Ann Ist Super Sanità 2004;40(2):173-179

Accounting for historical information in designing experiments: the Bayesian approach

Fulvio DE SANTIS and Marco PERONE PACIFICO

Dipartimento di Statistica, Probabilità e Statistiche Applicate, Università degli Studi "La Sapienza", Rome, Italy

Summary. - Two of the most important statistical problems in human and animal experimentation are the selection of an appropriate number of units to include in a given study and the allocation of these units to the various treatments. Properly addressing these issues allows the units to be used as efficiently as possible, which can contribute to addressing the overall issue of reducing the number of subjects in experimentation. To do so, reliable historical information is of particular importance. In the present paper, we describe the Bayesian approach to determining sample size and allocating units, with particular regard to how the use of historical data can be optimised. The paper focuses on two specific problems: the comparison of normal means and of binomial proportions.

Key words: allocation, Bayesian design, credible interval, clinical trials, interval estimation, sample size, testing.

Riassunto *(Uso di dati storici nei disegni sperimentali: l'approccio Bayesiano).* - Tra i principali problemi statistici che sorgono nella sperimentazione umana e animale, vi sono la scelta del numero di unità statistiche da includere nello studio e l'assegnazione dei diversi trattamenti a ciascuna unità. In questi studi è infatti di particolare importanza, sia dal punto di vista etico che economico, la limitazione del numero di soggetti da coinvolgere. Per questo motivo risulta rilevante l'uso efficiente dell'informazione pre-sperimentale sul fenomeno oggetto di studio. Impostare tali problemi in un'ottica Bayesiana consente di formalizzare e sfruttare l'informazione contenuta in dati provenienti da studi precedenti. Il presente lavoro rivisita alcuni metodi Bayesiani proposti in letteratura sul tema della scelta della numerosità campionaria e della assegnazione di trattamento, con particolare attenzione a due casi rilevanti: il confronto tra medie di popolazioni normali e tra proporzioni in modelli binomiali.

Parole chiave: assegnazione di trattamento, disegni sperimentali Bayesiani, insiemi di credibilità, numerosità campionaria, prove cliniche, stima per intervallo, test.

Introduction and motivations

In planning human and animal studies on the relative effectiveness of a novel treatment, compared to an established treatment, several statistical issues need to be addressed. Two important basic issues are the number of statistical units and their allocation to the treatment groups (e.g., treatment and control). Properly addressing these issues allows the units to be used as efficiently as possible, which can contribute to reducing the number of subjects necessary for performing an experiment. To do so, reliable information on both the novel treatment and the control treatment, which generally consists of historical information from previous studies, must be available. This information can contribute not only to reducing the overall size of an experiment but also to efficiently allocating the experimental units, with more individuals assigned to the novel treatment, for which it is assumed that less information is available.

Historical information can be formalised and incorporated into an analysis by adopting the Bayesian approach. For instance, in comparing two unknown parameters (θ_1 and θ_2) that represent the mean effectiveness of two treatments, formalising the information on these quantities through probability distributions has two immediate advantages. The first advantage is practical: assigning a prior distribution to the unknown quantities allows different plausible scenarios to be taken into consideration. Technically speaking, this allows local optimality to be avoided (see examples below). Moreover, in allocating units, the Bayesian approach allows for the use of flexible rules, which reflect the actual knowledge on the

Indirizzo per la corrispondenza (Address for correspondence): Fulvio De Santis, Dipartimento di Statistica, Università degli Studi "La Sapienza", P.le Aldo Moro 5, 00185 Rome, Italy. E-mail: fulvio.desantis@uniroma1.it.

phenomenon before performing the experiment. The second main advantage to the Bayesian approach is that it addresses additional unknown quantities that are not of direct scientific interest (i.e., "nuisance parameters"), such as the parameters that measure the variability of the data. These points are illustrated in the examples below.

Example 1. - Suppose that we are interested in evaluating the relative effectiveness of a novel treatment, compared to a standard treatment. If formalising the problem as an interval estimation of the difference in the means of independent normal random variables with equal unknown variances, σ^2, it can be easily determined that the width of the 1-α confidence interval based on two independent samples of sizes n_1 and n_2 (see, for instance, [1]) is

$$2t_{n-2;1-\frac{\alpha}{2}} S \sqrt{\frac{1}{n_1}+\frac{1}{n_2}},$$

where S is the pooled standard deviation, $n = n_1 + n_2$, and $t_{n\text{-}2;1\text{-}\frac{\alpha}{2}}$ is the $1-\frac{\alpha}{2}$ percentile of the Student distribution with n - 2 degrees of freedom. Note that in planning the analysis (i.e., when the data are yet to be obtained), the above quantity depends on the random variable S. To determine n_1 and n_2, the standard procedure (see, for example, [2]) is to require the expected width of the random interval to be less than a chosen threshold, ℓ. However, since the expected value of S depends on the unknown value of σ, a guess value of this nuisance parameter must be chosen to select values for n_1 and n_2.

