Sample Sizes To Estimate Proportions And Correlation

Brian A R de Melo, Raony C C Cesar and Carlos A B Pereira

Institute of Mathematics and Statistics - University of São Paulo

Abstract. Very often statisticians have to deal with the problem of sample size calculation. However this is one of the most difficult questions to be answered since it depends on many considerations, assumptions and restrictions on the real problem being solved. In this paper we develop some methods, based on credible intervals, to calculate sample sizes for proportions and correlations.

Keywords: bayesian inference, sample sizes, proportions, correlation
PACS: 02.50.-r

INTRODUCTION

It is not rare we, statisticians, have to answer the following question: "What sample size must we consider?". The sample size is an important topic and it is influenced by several factors such as cost of sampling, time needed to collect, ethical issues (in the case of research with humans or animals), among others. Furthermore, when performing the calculation of sample size for an experiment, it is necessary to consider the error that the researcher is willing to take, since there is always the possibility of errors working in statistics.

In this paper, the interest is the calculation of sample sizes for some situations commonly found in several studies, using Bayesian inference techniques. For this we consider the length of the $(1-\alpha)100\%$ credible interval as a criterion for the accuracy of estimates of the proportion. In the second and third sections we develope some methods to caluculate sample sizes when the main interest is to make inference about one or two proportions, respectively. Section four shows a method to calculate a sample size when we want to estimate the correlation coefficient.

ESTIMATING A PROPORTION

Whereas the researcher's interest is to make inference about a proportion, we can consider two approaches: finite population or superpopulation. When we know the size N of the population of interest, we use techniques of finite population, and consider the superpopulation when N is unknown.

Superpopulation

Let X be a random variable following a binomial distribution of parameters n and p, ie $X|(n,p) \sim Bin(n,p)$ and suppose, the prior $p \sim Beta(a,b)$. Therefore, the posterior distribution will be $p|(X=x,n) \sim Beta(A,B)$, for $A = a+x$ and $B = b+n+x$. The worst situation to make inference about the parameter p is obtained with a non-informative prior distribution, $a = b = 1$, and a symmetric posterior distribution, $A = B$. In this case, the variance of the posterior will be the largest possible (for a sample of size n) and hence the length of the HPD credible interval is also the largest possible.

Fixing the credibility level, $1-\alpha$, and length, l, of the credible interval, we can calculate the value of n that will produce a credible interval of length less than or equal to l, after the value $X = x$ is observed. This idea is used to calculate the sample size for a situation in which we sampled n individuals and we want to make inferences about the proportion of people who have certain characteristic C. Note that this is a conservative method, because we consider the largest sample satisfying the constraints, ie, it is guaranteed that, for any sample with size n obtained, the length of the credible interval for p will be smaller than or equal to the length l previously calculated. For example, if we set the credibility as $1-\alpha = 0.95$ and we want a interval with length smaller than $l = 0.1$ we must take a sample of $n = 382$ units, if we consider a uniform prior.

Finite Population

Consider a population A consisting of N units, that is, $A = \{1,2,\ldots,N\}$. Our interest now is to estimate the number θ of units with certain characteristic C of interest. Clearly, θ is a positive integer with unknown value. For the i-th populational unit we associate a variable U_i that takes the value 1 if the unit i has the characteristic C and 0 if it does not have the feature. Thus, the vector of population values is represented by $\psi = (U_1,\ldots,U_N)$ and $\theta = U_1 + \ldots + U_N$. It is natural to assume that changing the order of the units does not alter the expected distribution of the vector ψ. We therefore consider that the distribution of ψ is exchangeable, ie,

$$P(U_1 = u_1,\ldots,U_N = u_N) = P(U_{v_1} = u_1,\ldots,U_{v_n} = u_n), \tag{1}$$

$(v_1,\ldots,v_n)$ is a permutation of $(1,\ldots,N)$ and $(u_1,\ldots,u_N)$ is a vector whose elements are either equal to zero or one .

Under the condition of exchangeability we can always say that the vector ψ is a finite observation of an exchangeable process of zeros and ones. With this basic restriction and, by the representation theorem of de Finetti (1937), there exists a constant $0 < p < 1$ such that, conditional on p, the sequence of random variables $U_1, U_2, \ldots$ is a Bernoulli process with common success probability eual to p. Thus, it is natural to assume that $\theta|p$ follows a binomial distribution with parameters (N,p).

