 21 | 37 | | | | | | | | | | | | |
| .5 | 12 | 17 | 23 | 34 | 50 | | | | | | | | | | | | | |
| | 9 | 12 | 16 | 23 | 35 | | | | | | | | | | | | | |
| | 4 | 5 | 8 | 11 | 15 | 22 | 42 | | | | | | | | | | | |
| .45 | 10 | 14 | 19 | 25 | 35 | | | | | | | | | | | | | |
| | 8 | 9 | 14 | 17 | 24 | 36 | | | | | | | | | | | | |
| | 4 | 4 | 6 | 7 | 10 | 14 | 23 | 43 | | | | | | | | | | |
| .4 | 9 | 12 | 15 | 20 | 26 | 37 | | | | | | | | | | | | |
| | 7 | 9 | 10 | 14 | 20 | 25 | 36 | | | | | | | | | | | |
| | 4 | 4 | 5 | 6 | 7 | 12 | 14 | 23 | 43 | | | | | | | | | |
| .35 | 8 | 9 | 13 | 16 | 21 | 28 | 36 | | | | | | | | | | | |
| | 5 | 7 | 9 | 12 | 14 | 20 | 27 | 36 | | | | | | | | | | |
| | 4 | 4 | 4 | 5 | 6 | 7 | 12 | 14 | 23 | 42 | | | | | | | | |
| .3 | 6 | 9 | 10 | 13 | 16 | 21 | 28 | 37 | | | | | | | | | | |
| | 5 | 6 | 7 | 9 | 12 | 14 | 20 | 25 | 36 | | | | | | | | | |
| | 4 | 4 | 4 | 4 | 6 | 6 | 7 | 12 | 14 | 22 | 37 | | | | | | | |
| .25 | 6 | 8 | 9 | 11 | 13 | 16 | 21 | 26 | 35 | 50 | | | | | | | | |
| | 4 | 5 | 6 | 7 | 9 | 12 | 14 | 20 | 24 | 35 | | | | | | | | |
| | 2 | 4 | 4 | 4 | 4 | 6 | 6 | 7 | 10 | 15 | 21 | 35 | | | | | | |
| .2 | 5 | 6 | 7 | 9 | 11 | 13 | 16 | 20 | 25 | 34 | 46 | | | | | | | |
| | 4 | 5 | 6 | 6 | 7 | 9 | 12 | 14 | 17 | 23 | 31 | 47 | | | | | | |
| | 2 | 2 | 4 | 4 | 4 | 4 | 5 | 6 | 7 | 11 | 14 | 19 | 37 | | | | | |
| .15 | 4 | 6 | 6 | 7 | 9 | 10 | 13 | 15 | 19 | 23 | 30 | 42 | | | | | | |
| | 4 | 4 | 5 | 6 | 6 | 7 | 9 | 10 | 14 | 16 | 21 | 29 | 44 | | | | | |
| | 3 | 2 | 2 | 4 | 4 | 4 | 4 | 5 | 6 | 8 | 9 | 13 | 18 | 30 | | | | |
| .1 | 4 | 5 | 6 | 6 | 8 | 9 | 9 | 12 | 14 | 17 | 20 | 28 | 38 | | | | | |
| | 4 | 4 | 4 | 5 | 5 | 6 | 7 | 9 | 9 | 12 | 16 | 19 | 27 | 39 | | | | |
| | 3 | 3 | 2 | 2 | 4 | 4 | 4 | 4 | 4 | 5 | 6 | 9 | 11 | 17 | 24 | 47 | | |
| .05 | 4 | 4 | 4 | 5 | 6 | 6 | 8 | 9 | 10 | 12 | 14 | 18 | 22 | 30 | 41 | | | |
| | 2 | 4 | 4 | 4 | 4 | 5 | 5 | 7 | 8 | 9 | 10 | 13 | 18 | 22 | 30 | 45 | | |
| | 3 | 3 | 3 | 2 | 2 | 4 | 4 | 4 | 4 | 4 | 4 | 5 | 6 | 11 | 16 | 22 | 41 | |

Upper figure: power = $1 - \beta = 0.90$
Middle figure: power = $1 - \beta = 0.80$
Lower figure: power = $1 - \beta = 0.50$