Example 2. - Suppose that we are interested in comparing two probabilities, the unknown parameters of two independent binomial distributions. Let θ_1 and θ_2 denote these unknown parameters, with the former referring to the established treatment and the latter referring to the novel treatment. In particular, suppose that we want to estimate the unknown log-odds-ratio $\log\left(\frac{\theta_2}{1-\theta_2}\Big/\frac{\theta_1}{1-\theta_1}\right)$ using the standard 1 - α confidence interval based on two independent samples whose sizes are indicated by n_1 and n_2 . Suppose also that the objective is that of selecting the size of the sample so as to have a narrow interval. The most commonly used frequentist approach [3] is to choose the minimal sample size so that the width of the confidence interval is no greater than ℓ. Since the interval's width depends on the unknown parameters (θ_1, θ_2), the criterion requires preliminary estimates $\tilde{\theta}_1, \tilde{\theta}_2$. Denoting with $z_{1-\frac{\alpha}{2}}$ the $1-\frac{\alpha}{2}$ percentile of the standard normal distribution, the resulting total sample size, $n = n_1 + n_2$, is

$$n = \frac{4z^2_{1-\frac{\alpha}{2}}}{\ell^2}\left(\frac{1}{\frac{n_2}{n}\tilde{\theta}_2(1-\tilde{\theta}_2)}+\frac{1}{(1-\frac{n_2}{n})\tilde{\theta}_1(1-\tilde{\theta}_1)}\right)$$

where the optimal proportion of cases

$$\frac{n_2}{n} = \left(1+\sqrt{\frac{\tilde{\theta}_2(1-\tilde{\theta}_2)}{\tilde{\theta}_1(1-\tilde{\theta}_1)}}\right)^{-1}$$

is obtained minimising the asymptotic variance of the maximum likelihood estimator (see [4]). If the observed proportions match the initial estimates $(\tilde{\theta}_1, \tilde{\theta}_2)$, then the width of the confidence interval would be ℓ. Clearly, inaccurate preliminary estimates could lead to excessively wide confidence intervals, which is the typical "local optimality" problem of procedures for determining the standard sample size.

The above two examples reveal that, even in the simplest settings, the standard procedures for determining the sample size are only locally optimal. In the following sections, the Bayesian approach to determining the sample size is outlined, and the use of historical information to model uncertainty regarding the unknown parameters is demonstrated (see Example 3). The paper is organised as follows. In the following section, the Bayesian approach to determining sample size is described, and the criteria cited throughout the paper are introduced, distinguishing between estimation criteria and testing criteria. Then we focus on the problem of allocating units. Finally, the proposed criteria are applied to the standard problem of comparing normal means and to determining sample size for inference in binomial experiments.

Bayesian criteria for determining sample size and allocating units

We consider experiments in which two treatments are compared by observing two independent samples, $\mathbf{X}_{n_1}$ and $\mathbf{X}_{n_2}$. Denoting with $f(\cdot|\,\theta_j)$ the density function of $\mathbf{X}_{n_j}$ (j = 1,2), the posterior density of parameter θ_j relative to the observed sample $\mathbf{x}_{n_j}$ is

$$\pi(\theta_j \mid \mathbf{x}_{n_j}) = \frac{f(\mathbf{x}_{n_j} \mid \theta_j)\pi(\theta_j)}{m_{n_j}(\mathbf{x}_{n_j})} \qquad (1)$$

where π is the prior density of θ_j and

$$m_{n_j}(\mathbf{x}_{n_j}) = \int_{\Theta_j} f(\mathbf{x}_{n_j} \mid \theta_j)\pi(\theta_j)\, d\theta_j \qquad (2)$$

is the predictive density of $\mathbf{x}_{nj}$. If nuisance parameters are present, the posterior density of the parameter of interest can be obtained by integrating the joint posterior distribution.

In this paper, we discuss how to determine the size of the two samples, n_1 and n_2, for inference on a parameter $\delta = \delta(\theta_1, \theta_2)$, a function of the parameters of the two sampling distributions. The underlying concept in determining the sample size is that, before

performing the experiment that will yield the data, we want to choose the minimal sample size that satisfies a selected criterion. Thus, unlike standard Bayesian posterior analysis, the randomness of the data must be taken into account. Specifically, all of the posterior quantities considered are random, given that they are a function of $\mathbf{X}_{n_1}$ and $\mathbf{X}_{n_2}$, which are independent random variables, with the distribution of each obtained with (2). In comparative trials, determining the sample size is a two-fold problem, that is, it is necessary to choose both the total size, n, and the number of individuals allocated to the two treatment groups (n_1 and n_2). In the following subsections, we address these issues for both estimation problems and hypothesis testing (for additional reading on the Bayesian approach to the issue of sample size, see [5-8]).