In order to estimate the parameter of interest θ, we consider a sample of size n of the population A. Without loss of generality we consider that this sample is composed of the n first individuals of the population. Denoting by $X = U_1 + \ldots + U_n$ the total sample,

we have a consequence of the theorem of De Finnet: $P(X = x, \theta - X = k|p) = Pr(X = x|p)Pr(\theta - X = k|p)$. That is, X and $\theta - X$ are conditionally independent, given p.

Remember that p is just a nuisance parameter, that came in for the convenience of simplify the theory developed here.

Observing the value $X = x$ in the sample, the parameter of interest is now the amount $\theta - X$. With the same notation of the superpopulation case we have, if $p \sim Beta(a,b)$ then $p|X = x \sim Beta(A,B)$. As our interest is in the distribution of $(\theta - X)|X$ we use the distribution $(\theta - X)|(X,p) \sim (\theta - X)|p \sim Bin(N-n,p)$. Finally we eliminate the nuisance parameter p by calculating the predictive distribution of $\theta - X$ through the posterior $Beta(A,B)$, reaching the following results: $(\theta - X)|X \sim BetaBin(N-n,A,B)$. The mean and variance of this distribution are given by:

$$E(\theta - X|X) = (N-n)e_n \text{ and } Var(\theta - X|X) = \frac{(N-n)(N+n_0)}{n+n_0+1}e_n(1-e_n),$$

for $n_0 = a+b$ and $e_n = \frac{A}{A+B}$ is the mean of the posterior distribution of $p|X = x$.

Again we note that the case with largest interval occurs when $e_n = 0.5$. Consider now a new parameter defined by $\delta = \theta/N$. Let us focus on this ratio to establish the length of a credible interval for δ that does not exceed a predetermined value, say $2\varepsilon = 0.1$. Note that the posterior mean of δ is e_n and its posterior variance is:

$$v_n = Var(\delta|X = x) = \frac{(N-n)(N+n_0)e_n(1-e_n)}{(n+n_0+1)N^2} \leq \frac{(N-n)(N+n_0)}{4(n+n_0+1)N^2} \tag{2}$$

Now we will seek an interval (I_1, I_2) such that $P(I_1 \leq \delta \leq I_2|X = x) = 0.95$. That is, we find the value of t such that $I_1 = e_n - td$ and $I_2 = e_n + td$. Here, t is a multiplier of the standard deviation, d, of the posterior distribution of δ. Setting the standard deviation value of the parameter δ, we find the value of n necessary to reach the limit set. For example, suppose we fix $td = 2d = 0.05$. In this case, given $d = 0.025$ and $a = b = 1$ for $N = 5000$, we would have $n = 394$. Therefore, to calculate the required sample size we only need to set the precision ε and determine the parameters of the prior a and b, since the population size N is known.

As an example, suppose we wish to consider a population of $N = 50$ community banks. Consider a sample of $n = 15$ banks to study the amount of banks that meet a certain characteristic C. In our case in particular we find $X = 5$. We conclude here that $\theta_0 = \{10, 11, ..., 25\}$ is a set with 90.69% of credibility for θ. That is:

$$P(\theta \in \theta_0) = P(0.2 < \delta < 0.5) = 0.9069 \tag{3}$$

ESTIMATION AND COMPARISON OF TWO PROPORTIONS

Difference of two proportions

Let X and Y be two independent random variables such that $X|p_1 \sim Bin(n,p_1)$ and $Y|p_2 \sim Bin(m,p_2)$. In this situation, we may be interested in knowing whether the odds of success in the two samples is the same, in which case we can consider the difference between the two proportions $P = p_1 - p_2$.

First we consider the prior distributions: $p_i \sim Beta(a_i, b_i)$, $i = 1, 2$. Thus the posterior distributions will be: $p_1|(n,x) \sim Beta(A_1, B_1)$ and $p_2|(m,y) \sim Beta(A_2, B_2)$, for $A_1 = a_1 + x$, $B_1 = b_1 + n - x$, $A_2 = a_2 + y$ and $B_2 = b_2 + m - y$. Considering the difference between the proportions we can build, through simulation procedures, credible intervals for P and establish a relationship between the length of the interval and the sample sizes, n and m.

