Revista Colombiana de Estadística
Nº 11 - 1985

# THE USE OF PILOT STUDY DATA IN THE ESTIMATION OF SAMPLE SIZE

*Bernard Roser*[*] *Alvaro Muñoz*[**]

**Sumary.** In the two sample binomial case, one approach to the estimation of sample size is to conduct a pilot study and assume that the observed proportion in the pilot study have no sampling error and are in fact true population parameters which can be used directly in standard sample size formulas. This approach has conceptual difficulties when such a pilot study is small since there is typically considerable

[*] Chaning Laboratory, Department of Preventive Medicine and Clinical Epidemiology, Harvard Medical School and Peter Bent Brigham Hospital. Division of Brigham and Women's Hospital, Boston, MA 02115 USA.

[**] Department of Epidemiology and Department of Biostatistic Johns Hopkins University School of Hygiene and Public Health Baltimore, MD 21205 USA.

error in the observed proportions. In this paper, we propose an alternative method which takes into account the sampling error in pilot study data in the estimation of sample size for a larger study. Tables are provided comparing these two methods and it is shown that the former deterministic method may provide a grossly inaccurate estimate of the appropriate sample size for a larger study, particularly for small pilot studies.

*Keywords:* Sample size; power curves; binomial distribution; bayesian inference; pilot study.

## 1. **Introduction.**

It is sometimes the case in planning large clinical studies to first conduct a pilot study for the purpose of (a) establishing the feasibility of a large study and (b) estimating the appropriate sample size for the large study in case the study is feasible. The idea of using pilot study data in the estimation of sample size has been discussed generally in Armitage (1973, p.187) and Hill (1977, p.286). In this paper, we focus more specifically on quantifying how to use pilot study data where the purpose of the large study is to compare proportions in two independent samples.

Specifically, we wish to test the hypothesis $H_o:p_o = p_1$ vs. $H_1:p_o \neq p_1$. Suppose we have obtained the sample proportions $\hat{p}_o = x_o/n_o$, $\hat{p}_1 = x_1/n_1$ for the control and treatment group respectively in the pilot study and wish to use $\hat{p}_o$, $\hat{p}_1$ to estimate the sample size in the large study. A widely used estimator (Snedecor and Cochran (1980, p.129)) for the appropriate sample size $N_o = N_1 = N$ for each group in the large study is given by the formula

$$N = (z_{\alpha/2}+z_{\beta})^2(p_oq_o+p_1q_1)/(p_o-p_1)^2 \qquad (1.1)$$

where $q_i = 1-p_i$, $i = 0,1$ and $z_p$ is the $100\times(1-p)$th percentile of a standard normal distribution. In practice, since the $p_i$'s are generally not known in advance, the investigador usually either (a) provides an educated guess as to their magnitude based on (i) previuos work or (ii) as assessment of what would constitute a meaningful therapeutic effect or (b) substitutes $\hat{p}_i$ for $p_i$ if a pilot study has been performed. If the pilot study is small in magnitude and the resulting standard errors of the $\hat{p}_i$ are large, then the latter approach can potentially lead to serious errors in the sample size estimates since the $\hat{p}_i$ will be poor estimates of the $p_i$. In Section 2, we present a more realistic method for using pilot study material to estimate sample size which takes account of (a) the estimation error in the $\hat{p}_i$, and (b) the prior information regarding (i) the underlying rate in the control group and (ii) the magnitude of what is considered a meaningful therapeutic effect. In Section 3, power comparisons are given comparing our estimates with those provided by (1.1). An example is given in Section 4 illustrating the use of these methods.