Criteria for estimation problems

For a given sample ($\mathbf{X}_{n_1} = \mathbf{x}_{n_1}$, $\mathbf{X}_{n_2} = \mathbf{x}_{n_2}$) the two most common Bayesian interval estimates for the parameter of interest are equal-tails intervals and highest posterior density sets (HPD). The $1-\alpha$ level equal-tails interval has as its endpoints the $\frac{\alpha}{2}$ and $1-\frac{\alpha}{2}$ quantiles of the posterior distribution. The $1-\alpha$ level HPD set is the subset of the parameter space with the highest posterior density and a posterior probability greater than or equal to $1-\alpha$. Each of these estimates has advantages and disadvantages to its use. For instance, HPD sets have the shortest length of all of the sets of level $1-\alpha$, yet they are not invariant under parametric transformation and are computationally difficult to derive. Equal-tails intervals are easy to derive, yet they are not of minimal width. Below, we focus on equal-tails intervals, given that they are easier to derive than HPD sets and are more commonly used. Furthermore, for larger samples, the posterior distribution is often approximately normal, hence equal-tails intervals coincide with HPD and are thus optimal.

To determine the sample size, based on the features of the interval estimates of the parameter of interest, two criteria are considered: Average Length Criterion (ALC) [9] and Length Probability Criterion (LPC) [9, 10].

Denoting with $L(\mathbf{X}_{n_1}, \mathbf{X}_{n_2})$ the random width of the Bayesian $1-\alpha$ level interval estimate of δ, when using ALC one looks for the smallest sample size so that the expected width of the interval is not greater than a chosen threshold, ℓ

$$\mathbf{E}[L(\mathbf{X}_{n_1}, \mathbf{X}_{n_2})] \leq \ell \qquad (3)$$

the expectation being with respect to the joint predictive distribution of ($\mathbf{X}_{n_1}$, $\mathbf{X}_{n_2}$) obtained from (2). The rationale behind ALC is that of avoiding wide interval estimates, which are uninformative. However, ALC does not control the sampling variability in the random width of the intervals. To take variability into account, LPC can be used: the smallest sample size is selected so that the probability of having an estimated interval whose width is greater than or equal to a given threshold, ℓ, is limited by a chosen level, $\gamma \in (0,1)$:

$$\mathbf{P}[L(\mathbf{X}_{n_1}, \mathbf{X}_{n_2}) \geq \ell] \leq \gamma \qquad (4)$$

A criterion for hypothesis testing

For testing two alternative hypotheses,

$$H_0 : \delta \in \Delta_0 \quad \text{vs.} \quad H_1 : \delta \in \Delta_1 = \Delta_0^c$$

we propose selecting the sample size using the criterion proposed by Verdinelli [11] and related to criteria introduced by Weiss [12] in the Bayesian framework (see also Royall [13], who has developed a theory of statistical evidence based on the likelihood function). The basic idea is that data are collected to produce substantial evidence in favour of either H_0 or H_1. Unless a formal decision theoretic approach is adopted, in the Bayesian scenario, one of the two hypotheses is selected based on their posterior probabilities

$$\Pi(H_j \mid \mathbf{x}_{n_1}, \mathbf{x}_{n_2}) = \int_{\Delta_j} \pi(\delta \mid \mathbf{x}_{n_1}, \mathbf{x}_{n_2})\, d\delta \qquad j = 0,1$$

where $\pi(\delta \mid \mathbf{x}_{n_1}, \mathbf{x}_{n_2})$ is the posterior density of δ. If $\Pi(H_j \mid \mathbf{x}_{n_1}, \mathbf{x}_{n_2}) \geq \rho$ where $\rho \in (\frac{1}{2}, 1)$ is a chosen probability level, then there is strong evidence in favour of H_j, whereas if $\Pi(H_0 \mid \mathbf{x}_{n_1}, \mathbf{x}_{n_2}) \in (1-\rho, \rho)$ neither H_0 nor H_1 are strongly supported by the observed data. The criterion for choosing the sample size thus consists of considering the smallest size such that the probability that neither H_0 nor H_1 are strongly supported is less than a certain threshold, γ. Therefore, we choose the smallest sample size such that

$$\mathbf{P}[1-\rho \leq \Pi(H_0 \mid \mathbf{x}_{n_1}, \mathbf{x}_{n_2}) \leq \rho] \leq \gamma \qquad (5)$$

Allocation of units

In sample size problems, ALC, LPC and the above-described test criterion each provides a condition sufficient for determining sample size. However, the problem considered here also requires choosing the number of units to be assigned to each treatment (n_1 and n_2). Since conditions (3), (4), and (5) for determining sample size depend on both n_1 and n_2, the criteria must be integrated to obtain a single value for each of the two sample sizes.