Simulation

To build the simulation procedure we consider that we have no information about any of the two ratios we are interested in estimating. Thus, we have a priori $p_i \sim Beta(1, 1)$, $i = 1, 2$. The simulation is then performed as follows:

- Simulate $X \sim Bin(n, 0.5)$ and $Y \sim Bin(m, 0.5)$, obtaining the values $X = x$ and $Y = y$;
- Using the values of x and y we simulate a sample of size ten thousand of the posterior distribution of p_1 and p_2, $p_1|(n,x) \sim Beta(A_1, B_1)$ and $p_2|(m,y) \sim Beta(A_2, B_2)$;
- With these samples we generate a sample of the proportion's difference, $P = p_1 - p_2$;
- We built an HPD interval with 95% of credibility and calculate the length of this interval.

This is done a thousand times for each pair $(n,m) \in \{1, \ldots, 100\}^2$. Note that the probability equal to 0.5 is used to simulate the binomial distributions because in this case the variance of the binomial distribution is as high as possible, and the probability that the Beta distribution (posterior distribution) is symmetric (and thus have greater variance) is higher. Thus is obtained a relation between the length of the credible interval and the pair of sample sizes (n,m).

In Figure 1 are given the surface that shows the maximum lengths of the credible interval as a function of the samples sizes n and m and the contour of the surfaces for values of n and m between 1 and 20 in which we can see that, for example, if we want the length of the credible interval to be smaller or equal to 0.7 we consider $n = m = 13$ or $n = 20$ and $m = 10$. We also noticed that, for a fixed length, the total sample size $(n + m)$ is smaller when the values of n and m are close.

Logistic Normal model

In this section we propose a new approach for estimating the size of the samples, considering the logarithm of the odds ratio, $LO = log\frac{p_1(1-p_2)}{p_2(1-p_1)}$ instead of the difference between the two proportions.

We know that $\log\frac{p}{1-p}$ is approximately $Normal(\psi(a) - \psi(b), \psi'(a) + \psi'(b))$, for $p \sim Beta(a,b)$, $\psi(.)$ is the digamma function and $\psi'(.)$ is the trigamma function.

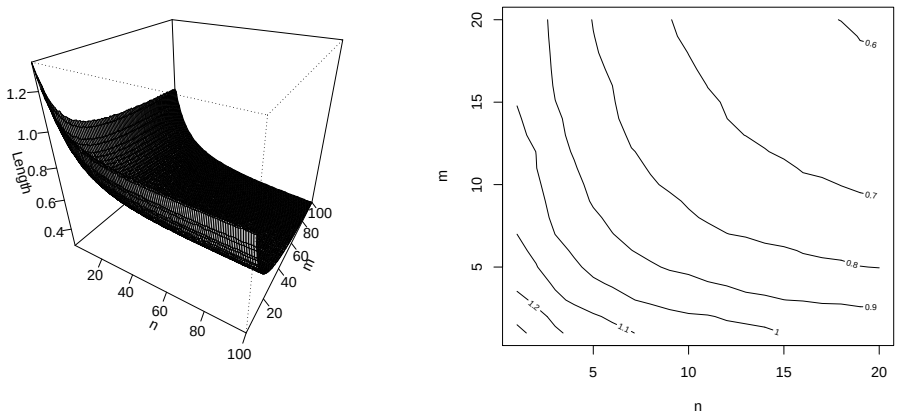

FIGURE 1. Length of the 95% credible interval for P depending on the sample sizes n e m and their contours.

For more details on this approach see Aitchison & Shen (1980) and Pereira & Stern (2008). The distribution of LO will then be approximately $Normal(\mu, \sigma^2)$, with $\mu = \psi(A_1) - \psi(B_1) - \psi(A_2) + \psi(B_2)$ e $\sigma^2 = \psi'(A_1) + \psi'(B_1) + \psi'(A_2) + \psi'(B_2)$.

Let $[\mu \pm z_{1-\frac{\alpha}{2}}\sigma]$ be the credible interval with $(1-\alpha)100\%$ of credibility for a random variable $X \sim Normal(\mu, \sigma^2)$, then the length of the interval is $2z_{1-\frac{\alpha}{2}}\sigma$. That is, when we consider the logarithm of the odds ratio, the length of the credible interval depends only on the variance, which is a function of A_i and B_i, $i = 1,2$. For fixed values of m and n we construct all possible samples and, as in the case of one proportion, we consider the maximum variance of the posterior distribution to ensure that the established accuracy is achieved after the sample is observed.