TABLE A.2

SAMPLE SIZE $n=n_1=n_2$ REQUIRED TO OBTAIN A SPECIFIED POWER WHEN $\alpha = 0.05$

| $P_2$ | .95 | .9 | .85 | .8 | .75 | .7 | .65 | .6 | .55 | .5 | .45 | .4 | .35 | .3 | .25 | .2 | .15 | .1 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| .85 | | | | | | | | | | | | | | | | | | |
| | | | | | | | | | | | | | | | | | | |
| .8 | | | | | | | | | | | | | | | | | | |
| | 30 | | | | | | | | | | | | | | | | | |
| .75 | | | | | | | | | | | | | | | | | | |
| | 42 | | | | | | | | | | | | | | | | | |
| | 23 | 37 | | | | | | | | | | | | | | | | |
| .7 | 38 | | | | | | | | | | | | | | | | | |
| | 28 | 50 | | | | | | | | | | | | | | | | |
| | 15 | 25 | 46 | | | | | | | | | | | | | | | |
| .65 | 28 | 47 | | | | | | | | | | | | | | | | |
| | 23 | 35 | | | | | | | | | | | | | | | | |
| | 12 | 18 | 27 | 50 | | | | | | | | | | | | | | |
| .6 | 23 | 35 | | | | | | | | | | | | | | | | |
| | 18 | 26 | 39 | | | | | | | | | | | | | | | |
| | 8 | 12 | 19 | 29 | | | | | | | | | | | | | | |
| .55 | 19 | 27 | 39 | | | | | | | | | | | | | | | |
| | 14 | 19 | 28 | 44 | | | | | | | | | | | | | | |
| | 7 | 10 | 14 | 19 | 33 | | | | | | | | | | | | | |
| .5 | 16 | 21 | 29 | 43 | | | | | | | | | | | | | | |
| | 12 | 17 | 22 | 32 | 46 | | | | | | | | | | | | | |
| | 6 | 8 | 11 | 15 | 22 | 34 | | | | | | | | | | | | |
| .45 | 12 | 17 | 24 | 32 | 45 | | | | | | | | | | | | | |
| | 9 | 12 | 17 | 24 | 33 | 50 | | | | | | | | | | | | |
| | 6 | 7 | 9 | 11 | 16 | 23 | 37 | | | | | | | | | | | |
| .4 | 11 | 15 | 19 | 25 | 34 | 48 | | | | | | | | | | | | |
| | 8 | 11 | 14 | 19 | 25 | 35 | | | | | | | | | | | | |
| | 5 | 6 | 8 | 9 | 12 | 17 | 22 | 38 | | | | | | | | | | |
| .35 | 9 | 13 | 15 | 21 | 27 | 36 | 49 | | | | | | | | | | | |
| | 7 | 9 | 12 | 16 | 20 | 27 | 37 | | | | | | | | | | | |
| | 5 | 6 | 7 | 8 | 9 | 13 | 15 | 22 | 37 | | | | | | | | | |
| .3 | 8 | 10 | 14 | 17 | 21 | 28 | 36 | 48 | | | | | | | | | | |
| | 7 | 8 | 9 | 13 | 16 | 20 | 27 | 35 | 50 | | | | | | | | | |
| | 4 | 5 | 6 | 7 | 8 | 9 | 13 | 17 | 23 | 34 | | | | | | | | |
| .25 | 7 | 9 | 11 | 14 | 18 | 21 | 27 | 34 | 45 | | | | | | | | | |
| | 6 | 7 | 8 | 9 | 13 | 16 | 20 | 25 | 33 | 46 | | | | | | | | |
| | 2 | 4 | 4 | 6 | 7 | 8 | 9 | 12 | 16 | 22 | 33 | | | | | | | |
| .2 | 7 | 8 | 9 | 12 | 14 | 17 | 21 | 25 | 32 | 43 | | | | | | | | |
| | 5 | 7 | 8 | 8 | 9 | 13 | 16 | 19 | 24 | 32 | 44 | | | | | | | |
| | 2 | 2 | 4 | 4 | 6 | 7 | 8 | 9 | 11 | 15 | 19 | 29 | 50 | | | | | |
| .15 | 6 | 7 | 8 | 9 | 11 | 14 | 15 | 19 | 24 | 29 | 39 | | | | | | | |
| | 4 | 6 | 7 | 8 | 8 | 9 | 12 | 14 | 17 | 22 | 28 | 39 | | | | | | |
| | 3 | 2 | 2 | 4 | 4 | 6 | 7 | 8 | 9 | 11 | 14 | 19 | 27 | 46 | | | | |
| .1 | 5 | 6 | 7 | 8 | 9 | 10 | 13 | 15 | 17 | 21 | 27 | 35 | 47 | | | | | |
| | 4 | 4 | 6 | 7 | 7 | 8 | 9 | 11 | 12 | 17 | 19 | 26 | 35 | 50 | | | | |
| | 3 | 3 | 2 | 2 | 4 | 5 | 6 | 6 | 7 | 8 | 10 | 12 | 18 | 25 | 37 | | | |
| .05 | 4 | 5 | 6 | 7 | 7 | 8 | 9 | 11 | 12 | 16 | 19 | 23 | 28 | 38 | | | | |
| | 2 | 4 | 4 | 5 | 6 | 7 | 7 | 8 | 9 | 12 | 14 | 18 | 23 | 28 | 42 | | | |
| | 3 | 3 | 3 | 2 | 2 | 4 | 5 | 5 | 6 | 6 | 7 | 8 | 12 | 15 | 23 | 30 | | |