For each possible total sample size, n, the optimal allocation (n_1, n_2) is that which minimises the probability (4) of having a wide interval estimate or, if the test criterion is adopted, the probability (5) of not having substantial evidence in favour of either hypothesis. As demonstrated in Section 3, this criterion can be used, for instance, for the normal model with conjugate prior distribution. Nonetheless, in the general case, such an allocation may not exist, or it could be computationally intensive.

When prior information on one of the two treatments is more accurate than the information on the other treatment, it may be a good idea to use the following alternative allocation rule: choose n_1 and n_2 so that the posterior uncertainty regarding the parameters in the two populations is approximately equal. This usually results in fewer units being assigned to the better known treatment. A means of imposing equal posterior uncertainty is that of requiring the expectation of the posterior variances of θ_1 and θ_2 to be equal. Thus, for a given total size n, one can choose n_1 as the solution to the following equation (as n_1 varies in $\{1,\cdots,n-1\}$)

$$\mathbf{E}[Var(\theta_1 \mid \mathbf{X}_{n_1})] = \mathbf{E}[Var(\theta_2 \mid \mathbf{X}_{n-n_1})] \qquad (6)$$

If the root of equation (6) is not an integer, choose $n_1 \in \{1,\cdots, n-1\}$ so that the left-hand-side and the right-hand-side are as close as possible.

The following section focuses on situations in which the expected posterior variance has a closed form, thus allowing the allocation criterion (6) to be easily adopted.

Two relevant problems

Choosing the sample size for normal and binomial experiments is an important starting point for gaining insight into more complex cases and different models.

Difference of normal means

Suppose that the effect of each treatment on the units can be modelled as a normal random variable. Hence, each sample $\mathbf{X}_{n_j}$ (j = 1,2) is drawn from a normal distribution with an unknown mean μ_1 and unknown precision $\lambda = \sigma^{-2}$ (the precision is supposed to be the same for the two samples). We consider the usual conjugate prior for the parameters (μ_1, μ_2, λ) a gamma distribution with fixed parameters (ν, β) for the precision λ and, conditionally on λ, a normal distribution for each mean μ_j with prior mean μ_{0j} and precision $n_{0j}\lambda$. Using, as in Bernardo and Smith [14], $\mathrm{N}(\cdot \mid \mu, \lambda)$ and $\mathrm{Ga}(\cdot \mid \nu, \beta)$ to denote normal and gamma density, respectively, the joint prior for the three parameters is

$$\pi(\mu_1,\mu_2,\lambda) = \mathrm{N}(\mu_1 \mid \mu_{01}, n_{01}\lambda)\ \mathrm{N}(\mu_2 \mid \mu_{02}, n_{02}\lambda)\ \mathrm{Ga}(\lambda \mid \nu, \beta)$$

The parameter of interest here is $\delta = \mu_1 - \mu_2$. The above prior assumptions are equivalent to assigning

$$\pi(\delta,\lambda) = \mathrm{N}(\delta \mid \mu_{01} - \mu_{02}, (n_{01}^{-1} + n_{02}^{-1})^{-1}\lambda)\ \mathrm{Ga}(\lambda \mid \nu, \beta)$$

The posterior distribution of δ is known (see, for instance, [14]). De Santis and Perone Pacifico [10] provide the quantities needed to choose the sample size according to ALC and LPC

$$\mathbf{E}[L(\mathbf{X}_{n_1}, \mathbf{X}_{n_2})] = k \cdot \frac{\Gamma(\nu + \frac{n_1+n_2}{2})\Gamma(\nu - \frac{1}{2})}{\Gamma(\nu)\Gamma(\nu - \frac{1}{2} + \frac{n_1+n_2}{2})}$$

$$\mathbf{P}[L(\mathbf{X}_{n_1}, \mathbf{X}_{n_2}) \geq \ell] = \int_0^{(k/\ell)^2} \mathrm{Be}(t \mid \nu, \tfrac{n_1+n_2}{2})\, dt$$

where $\mathrm{Be}(\cdot \mid \nu, \frac{n}{2})$ is the Beta density with parameters $(\nu, \frac{n}{2})$ and

$$k = 2t_{2\nu+n_1+n_2;1-\frac{\alpha}{2}} \cdot \sqrt{\frac{2\beta(n_1 + n_{01} + n_2 + n_{02})}{(n_1 + n_2 + 2\nu)(n_1 + n_{01})(n_2 + n_{02})}}$$