## 2. Theory.

We adopt a Bayesian approach to this problem. In particular, if one a priori expects $p_i$ to be $\pi_i$ and ranging from $m_i\pi_i$ to $M_i\pi_i$ then one may parametrize the prior distribution as a Beta with expected value $\pi_i$ and standard deviation $\sigma_i$ where $q\sigma_i = M_i\pi_i - m_i\pi_i$ for some pre-specified $q, m_i, M_i$, $i = 0,1$. One can interpret $\pi_1/\pi_o$ = Relative Risk as an expression of the expected therapeutic effect in comparing the treatment with the control group. In addition, $q$ is the number of standard deviations equal to the range of $p_i$. It can be easily shown that the parameters $a_i$, $b_i$ of the above Beta distributions are given by

$$\frac{a_i}{\pi_i} = \frac{b_i}{1-\pi_i} = \frac{q^2(1-\pi_i)}{(M_i-m_i)^2\pi_i} - 1, \qquad i = 0,1. \tag{2.1}$$

It then follows immediately from the properties of the binomial distribution that the posterior distribution of $p_i$ given the pilot study dada is:

$$g_i(p_i|\hat{p}_i) = p_i^{x_i+a_i-1}(1-p_i)^{n_i-x_i+b_i-1} \Big/ \int_{p_{i=0}}^{1} p_i^{x_i+a_i-1} \cdot (1-p_i)^{n_i-x_i+b_i-1}\, dp_i. \qquad i = 0,1 \tag{2.2}$$

i.e. $p_i$ given $\hat{p}_i$ follows a Beta distribution with parameters $x_i+a_i$ and $n_i-x_i+b_i$. If one uses standard power calculations for the two sample binomial problem, then for a specific $N$ and $\alpha$ the power $(\pi(N,\alpha|p_0,p_1))$ conditional on $p_0$ and $p_1$ can be expressed in the form:

$$\pi(N,\alpha|p_0,p_1) = \Phi(-z_{\alpha/2} + \Delta\sqrt{N} \,/\, \sqrt{p_0q_0+p_1q_1}\,) + \Phi(-z_{\alpha/2} - \Delta\sqrt{N} \,/\, \sqrt{p_0q_0+p_1q_1}\,) \tag{2.3}$$

where $\Delta = p_0 - p_1$. It then follows immediately from (2.2) and (2.3) that the expected posterior power $\lambda(N,\underset{\sim}{n},\underset{\sim}{\hat{p}},\alpha)$ is given by:

$$\lambda(N,\underset{\sim}{n},\underset{\sim}{\hat{p}},\alpha) = \int_{p_o=0}^{1}\int_{p_1=0}^{1} \pi(N,\alpha|p_o,p_1)g_o(p_o|\hat{p}_o)\cdot \cdot g_1(p_1|\hat{p}_1)dp_1 dp_o . \tag{2.4}$$

We will subsequently refer to $\lambda(N,\underset{\sim}{n},\underset{\sim}{\hat{p}},\alpha)$ as the "probabilistic" power in comparison to the "deterministic" power obtained from substituting $\hat{p}_o$ for $p_o$ and $\hat{p}_1$ for $p_1$ in (2.3) as follows:

$$\lambda^*(N,\underset{\sim}{\hat{p}},\alpha) = \Phi(-z_{\alpha/2}+\hat{\Delta}\sqrt{N}/\sqrt{\hat{p}_o\hat{q}_o+\hat{p}_1\hat{q}_1})$$

$$+ \Phi(-z_{\alpha/2}-\hat{\Delta}\sqrt{N}/\sqrt{\hat{p}_o\hat{q}_o+\hat{p}_1\hat{q}_1}) \tag{2.5}$$

where $\hat{\Delta} = \hat{p}_o-\hat{p}_1$. We note that $\lambda$ is a function of $N,\underset{\sim}{n},\underset{\sim}{\hat{p}}$ and $\alpha$ while $\lambda^*$ is only a function of $N,\underset{\sim}{\hat{p}}$, $\alpha$ since the deterministic power is not affected by the sample size used in the pilot study. It is of interest to note that the deterministic power is the probabilistic power for the case when the posterior distribution of $p_i$ has all its mass at $\hat{p}_i$.