Figure 2 illustrates the lengths of the credible intervals obtained by using uniform prior distributions, for different sample sizes n and m and the contour. We note that the behavior of the intervals are very similar to those of Figure 1, except for the amount of lengths. That is because the logarithm of the odds ratio can take, theoretically, any value on the real line, but the difference between the proportions assumes values only in the interval $[-1,1]$. Table 1 illustrates the behavior of the length of the credible intervals for a uniform priori and some informative prior distributions, considering the same size of the two groups. By setting a value of $n = m = 20$ we obtain a maximum interval of size 7.215 using the log odds, whereas when we consider the difference between the proportions the maximum length is equal to 0.587, for the uniform priors. We also note that the size of the intervals decreases when we use informative priors. For example, the size of the interval of the log odds, with $n = m = 20$, goes from 7.215 to $1,762$ if we consider $a_1 = 10$, $b_1 = 30$ and $a_2 = b_2 = 20$ instead of a uniform prior.

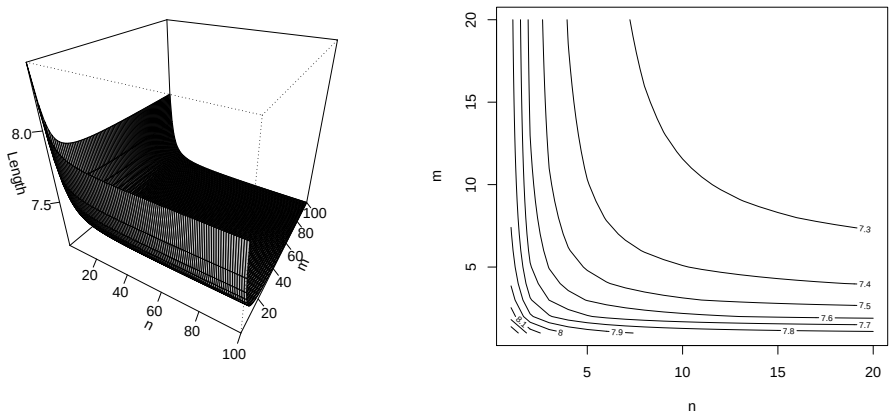

FIGURE 2. Length of the 95% credible interval for *LO* depending on the sample sizes n e m and their contours.

TABLE 1. Sample sizes and length of the 95% credible intervals for several prior distributions

Sample	**Difference of proportions**			**Log odds**		
n=m	a_1=b_1=1 a_2=b_2=1	a_1=4 b_1=8 a_2=b_2=6	a_1=10 b_1=30 a_2=b_2=20	a_1=b_1=1 a_2=b_2=1	a_1=4 b_1=8 a_2=b_2=6	a_1=10 b_1=30 a_2=b_2=20
10	0.776	0.587	0.387	7.313	3.003	1.820
20	0.587	0.494	0.358	7.215	2.882	1.762
30	0.492	0.432	0.332	7.180	2.827	1.724
40	0.429	0.386	0.311	7.163	2.794	1.698
50	0.390	0.357	0.295	7.153	2.773	1.679
60	0.356	0.331	0.281	7.146	2.759	1.664
70	0.331	0.310	0.269	7.141	2.748	1.652
80	0.312	0.294	0.257	7.137	2.739	1.643
90	0.294	0.281	0.246	7.134	2.732	1.635
100	0.279	0.266	0.239	7.131	2.727	1.628

ESTIMATING THE CORRELATION COEFFICIENT

In this section we consider the situation in which a researcher is interested in studying the relationship between two variables and for this he must select a sample size. We will denote by ρ the Pearson correlation coefficient and $\zeta = \frac{1}{2} log\left(\frac{1+\rho}{1-\rho}\right)$ a normalized parameterization of the correlation. An important result for this situation, shown by Fisher (1925) says: If R is the sample correlation coefficient obtained from a sample of size n then, the transformation $Z = \frac{1}{2} log\left(\frac{1+R}{1-R}\right)$, asymptotically follows a normal distribution with mean ζ and variance $\frac{1}{n-3}$. After we collect the sample, the values actually obtained are represented by r and $z = Z(r)$. Thus, in order to calculate an adequate sample size and performing inference on the correlation ρ between the two

variables of interest, to simplify the theory developed, we will work with the normalized parameter ζ, instead of ρ, and at the end, we can recover the correlation value using the transformation $\rho = \frac{(e^{2\zeta}-1)}{(e^{2\zeta}+1)}$.