Note that the distribution and expectation of $L(\mathbf{X}_{n_1}$ and $\mathbf{X}_{n_2})$ do not depend on the prior means μ_{01} and μ_{02}. Moreover, it is easy to verify that the probability in (4) depends on the allocation of n in n_1 and n_2 based only on the quantity k. Since k is minimal when $n_1 + n_{01} = n_2 + n_{02}$, choosing

$$n_1 = \frac{n + n_{02} - n_{01}}{2} \quad \text{and} \quad n_2 = \frac{n + n_{01} - n_{02}}{2}$$

results in the optimal allocation of units to the treatments. Since n_{0j} is usually interpreted as the strength of prior information on μ_j, then $n_j + n_{0j}$ can be considered as the amount of posterior information on the same parameter. Hence the allocation that requires $n_1 + n_{01} = n_2 + n_{02}$ in some sense tends to balance the uncertainty regarding the two treatments.

Taking into account the above constraints, the criteria in (3) and (4) depend on the total size n only.

One can easily use LPC (or ALC) by computing $\mathbf{P}[L(\mathbf{X}_{n_1}, \mathbf{X}_{n_2}) \geq \ell]$ (or $\mathbf{E}[L(\mathbf{X}_{n_1}, \mathbf{X}_{n_2})]$) for several values of n.

Regarding the impact of the strength of prior information on the above criteria, it can be shown that, for any fixed value of α and of γ (for LPC), the "prior" sample size and the sample size of the new experiment play a concurring role: the stronger the prior information (i.e., the larger $n_{01} + n_{02}$), the smaller the number of new observations needed to satisfy either ALC or LPC. Therefore, the more accurate the prior information, the lower the number of experimental units necessary for a new study.

Let us now turn to the test criterion and consider the following one-sided testing problem:

$$H_0 : \delta \le \delta_0 \quad \text{vs} \quad H_1 : \delta > \delta_0$$

Assume, for simplicity, that the sampling precision, λ, is known and that

$$\pi(\delta) = \mathrm{N}(\delta \mid \delta_0, (n_{01}^{-1} + n_{02}^{-1})^{-1} \lambda)$$

where δ_0 is the value that separates the null and the alternative hypotheses. Under the above assumptions

$$\mathbf{P}[1\text{-}\rho \le \Pi(H_0 \mid \mathbf{x}_{n_1}, \mathbf{x}_{n_2}) \le \rho] = \Phi\left(\xi(n_1, n_2) \cdot z_{\frac{\rho \Pi_0}{(1-\rho)(1-\Pi_0)+\rho\Pi_0}} \right) - \Phi\left(\xi(n_1, n_2) \cdot z_{\frac{(1-\rho)\Pi_0}{(1-\rho)\Pi_0+\rho(1-\Pi_0)}} \right) \quad (7)$$

where z_α and $\Phi(\alpha)$ are, respectively, the α percentile and the cumulative distribution function of the standard normal, Π_0 is the prior probability of the null hypothesis and where

$$\xi(n_1, n_2) = \sqrt{\frac{n_1 + n_2}{n_1 n_2} \frac{n_{01} n_{02}}{n_{01} + n_{02}}}$$

The probability (7) does not depend on λ and is a decreasing function of the quantity $\xi(n_1, n_2)$, which measures the relative strength of the prior information compared to the experimental information (for instance, if $n_1 = n_2 = n/2$ and $n_{01} = n_{02} = n_0/2$, then $\xi(n_1, n_2) = \sqrt{n_0 / n}$).

Therefore, when using the test criterion, the "prior" sample size and the sample size of the new experiment play a contrasting role: to achieve a certain level in the predictive probability of obtaining weak evidence, the smaller the prior sample size, the smaller the number of new units required. This is of course in contrast with what has been seen for ALC and LPC. This has to do with the fact that, in the testing criterion, the effect of the observations is two-fold: they contribute both to reducing posterior variance and shifting the posterior distribution of the parameter away from the prior mean. Therefore, the smaller the prior precision, the smaller the number of new observations needed to shift the posterior distribution and to provide decisive evidence in favour of either the null or the alternative hypothesis.

Two comments are in order. First, note that the above results can be easily extended by considering for δ a prior mean different from δ_0, yet obtaining closed-form formulas for the probability of weak evidence. Second, more realistic situations in which analytical results are not available (unknown sampling precision, non-conjugate priors) can be easily addressed by resorting to a simulation scheme to approximate the probabilities of obtaining weak evidence. For details, see De Santis [15].