## 3. Power studies.

In this section, we present the results of power studies for the case $\pi_o = .10$, $\pi_1 = .05$,

$m_i = \frac{1}{2}$, $M_i = 2$, $i = 0,1$, $q = 4$ (i.e., the expected value of $p_i$ is $\pi_i$ and four times the standard deviation of $p_i$ is equal to $2\pi_i - \pi_i/2$). Specifically, we evaluate $\lambda(N,\underset{\sim}{n},\underset{\sim}{\hat{p}},\alpha)$ in (2.4) and $\lambda^*(N,\underset{\sim}{\hat{p}},\alpha)$ in (2.5) for $n_1 = n_2 = 20,40$; $N = 200(200)1000$; $\hat{p}_0 = .05(.05).30$, $\hat{p}_1 = .05(.05)\hat{p}_0$, $\alpha = .05$. The IMSL double precision subroutines DLGAMMA, MDNORD and DBCODU (International Mathematical and Statistical Libraries, 1979) were used to compute the powers in (2.4) and (2.5). These powers are presented in Table 1.

## TABLE 1

Probabilistic and deterministic power for $N = 200(200)1000$, $\hat{p}_0 = .05(.05).30$, $\hat{p}_1 = .05(.05)\hat{p}_0$ where for the probabilistic power $n_0 = n_1 = 20$, 40 and the prior distribution for group $i$ is asssumed to be Beta with parameters $a_i$, $b_i$, $i = 0,1$*

| | | | N | | | | |
|---|---|---|---|---|---|---|---|
| $\hat{p}_0$ | $\hat{p}_1$ | | 200 | 400 | 600 | 800 | 1000 |
| 0.05 | 0.05 | prob20+ | 0.42 | 0.57 | 0.65 | 0.70 | 0.73 |
| | | prob40 | 0.36 | 0.51 | 0.60 | 0.65 | 0.69 |
| | | det | 0.05 | 0.05 | 0.05 | 0.05 | 0.05 |
| 0.10 | 0.05 | prob20 | 0.49 | 0.65 | 0.73 | 0.77 | 0.80 |
| | | prob40 | 0.49 | 0.66 | 0.73 | 0.78 | 0.81 |
| | | det | 0.48 | 0.77 | 0.91 | 0.97 | 0.99 |
| 0.10 | 0.10 | prob20 | 0.42 | 0.58 | 0.66 | 0.71 | 0.74 |
| | | prob40 | 0.37 | 0.52 | 0.61 | 0.66 | 0.70 |
| | | det | 0.05 | 0.05 | 0.05 | 0.05 | 0.05 |

