CS-BIGS 3(1): 18-22
http://www.bentley.edu/csbigs/documents/huson.pdf

The Bayesian Bootstrap in a Predictive Power Analysis

L. W. Huson
Medical Statistics Consultancy Service, GNB Limited, UK

In the planning and design of new clinical trials, calculation of the required sample size and power is a critical part of the process. Power calculations are usually based on quantities estimated from analysis of historical data and are therefore subject to uncertainty. In many cases this is addressed by sensitivity analysis, but simple sensitivity analysis gives an incomplete picture of the uncertainty involved in estimates of power. Here we describe an analysis of historical clinical trial data using the Bayesian Bootstrap, which gives - by generation of the predictive power distribution - a fully probabilistic description of the uncertainty in a power calculation.

1. Introduction

The ICH E9 Guideline on Statistical Principles for Clinical Trials (ICH; 1998) states that "*the number of subjects in a clinical trial should always be large enough to provide a reliable answer to the questions addressed*". By convention, in most pivotal clinical trials, a sample size is chosen which gives at least 80% power, and to derive this, relevant historical data are generally analyzed – for example to give estimates of mean levels of response on test and control treatment arms, and to give estimates of variability.

There are, however, problems associated with conventional power calculations. The most obvious of these problems is that a single estimate of power, based on historical data, is inevitably subject to uncertainty. In many cases there will be relatively little historical data, and hence estimates taken from such data are only approximations at best. Regulatory guidelines recognise this uncertainty and indicate that it should be appropriately addressed. The ICH E9 Guideline, for example, states that: "*it is important to investigate the sensitivity of the sample size estimate to a variety of deviations from these assumptions.* " (ICH; 1998).

In this report we illustrate the use of the Bayesian Bootstrap, in an analysis of historical data, to provide a predictive distribution for the power of a planned future clinical trial. The predictive power distribution gives a complete description of what it is reasonable to believe about the power of the planned trial, and is preferable both to a single power estimate and to a simple sensitivity analysis. We compare the predictive power distribution derived from the Bayesian Bootstrap with one obtained using a more conventional Bayesian analysis implemented in BUGS (Spiegelhalter et. al., 2005).

2. Predictive Power Distributions

Spiegelhalter et. al. (2004, see Section 6.5.5.) give a simple explanation and example of the concept of a predictive power distribution. If the uncertainty about the quantities entering a conventional power calculation (for example mean values and standard deviation) is expressed in the form of prior distributions, then the predictive power can be considered simply as the distribution that is directly induced or implied by these priors: a function of the conditional power and the priors. Spiegelhalter et.al. give an example of how this can be derived by Monte Carlo simulation in a simplified case. This basic idea can be extended to include both prior belief about the uncertain quantities, and also information obtained from historical data, and then it becomes a fully Bayesian procedure. The predictive power distribution provides a complete summary of what a Bayesian would regard as being reasonable to believe about the study power. Rubin & Stern (1998) provide a more detailed theoretical discussion of the use and interpretation of predictive power distributions.

3. The Bootstrap and the Bayesian Bootstrap

Let X_1 , X_2 , . . . X_n be a sample of independent, identically distributed random variables having an arbitrary and unknown distribution. Typically we wish to calculate an estimate of some parameter of this distribution – for example, the mean value. The bootstrap forms an empirical estimate of the sampling distribution of the parameter of interest by drawing n items from the sample data, at random and with replacement, and computing a point estimate from this sample. This process is repeated many times and the resulting collection of estimates constitutes an empirical summary of the sampling distribution. It should be noted that there are some important differences between this technique and the related concept of permutation testing. Efron & Tibshirani (1993; see Section 16.3) specifically discuss this relationship and the differences between the two techniques. In the simple bootstrap process, it can be seen that in each random sample from the $\{X_i\}$, each individual value is drawn with a frequency of between 0 and n times. These frequencies can be regarded as weights in the estimation of, for example, a mean value. The simplest description of the Bayesian Bootstrap is that the integer weights implicit in the conventional bootstrap are replaced by specially-chosen non-integer weights. To form these non-integer weights the following process is described by Rubin (1981):

1. Generate n-1 independent identically distributed random variables from the uniform distribution on the interval 0,1.

2. Order these random variables and denote the ordered sequence as $U_1 \ldots U_{n-1}$. Form the sequence of differences:

$$\begin{aligned} \Delta_1 &= U_1 - 0, \\ \Delta_2 &= U_2 - U_1, \\ \Delta_3 &= U_3 - U_2, \\ &\ldots. \\ \Delta_n &= 1 - U_{n-1}. \end{aligned}$$

Then, proceed as in the conventional bootstrap, but sample values from the data $\{X_i\}$ using the $\{\Delta_i\}$ as weights i.e. data value X_1 is assigned weight Δ_1, which is equivalent to selecting value X_1 with probability Δ_1. Using this weighted sample, calculate the estimate of interest – e.g. the mean value.