Binomial experiments

Suppose now that the response to the treatment of each unit is binary (e.g., positive or negative); thus $\mathbf{X}_{n_j}$ denotes the number of positive responses among the n_j units assigned to treatment j (with $j = 0,1$). Under independence assumptions, the two samples $\mathbf{X}_{n_1}$ and have $\mathbf{X}_{n_2}$ independent binomial distributions with parameters (n_1,θ_1) and (n_2,θ_2), respectively. The two unknown parameters θ_1 and θ_2 denote the probability of success of each treatment. The parameter of interest is a function $\delta = \delta(\theta_1, \theta_2)$ often the odds-ratio $\frac{\theta_2}{1-\theta_2} \Big/ \frac{\theta_1}{1-\theta_1}$ or its logarithmic transformation.

We consider the conjugate priors for the unknown parameters θ_1 and θ_2 to be independent and thus assume that each θ_j has a beta distribution as a beta with fixed hyperparameters (α_j, β_j); the corresponding posterior density of θ_j relative to the observed value $\mathbf{x}_{n_j}$ is still beta with parameters $(\alpha_j + \mathbf{x}_{n_j}, \beta_j + n_j - \mathbf{x}_{n_j})$. Prior independence implies posterior independence of θ_1 and θ_2 as well as independence of the predictive distribution of $\mathbf{X}_{n_1}$ and $\mathbf{X}_{n_2}$. From standard conjugate analysis (see, for instance, [14]) it follows that the predictive distribution of $\mathbf{X}_{n_j}$ is binomial-beta with parameters (α_j, β_j, n_j):

$$m_{n_j}(\mathbf{x}_{n_j}) = \binom{n_j}{\mathbf{x}_{n_j}} \frac{B(\alpha_j + \mathbf{x}_{n_j}, \beta_j + n_j - \mathbf{x}_{n_j})}{B(\alpha_j, \beta_j)} \qquad \mathbf{x}_{n_j} = 0, \cdots, n_j \quad (8)$$

where B denotes the standard beta function. Under these assumptions, the expected posterior variance of θ_j is

$$\mathbf{E}[Var(\theta_j \mid \mathbf{x}_{n_j})] = \frac{\alpha_j \beta_j}{(\alpha_j + \beta_j)(\alpha_j + \beta_j + 1)(\alpha_j + \beta_j + n_j)} \quad (9)$$

In addition to computational convenience, there are several other reasons for considering beta prior distributions for the unknown parameters. First of all, the class of beta priors is often rich enough to represent a wide range of pre-experimental information on the unknown proportions θ_1 and θ_2. Moreover, in our problem, it is only necessary to specify the hyperparameters (α_j, β_j), which can be interpreted in a straightforward manner, as shown in Example 3.

Although we assume prior independence essentially to simplify computations, there may be other justifications for doing so. For example, as pointed out by Joseph, du Berger, and Belisle [16], when the optimal sample size selected using independent priors is relatively large with respect to the weight of prior information, switching to dependent priors does not have a strong impact and the priors are expected to have a limited effect on the analysis; however, for small or moderate sample sizes, Bayesian criteria are significantly affected by prior dependence. The sensitivity of the optimal

sample size to prior assumptions is an important issue, and we plan to elaborate on it in future studies.

In choosing the sample size according to LPC, the sampling distribution of $L(\mathbf{X}_{n_1}, \mathbf{X}_{n_2})$ needs to be evaluated, given that the two samples vary in range according to their independent binomial-beta distributions. To determine whether or not a candidate value n satisfies (4), the following steps are used:

i) using allocation criterion (6), compute, for $j = 1,2$, the expected posterior variances (9) and choose n_1 that minimises their absolute difference, let $n_2 = n - n_1$;

ii) compute $L(\mathbf{x}_{n_1}, \mathbf{x}_{n_2})$, for all $\mathbf{x}_{n_1} = 0,\cdots,n_1$ and $\mathbf{x}_{n_2} = 0,\cdots,n_2$;

iii) select all the pairs $(\mathbf{x}_{n_1}, \mathbf{x}_{n_2})$ such that $L(\mathbf{x}_{n_1}, \mathbf{x}_{n_2}) \le \ell$ and compute their probabilities $m_{n_1}(\mathbf{x}_{n_1}) \cdot m_{n_2}(\mathbf{x}_{n_2})$ using (8); if the sum of the probabilities is not greater than γ, then n satisfies (4).