| | | | *N* | | | | |
|---|---|---|---|---|---|---|---|
| $\hat{p}_0$ | $\hat{p}_1$ | | 200 | 400 | 600 | 700 | 1000 |
| 0.15 | 0.05 | prob20 | 0.56 | 0.72 | 0.79 | 0.83 | 0.85 |
| | | prob40 | 0.61 | 0.78 | 0.84 | 0.87 | 0.89 |
| | | det | 0.92 | 1.00 | 1.00 | 1.00 | 1.00 |
| 0.15 | 0.10 | prob20 | 0.49 | 0.65 | 0.73 | 0.77 | 0.80 |
| | | prob40 | 0.49 | 0.66 | 0.73 | 0.78 | 0.81 |
| | | det | 0.33 | 0.57 | 0.75 | 0.86 | 0.92 |
| 0.15 | 0.15 | prob20 | 0.43 | 0.59 | 0.66 | 0.71 | 0.74 |
| | | prob40 | 0.38 | 0.53 | 0.62 | 0.67 | 0.71 |
| | | det | 0.05 | 0.05 | 0.05 | 0.05 | 0.05 |
| 0.20 | 0.05 | prob20 | 0.63 | 0.78 | 0.84 | 0.87 | 0.89 |
| | | prob40 | 0.72 | 0.86 | 0.91 | 0.93 | 0.95 |
| | | det | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| 0.20 | 0.10 | prob20 | 0.56 | 0.72 | 0.79 | 0.83 | 0.85 |
| | | prob40 | 0.61 | 0.77 | 0.84 | 0.87 | 0.89 |
| | | det | 0.81 | 0.98 | 1.00 | 1.00 | 1.00 |
| 0.20 | 0.15 | prob20 | 0.50 | 0.66 | 0.73 | 0.77 | 0.80 |
| | | prob40 | 0.49 | 0.66 | 0.74 | 0.78 | 0.81 |
| | | det | 0.26 | 0.46 | 0.63 | 0.75 | 0.84 |
| 0.20 | 0.20 | prob20 | 0.44 | 0.59 | 0.67 | 0.72 | 0.75 |
| | | prob40 | 0.39 | 0.55 | 0.63 | 0.68 | 0.72 |
| | | det | 0.05 | 0.05 | 0.05 | 0.05 | 0.05 |
| 0.25 | 0.05 | prob20 | 0.69 | 0.84 | 0.89 | 0.91 | 0.93 |
| | | prob40 | 0.81 | 0.92 | 0.95 | 0.97 | 0.97 |
| | | det | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| 0.25 | 0.10 | prob20 | 0.63 | 0.78 | 0.84 | 0.87 | 0.89 |
| | | prob40 | 0.71 | 0.86 | 0.91 | 0.93 | 0.94 |
| | | det | 0.98 | 1.00 | 1.00 | 1.00 | 1.00 |
| 0.25 | 0.15 | prob20 | 0.56 | 0.72 | 0.79 | 0.83 | 0.85 |
| | | prob20 | 0.61 | 0.77 | 0.83 | 0.87 | 0.89 |
| | | det | 0.71 | 0.95 | 0.99 | 1.00 | 1.00 |
| 0.15 | 0.20 | prob20 | 0.50 | 0.66 | 0.74 | 0.78 | 0.81 |
| | | prob40 | 0.50 | 0.67 | 0.74 | 0.79 | 0.82 |
| | | det | 0.22 | 0.40 | 0.55 | 0.67 | 0.76 |
| 0.25 | 0.25 | prob20 | 0.44 | 0.60 | 0.68 | 0.73 | 0.76 |
| | | prob40 | 0.40 | 0.56 | 0.64 | 0.69 | 0.73 |
| | | det | 0.05 | 0.05 | 0.05 | 0.05 | 0.05 |

| | | | $N$ | | | | |
|---|---|---|---|---|---|---|---|
| $\hat{p}_0$ | $\hat{p}_1$ | | 200 | 300 | 400 | 600 | 1000 |
| 0.30 | 0.05 | prob20 | 0.75 | 0.88 | 0.92 | 0.94 | 0.95 |
| | | prob40 | 0.88 | 0.96 | 0.98 | 0.98 | 0.99 |
| | | det | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| 0.30 | 0.10 | prob20 | 0.69 | 0.83 | 0.88 | 0.91 | 0.92 |
| | | prob20 | 0.80 | 0.92 | 0.95 | 0.96 | 0.97 |
| | | det | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| 0.30 | 0.15 | prob20 | 0.63 | 0.78 | 0.84 | 0.87 | 0.89 |
| | | prob40 | 0.71 | 0.85 | 0.90 | 0.93 | 0.94 |
| | | det | 0.95 | 1.00 | 1.00 | 1.00 | 1.00 |
| 0.30 | 0.20 | prob20 | 0.57 | 0.73 | 0.79 | 0.83 | 0.85 |
| | | prob40 | 0.61 | 0.77 | 0.84 | 0.87 | 0.89 |
| | | det | 0.64 | 0.91 | 0.98 | 1.00 | 1.00 |
| 0.30 | 0.25 | prob20 | 0.51 | 0.67 | 0.74 | 0.78 | 0.81 |
| | | prob40 | 0.51 | 0.68 | 0.75 | 0.79 | 0.82 |
| | | det | 0.20 | 0.35 | 0.49 | 0.61 | 0.71 |
| 0.30 | 0.30 | prob20 | 0.45 | 0.61 | 0.69 | 0.73 | 0.76 |
| | | prob40 | 0.42 | 0.58 | 0.66 | 0.71 | 0.74 |
| | | det | 0.05 | 0.05 | 0.05 | 0.05 | 0.05 |

* $a_i/\pi_i = b_i(1-\pi_i) = q^2(1-\pi_i)/\{(M_i-m_i)^2\pi_i\}$, $i = 0,1.$

$m_i = \frac{1}{2}$, $M_i = 2$, $i = 0,1$, $q = 4$, $\pi_o = 0.10$, $\pi_1 = 0.05$.