Upper figure: power $1 - \beta = 0.90$
Middle figure: power $1 - \beta = 0.80$
Lower figure: power $1 - \beta = 0.50$

TABLE A.3

SAMPLE SIZE $n=n_1=n_2$ REQUIRED TO OBTAIN A SPECIFIED POWER WHEN

$\alpha = 0.01$

| $P_2$ | .95 | .9 | .85 | .8 | .75 | .7 | .65 | .6 | .55 | .5 | .45 | .4 | .35 | .3 | .25 | .2 | .15 | .1 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| .85 | | | | | | | | | | | | | | | | | | |
| | | | | | | | | | | | | | | | | | | |
| .8 | | | | | | | | | | | | | | | | | | |
| | | | | | | | | | | | | | | | | | | |
| .75 | | | | | | | | | | | | | | | | | | |
| | 36 | | | | | | | | | | | | | | | | | |
| .7 | | | | | | | | | | | | | | | | | | |
| | 44 | | | | | | | | | | | | | | | | | |
| | 27 | 45 | | | | | | | | | | | | | | | | |
| .65 | 43 | | | | | | | | | | | | | | | | | |
| | 34 | | | | | | | | | | | | | | | | | |
| | 21 | 31 | | | | | | | | | | | | | | | | |
| .6 | 34 | | | | | | | | | | | | | | | | | |
| | 27 | 40 | | | | | | | | | | | | | | | | |
| | 15 | 23 | 36 | | | | | | | | | | | | | | | |
| .55 | 27 | 39 | | | | | | | | | | | | | | | | |
| | 23 | 31 | 46 | | | | | | | | | | | | | | | |
| | 13 | 19 | 26 | 38 | | | | | | | | | | | | | | |
| .5 | 23 | 34 | 45 | | | | | | | | | | | | | | | |
| | 19 | 24 | 35 | | | | | | | | | | | | | | | |
| | 11 | 16 | 21 | 28 | 42 | | | | | | | | | | | | | |
| .45 | 18 | 26 | 35 | 48 | | | | | | | | | | | | | | |
| | 15 | 20 | 27 | 37 | | | | | | | | | | | | | | |
| | 10 | 12 | 17 | 21 | 29 | 43 | | | | | | | | | | | | |
| .4 | 17 | 22 | 29 | 37 | 50 | | | | | | | | | | | | | |
| | 13 | 17 | 22 | 29 | 40 | | | | | | | | | | | | | |
| | 9 | 11 | 14 | 17 | 23 | 31 | 46 | | | | | | | | | | | |
| .35 | 14 | 18 | 24 | 31 | 40 | | | | | | | | | | | | | |
| | 11 | 15 | 19 | 24 | 31 | 40 | | | | | | | | | | | | |
| | 8 | 10 | 11 | 16 | 18 | 22 | 32 | 46 | | | | | | | | | | |
| .3 | 12 | 16 | 20 | 25 | 32 | 40 | | | | | | | | | | | | |
| | 10 | 12 | 16 | 20 | 25 | 32 | 40 | | | | | | | | | | | |
| | 7 | 9 | 10 | 11 | 16 | 17 | 22 | 31 | 43 | | | | | | | | | |
| .25 | 11 | 13 | 17 | 20 | 25 | 32 | 40 | 50 | | | | | | | | | | |
| | 10 | 11 | 13 | 18 | 20 | 25 | 31 | 40 | | | | | | | | | | |
| | 7 | 8 | 9 | 10 | 11 | 16 | 18 | 23 | 29 | 42 | | | | | | | | |
| .2 | 10 | 11 | 14 | 18 | 20 | 25 | 31 | 37 | 48 | | | | | | | | | |
| | 9 | 10 | 11 | 14 | 18 | 20 | 24 | 29 | 37 | | | | | | | | | |
| | 6 | 8 | 8 | 9 | 10 | 11 | 16 | 17 | 21 | 38 | 38 | | | | | | | |
| .15 | 9 | 10 | 11 | 14 | 17 | 20 | 24 | 29 | 35 | 45 | | | | | | | | |
| | 7 | 9 | 10 | 11 | 13 | 16 | 19 | 22 | 27 | 35 | 46 | | | | | | | |
| | 3 | 6 | 8 | 8 | 9 | 10 | 11 | 14 | 17 | 21 | 26 | 36 | | | | | | |
| .1 | 8 | 9 | 10 | 11 | 13 | 16 | 18 | 22 | 26 | 34 | 39 | | | | | | | |
| | 8 | 8 | 9 | 10 | 11 | 12 | 15 | 17 | 20 | 24 | 31 | 40 | | | | | | |
| | 4 | 3 | 6 | 8 | 8 | 9 | 10 | 11 | 12 | 16 | 19 | 23 | 31 | 45 | | | | |
| .05 | 8 | 8 | 9 | 10 | 11 | 12 | 14 | 17 | 18 | 23 | 27 | 34 | 43 | | | | | |
| | 6 | 8 | 7 | 9 | 10 | 10 | 11 | 13 | 15 | 19 | 23 | 27 | 34 | 44 | | | | |
| | 3 | 3 | 5 | 7 | 8 | 8 | 9 | 9 | 10 | 11 | 12 | 15 | 21 | 27 | 36 | | | |