Computing $L(\mathbf{x}_{n_1}, \mathbf{x}_{n_2})$ in step (ii) may be time consuming. When the parameter of interest is the logarithm of the odds-ratio and the two sample sizes are reasonably high, one can use the normal approximation to the posterior distribution of δ, as in De Santis, Perone Pacifico and Sambucini [17]. For more general cases, when an analytic expression for $L(\mathbf{x}_{n_1}, \mathbf{x}_{n_2})$ is not available, for each given pair $(\mathbf{x}_{n_1}, \mathbf{x}_{n_2})$ the equal-tails set can be obtained numerically as follows:

a) generate a large sample from the joint posterior distribution of (θ_1, θ_2);

b) compute the corresponding values of $\delta(\theta_1, \theta_2)$;

c) order the sample of δ's and discard a fraction of $\alpha/2$ values on each side;

d) the equal-tails set is the interval containing all the retained values in the sample of δ's.

Test criterion (5) can be implemented with a similar procedure, computing the quantity $\Pi(H_0 \mid \mathbf{x}_{n_1}, \mathbf{x}_{n_2})$ instead of $L(\mathbf{x}_{n_1}, \mathbf{x}_{n_2})$ in step (ii).

Example 3. - This example was taken from De Santis, Perone Pacifico and Sambucini [17] (Section 5), where data from a previous case-control study were used to choose the sample size for a new experiment. Although referring to a different context, the example provides an idea of the difference between the frequentist and the Bayesian approaches to determining sample size in binomial experiments.

Denoting with $\tilde{\mathbf{x}}_{n_j}$ and $\tilde{n}_j$ the corresponding values, in the previous experiment, of the quantities $\mathbf{x}_{n_j}$ and n_j, the historical data used to fix the prior parameters α_j and β_j are

$$\tilde{\mathbf{x}}_{n_1} = 81, \quad \tilde{n}_1 = 741, \qquad \tilde{\mathbf{x}}_{n_2} = 61, \quad \tilde{n}_2 = 404. \tag{10}$$

Following the standard elicitation techniques in conjugate analysis described in Bernardo and Smith [14], for the binomial model, $\alpha_j/(\alpha_j + \beta_j)$ should be chosen as the prior estimate of θ_j and $\alpha_j + \beta_j$ as the strength of the prior information on θ_j (number of observations to which the prior information is considered equivalent). Relying on the historical data (10),

$$\frac{\alpha_j}{\alpha_j + \beta_j} = \frac{\tilde{\mathbf{x}}_{n_j}}{\tilde{n}_j}$$

provides the prior estimate of the proportions θ_j. Regarding the strength of the prior information, the same proportion between the two groups has been kept and, as an example, a total strength $\tilde{s} = 60$ observations has been considered, so that

$$\frac{\alpha_1 + \beta_1}{\alpha_2 + \beta_2} = \frac{\tilde{n}_1}{\tilde{n}_2} \quad \text{and} \quad (\alpha_1 + \beta_1) + (\alpha_2 + \beta_2) = \tilde{s} = 60$$

The resulting values of the hyperparameters are given by the total strength times the corresponding relative frequencies in the contingency table

$$\alpha_1 = \tilde{s}\,\frac{\tilde{\mathbf{x}}_{n_1}}{\tilde{n}_1 + \tilde{n}_2} = 4.24, \quad \beta_1 = \tilde{s}\,\frac{\tilde{n}_1 - \tilde{\mathbf{x}}_{n_1}}{\tilde{n}_1 + \tilde{n}_2} = 34.59,$$
$$\alpha_2 = \tilde{s}\,\frac{\tilde{\mathbf{x}}_{n_2}}{\tilde{n}_1 + \tilde{n}_2} = 3.20, \quad \beta_2 = \tilde{s}\,\frac{\tilde{n}_2 - \tilde{\mathbf{x}}_{n_2}}{\tilde{n}_1 + \tilde{n}_2} = 17.97.$$

LPC has been implemented using the numerical procedure described above with $\ell = 1.5$ and $\gamma = 0.05$ and considering the normal approximation to the density of the log-odds-ratio. The minimal sample size that ensures $P[L\mathbf{X}_{n_1},\mathbf{X}_{n_2}) \ge 1.5] \le 0.05$ is ($n_1 = 188$, $n_2 = 271$). The frequentist procedure reported in Example 2, with the same desired width $\ell = 1.5$ and the same prior information given in (10), requires ($n_1 = 132$, $n_2 = 114$).

Unlike the results obtained using LPC, the frequentist procedure requires that there be more observations for Population 1 than for Population 2, given that this procedure is only based on the proportion of positive responses in the two groups, whereas the Bayesian approach also takes into account the total amount of information regarding each group ($\tilde{n}_1$, $\tilde{n}_2$). Since ($\tilde{n}_1 = 741$, $\tilde{n}_2 = 404$) prior information on Population 1 is more accurate; thus LPC tends to balance such information, requiring a higher number of new observations for Population 2.