** prob20 = probabilistic power for $n_o = n_1 = n = 20$;

prob40 = probabilistic power for $n_o = n_1 = n = 40$;

det = deterministic power.

For a given $\hat{p}_o$, $N$, the deterministic power increases much more rapidly with increasing $\hat{p}_o - \hat{p}_1$ than either probabilistic power. The differences between the deterministic and probabilistic powers are dramatic in most instances. In particular, if $\hat{p}_o/\hat{p}_1 \geqslant \pi_o/\pi_1$, then the deterministic power is usually much larger than either probabilistic power. For example, if $\hat{p}_o = .10 = \pi_o$, $\hat{p}_1 = .05 = \pi_1$, $N = 1000$, then the deterministic power is 0.99 while the probabilistic powers are 0.80 and 0.81 for $n = 20$ and 40 respectively. Conversely, if $\hat{p}_o/\hat{p}_1 \ll \pi_o/\pi_1$, then the deterministic power is generally smaller than either probabilistic power. For example, if $\hat{p}_o = .20$, $\hat{p}_1 = .15$, $N = 400$, then the deterministic power if 0.46 while the probabilistic powers are 0.66 for both $n = 20$ and 40 respectively. Thus, if one uses the deterministic sample size method in (1.1) to assess the appropriate sample size to achieve a prespecified power, then one will tend to underestimate the appropriate sample size for large $\hat{p}_o/\hat{p}_1$ and overestimate the appropriate sample size for small $\hat{p}_o/\hat{p}_1$.

In general, the probabilistic powers are similar for $n = 20$ and 40. If $\hat{p}_o > \hat{p}_1$, then the probabilistic power for $n = 40$ is usually larger than for $n = 20$, while if $\hat{p}_o = \hat{p}_1$, then the opposite is true with the difference in the two powers being $\lesssim 0.13$ in all instances and $\lesssim 0.05$ in most instantes. The difference in the assessments of power using the deterministic and probabilistic methods were similar to those given above when $\underset{\sim}{\pi}, N, \underset{\sim}{n}$ and $\underset{\sim}{\hat{p}}$ were held fixed and $\underset{\sim}{m}$, $\underset{\sim}{M}$ and $q$ were allowed to vary.

## 4. Example.

An area of recent investigation has been the ascertainment of risk factors during the prenatal period to identity women destined to deliver low birthweight (<2500 gm) infants. One risk factor which has been extensively studied in low income populations is the presence of U. Urealyticum obtained from a vaginal culture at the first prenatal visit (Kass et al. (1981)). A clinical trial is planned in a prenatal clinic attended by women of a higher socioeconomic status based on women who are initially positive for U. Urealyticum. The investigators plan to randomize an equal number of women to (a) erythromycin, a drug which is expected to eliminate U. Urealyticum and reduce the proportion of low birthweight infants or (b) placebo. Since the prevalence rate of low birthweight varies inversely with socieconomic status, the results of the previous study could not be used to help estimate sample size in the proposed study and a pilot study was conducted for this purpose. The results of the pilot study indicated that of twenty women who were treated with erythromycin only one delivered a low birthweight infant ($\hat{p}_1 = 0.05$), while of twenty women who were treated with placebo two delivered a low birthweight infant ($\hat{p}_o = 0.10$).

We see from (2.5) that if we use the deterministic power calculation method, then we would need to recruit 430 pregnant women in each group in the large study in order to have an 80% chance of detecting a significant difference using a two-sided test with $\alpha = .05$. However, if we compute the probabilistic power from (2.4) using the same prior parameters as in Table 1, then we would require 1000 pregnant women in each group in order to achieve the same type I and type II error. Similar results would be obtained if 40 pregnant women obtained for each group in the pilot study and $\hat{p}_o$ and $\hat{p}_1$ are maintained at .10 and .05 respectively. Thus, we see that the deterministic method grossly underestimates the appropriate sample size in this case.