Upper figure: power = $1 - \beta = 0.90$
Middle figure: power = $1 - \beta = 0.80$
Lower figure: power = $1 - \beta = 0.50$

# REPORT DOCUMENTATION PAGE

| | |
|---|---|
| 1a. REPORT SECURITY CLASSIFICATION<br>UNCLASSIFIED | 1b. RESTRICTIVE MARKINGS |
| 2a. SECURITY CLASSIFICATION AUTHORITY | 3. DISTRIBUTION/AVAILABILITY OF REPORT<br>UNLIMITED |
| 2b. DECLASSIFICATION/DOWNGRADING SCHEDULE | |
| 4. PERFORMING ORGANIZATION REPORT NUMBER(S) | 5. MONITORING ORGANIZATION REPORT NUMBER(S) |
| 6a. NAME OF PERFORMING ORGANIZATION<br>The University of North Carolina<br>Department of Biostatistics<br>6b. OFFICE SYMBOL (If applicable) | 7a. NAME OF MONITORING ORGANIZATION<br>Statistics and Probability Program |
| 6c. ADDRESS (City, State, and ZIP Code)<br>Chapel Hill, North Carolina 27514 | 7b. ADDRESS (City, State, and ZIP Code)<br>Office of Naval Research<br>Arlington, VA. 22217 |
| 8a. NAME OF FUNDING/SPONSORING ORGANIZATION<br>Office of Naval Research<br>8b. OFFICE SYMBOL (If applicable) | 9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER<br>N00014-83-0387 |
| 8c. ADDRESS (City, State, and ZIP Code)<br>Arlington, VA. 22217 | 10. SOURCE OF FUNDING NUMBERS<br>PROGRAM ELEMENT NO. / PROJECT NO. / TASK NO. / WORK UNIT ACCESSION NO. |

11. TITLE (Include Security Classification)
A Review of Sample Size Determination in Comparative Studies (Unclassified)

12. PERSONAL AUTHOR(S)
M.K. Habib, K.M. Magruder-Habib, and L.L. Kupper

| 13a. TYPE OF REPORT | 13b. TIME COVERED | 14. DATE OF REPORT (Year, Month, Day) | 15. PAGE COUNT |
|---|---|---|---|
| Technical | FROM _____ TO _____ | 1987 November 20 | 26 |

16. SUPPLEMENTARY NOTATION

| 17. COSATI CODES | | | 18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number) |
|---|---|---|---|
| FIELD | GROUP | SUB-GROUP | |
| | | | |
| | | | |

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

Sample size considerations for small and large sample designs of 2x2 comparative studies are reviewed. A test developed by Boschloo (1970) is recommended for small sample studies. This test is a simple modification of Fisher's unconditional test (FU) which is less conservative and hence more powerful than the FU test. New sample size tables for Boschloo's test are given in this paper. It is also argued that the uncorrected $X^2$-test is quite satisfactory for large sample comparative studies except under extreme situations, such as when compared proportions are near zero or one.

| | |
|---|---|
| 20. DISTRIBUTION/AVAILABILITY OF ABSTRACT<br>☐ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT. ☐ DTIC USERS | 21. ABSTRACT SECURITY CLASSIFICATION |
| 22a. NAME OF RESPONSIBLE INDIVIDUAL | 22b. TELEPHONE (Include Area Code) / 22c. OFFICE SYMBOL |