Table 1 shows the behaviour of the LPC sample size as the strength of prior information increases from 30 to 90.

Whereas the frequentist optimal proportion and total size are not affected by the choice of the prior strength, with LPC the difference between the prior information on the two groups increases. Hence an increasing proportion of subjects from Population 2 is needed. Moreover, an increase in prior strength from 30 to 90 corresponds to a strong decrease in the sample size needed to satisfy LPC. Thus when using the Bayesian criterion, it is crucial to establish the reliability of the data set used for prior elicitation.

Table 1. - Sample size obtained using LPC for different values of the prior total strength

Prior strength	n_1	n_2	$\frac{n_2}{n_1+n_2}$
30	446	578	0.564
40	296	398	0.573
50	229	318	0.581
60	188	271	0.590
70	159	239	0.600
80	137	217	0.613
90	120	199	0.624

It should be mentioned that the frequentist procedure, although requiring a lower number of observations, could fail to prevent wide confidence intervals from being obtained: if the true proportions (θ_1, θ_2) differ significantly from their preliminary estimates ($\tilde{\theta}_1$, $\tilde{\theta}_2$), the resulting confidence interval could be quite wide. LPC does not suffer from this inconvenience, since it takes into account the variability of prior estimates. In particular, the frequentist optimal total sample size could be obtained through LPC with accurate prior information ($\tilde{S}$ large): in fact, increasing the total strength corresponds to concentrating the prior distribution of (θ_1, θ_2) on its mean ($\tilde{\theta}_1$, $\tilde{\theta}_2$) and thus to increasingly relying on the initial guess.

Acknowledgements

We wish to thank Maria Puopolo for inviting us to partecipate to this editorial series published within the Project no. 01/C "Biostatistical and ethological approaches for the promotion of welfare of laboratory animals and of the quality of esperimental data", to MP.

This research was partially supported by Progetto Giovani Ricercatori "La Sapienza" and by Cofinanziamento MIUR.

Submitted on invitation.
Accepted on 3 March 2004.

REFERENCES

1. Armitage P, Berry G, Matthews JNS. *Statistical methods in medical research*. (4th edition). Oxford: Blackwell Science; 2003.

2. Beal SL. Sample size determination for confidence intervals on the population mean and on the difference between two population means. *Biometrics* 1989;45:969-77.

3. O'Neill RT. Sample sizes for estimation of the odds ratio in unmatched case-control studies. *Am J Epidemiol* 1984;120:145-53.

4. Walter SD. Determination of significant relative risks and optimal sampling procedures in prospective and retrospective comparative studies of various sizes. *Am J Epidemiol* 1977;105:387-97.

5. Chaloner K, Verdinelli I. Bayesian experimental design: a review. *Statistical Science* 1995;10:237-304.

6. Piccinato L. *Metodi per le decisioni statistiche*. Milano: Springer-Verlag Italia; 1996.

7. Adcock J. Sample size determination: a review. *The Statistician* 1997;46:261-83.

8. Lindley DV. The choice of sample size. *The Statistician* 1997;46:129-38.

9. Joseph L, Wolfson D, du Berger R. Sample size calculation for binomial proportions via highest posterior density intervals. *The Statistician* 1995;44:143-54.

10. De Santis F, Perone Pacifico M. Two experimental settings in clinical trials: predictive criteria for choosing the sample size in interval estimation. In: *Applied Bayesian statistical studies in biology and medicine*. Boston: Kluwer; 2003.

11. Verdinelli I. *Bayesian design of experiments for the linear model*. (Ph. D. Thesis). Pittsburgh, PA: Department of Statistics, Carnegie Mellon University; 1996.

12. Weiss R. Bayesian sample size calculations for hypothesis testing. *The Statistician* 1997;46:209-26.

13. Royall RM. On the probability of observing misleading statistical evidence (with discussion). J Am Statistical Assoc 200; 95:760-80.

14. Bernardo JM, Smith AFM. *Bayesian theory*. New York: Wiley;1994.

15. De Santis F. Statistical evidence and sample size determination for Bayesian hypothesis testing. *J Statistical Plann Inference* (2004, in press).

16. Joseph L, du Berger R, Belisle P. Bayesian and mixed Bayesian/likelihood criteria for sample size determination. *Statistics Med* 1997;16:769-81.

17. De Santis F, Perone Pacifico M, Sambucini V. Optimal predictive sample size for case-control studies. *Journal of the Royal Statistical Society* Series C (2004, in press).