Suppose that in a pilot sample of size n_0 we get the sample correlation r_0 and hence $z_0 = Z(r_0)$ the standardized value. Using the pilot sample to obtain a prior information for ζ we consider as our prior density the likelihood function, which is the normal density with mean z_0 and variance $\frac{1}{n_0-3}$. That is, the prior $\zeta \sim N(z_0, \frac{1}{n_0-3})$. The posterior distribution of ζ, after observing an additional sample of size n, will be Normal with mean $m = \frac{z_0(n_0-3)+z(n-3)}{n_0+n-6}$ and variance $v = \frac{1}{n_0+n-6}$ (DeGroot, 1989).

Thus, a credible interval for ζ, with $(1-\alpha)100\%$ of credibility, has the lower limit $I_1 = m - z_{1-\frac{\alpha}{2}}\sqrt{v}$ and upper $I_2 = m + z_{1-\frac{\alpha}{2}}\sqrt{v}$, and z_α is the α-th quantile of the standart normal distribution. Therefore the length of the credible interval is $I_2 - I_1 = 2(z_{1-\frac{\alpha}{2}}\sqrt{v})$. Whereas in the less favorable case ($\rho = 0$) we want an interval for ρ of maximum length equal to 0.2. Then the length of the interval for ζ equals 0.20068. Thus, we have:

$$I_2 - I_1 = 0,20068 = 2 \times (z_{1-\frac{\alpha}{2}}\sqrt{v}) = \frac{2z_{1-\frac{\alpha}{2}}}{\sqrt{n_0+n-6}}. \tag{4}$$

Whereas in a pilot sample were obtained $n_0 = 20$, $r_0 = 0.78$ and $z_0 = 1.05$ then, the prior density is $\zeta \sim N(1.05, \frac{1}{17})$. Solving the equation (4) for these values and considering a credibility equal to 90% ($\alpha = 10\%$), we obtain the sample size $n \approx 254$. This sample is needed to ensure a range in length from 0.2 to maximum credibility of at least 90%, even for the less favorable case $\rho = 0$. To be convinced of the property of the method, we assume that the correlation obtained with the sample is $r = 0.9$. Here, the interval for ζ is $[1.3450, 1.5454]$. The interval corresponding to ρ would be $[0.8729, 0.9130]$ whose length is 0.0402. As a consequence of the normality of the posterior density, it is interesting to note that the lengths of the intervals for ζ are the same, but the length of the credible interval for ρ will decrease as the value of ρ diverges from zero.

CONCLUSION

This paper shows simple ways of calculating sample sizes in specific cases: proportion estimation, correlation estimation and comparison of proportions. For the estimation of a proportion we consider the two standard cases, superpopulation and finite population. If we know the population size, the sample size should guaranty a minimum fixed precision that is smaller than the precision for the case of superpopulation model. If the interest is to compare two proportions we consider either the difference of proportions, when we obtain the sample size through simulation, or the logarithm of the odds ratio, in which we have the advantage of working with the normal distribution. Considering the correlation we use a transformation introduced by Fisher and then calculate in a simply way the required sample size. The advantage of using Bayesian methods to calculate the sample size is that we do not need to work only with symmetric distributions for parameters in finite intervals, like p_1, p_2 and ρ. The symmetry is only used on the transformation. If

we have any prior information about the parameter of interest we may use it to reduce the sample size. Clearly we have to be secure about the origin of the information.

REFERENCES

1. B. de Finetti (1937), La prévision: ses lois logiques, ses sources subjectives. Ann. Inst. H. Poincaré, 7:1-68
2. C. A. B. Pereira; J. M. Stern (2008), Special characterizations of standard discrete distributions, REVSTAT Statistics Journal 6:199-230.
3. J. Aitchison, S. M. Shen (1980), Logistic-Normal Distributions: Some Properties and Uses, Biometrika, 67:261-272.
4. M. H. DeGroot (1989), Probability and Statistics, Second Edition, Addison-Wesley Publishing Company.
5. R. A. Fisher (1925), Statistical Methods for Research Workers. Edinburgh: Oliver and Boyd.