In contrast, suppose that 40 pregnant women had been obtained for each group in the pilot study with four of the placebo women ($\hat{p}_o = 0.10$) and three of the erythromycin women ($\hat{p}_1 = 0.075$) delivering a low birthweight infant. If one uses the same prior parameters as in Table 1 and the same type I and type II errors as above, then from (2.4) and (2.5) it follows that one would need approximately 1400 women in each group in the large study using the probabilistic method and 1550 women in each group using the deterministic method. Thus, in this case the deterministic method overestimates the appropriate sample size.

## 5. Discussion.

In this paper, a method is provided for using pilot study data in the estimation of sample size. This method enables one to combine prior information concerning (a) the underlying disease rate in the control group ($\pi_o$) and (b) the relative risk of disease in the treatment group as compared with the control group ($\pi_1/\pi_o$) with the results of the pilot study to obtain sample size estimates for a prespecified type I and type II error. These "probabilistic" sample size estimates have been compared with classical deterministic sample size estimates. The deterministic method generally underestimates the appropriate sample size for a prespecified type I and type II error for the larger proposed study of $\hat{p}_o/\hat{p}_1 \geqslant \pi_o/\pi_1$ and generally overestimates the appropriate sample size if $\hat{p}_o/\hat{p}_1 < \pi_o/\pi_1$. There is often at least a twofold difference in the sample size estimates using the two methods to achieve a given type I and type II error. Thus, this should make one cautious about using the deterministic sample size method based on pilot study material especially when the number of subjects studied in the pilot study is small.

In many instances, a large body of previously collected data exists which is similar enough to the proposed large study so that one can use this material directly for purposes of sample size estimation. In other instances, one will be able to specify what would be a meaningful therapeutic effect even in the absence of previous data directly applicable to the proposed study. In both cases, determistic sample size estimates obtained from inverting (2.5) for a prespecified $\alpha$ and $\beta$ are typically employed. In contrast, if no large body of previous data exists and in addition it is difficult to specify the magnitude of what woul constitute a meaningful therapeutic effect, then a small pilot study many be employed to augment existing information. In this case, deterministic sample size estimates can be be misleading and probabilistic sample size estimates obtained by inverting (2.4) for a prespecified $\alpha$ and $\beta$ will be most useful. On the other hand, it should be noted that if the size of the pilot study is very large as would effectively be the case if a large body of previously collected data exists and we regard this as a "pilot" study, then the probabilistic and deterministic sample size estimates will be very similar since the posterior distribution $g_i(p_i|\hat{p}_i)$ will be nearly the same as a distribution whose entire probability mass is at $\hat{p}_i$. Similarly, the closer $m_i$ and $M_i$ are (e.g., $M_i = 1.05$, $m_i = 0.95$) the closer $g_i(p_i|\hat{p}_i)$ will be to the distribucion whose entire probability mass is at $\pi_i$. Thus the methodology present ed here is flexible enough to permit the assessment of the sample size to be based mainly on the consideration of the minimal therapeutic effect $(\pi_i/\pi_o)$ which would be thought to be clinically significant.

*

## BIBLIOGRAFIA

Armitage, P., (1973). *Statistical Methods in Medical Research*. London: Wiley.

Hill, A.B., (1977). *A Short Textbook of Medical Statistic*, 10th ed. Philadelphia: J.B. Lippincott Company.

International Mathematical and Statistical Libraries (1979). *IMSL Reference Manual*, Vol. 2, 7th ed. Houston, IMSL.

Kass, E.H., Mccormack, W.M., Lin, J.S. and Rosner, B., (1981). Genital mycoplasmas as a hitherto unsuspected cause of excess premature delivery in the underpriveleged. *Clinical Research*, 29, 575a.

Snedecor, G. and Cochran, W.G., (1980). *Statistical Methods*, 7th ed. Ames, Iowa: Iowa State University Press.

* *