3. As with the conventional bootstrap, repeat this process many times to form an empirical distribution of the estimate of interest.

Rubin (1981) proved that this algorithm yields an approximation to the posterior distribution of the parameter of interest, which can be interpreted according to Bayesian principles. The prior that is implied in this process is essentially a non-informative prior for the probabilities of the $\{X_i\}$, under the approximating assumption that the $\{X_i\}$ have a discrete distribution

The theory underlying this prior is discussed in more detail, together with some alternatives, by Shao & Tu (1995), who also provide a full and detailed description of the algorithm and some of the theory underlying it. Lo (1987) reports on a detailed simulation study of the procedure, and Banks (1988) reports a further simulation study of both the Bayesian Bootstrap and some smoothed variants.

Since its introduction, the Bayesian Bootstrap has been little used in practice, despite being computationally more straightforward to implement than many other Bayesian techniques. The main advantage is that it is very easy to compute, using conventional software and programming languages. This means that it can readily be used in practice to form Bayesian posterior distributions without recourse to specialist Bayesian software such as BUGS, and it is therefore very convenient for common analyses such as power calculations.

4. The Data

The historical data we analyse come from a small pilot clinical trial in which a response (a summary of symptom severity scores) was measured on two groups of patients: one group treated with a new test treatment intended to reduce symptom severity, and the second receiving a conventional established treatment. The hope is that the new treatment will result in lower mean symptom severity scores than the established treatment, and a second and larger clinical trial was planned in order to further investigate this hypothesis. The objective of the analysis reported here was to provide Bayesian estimates of the power of this new clinical trial.

For reasons of confidentiality the details of the data and experimental treatment cannot be provided, and the data we utilise in this report have been simulated to have characteristics which closely match those of the original pilot clinical trial. Using this (simulated) historical data, we derive Bayesian estimates of the power of planned second clinical trial, in which the same two treatments will be compared using a larger sample of patients.

In the pilot clinical trial, 37 patients received the conventional treatment, and the mean symptom severity score was 128.1 (standard deviation=52.3). The test treatment arm consisted of 42 patients and the mean score was 112.8 with standard deviation = 41.7. A conventional power calculation (carried out using the package nQuery), using these mean values and assuming a common standard deviation of 47 units, indicates that the planned clinical study will require 150 patients per treatment arm to give 80% power to detect a statistically significant difference between the two treatment arms (using a two-sided alpha level of 0.05). But there is uncertainty about this estimate of power – it is based on the mean values and standard deviations seen in a relatively small sample of historical data, and clearly there is uncertainty about the true values of these means and standard deviations. The Bayesian Bootstrap can be used to summarize the consequences of this uncertainty by forming the predictive power distribution for the new clinical trial.

5. Predictive Power: Bayesian Bootstrap

To implement the Bayesian Bootstrap in a predictive power analysis, we apply the above algorithm to the historical data, to draw estimates of (a) the mean response on the test treatment (b) the mean response on the control treatment and (c) an estimate of the standard deviation, (calculated, for consistency with the original conventional power calculation reported above, by averaging the estimates for the two treatment arms). We then use these quantities in a conventional power calculation, with a fixed size of 150 patients per treatment arm, to derive an estimate of the study power. We repeat this process 5000 times in order to derive the empirical estimate of the predictive power distribution. This process is easily programmed, and we derived this estimate using the SAS® System. The SAS code for this case study is available from the author.

The numbers of replications required in a bootstrap application is always best judged in the context of a particular case by studying the behaviour of the estimated parameter as the number of replications increases. Efron & Tibshirani (1993; see Section 6.4) discuss this and conclude that, sometimes at least, even small numbers of replications can give good estimates. In the context of a Bayesian application, Huson & Kinnersley (2008) reported that their bootstrapped estimates converged satisfactorily after approximately 500 replications.

6. Predictive Power: BUGS

For the purposes of comparison, we also derived a predictive power distribution using the Bayesian modelling language BUGS (Spiegelhalter et. al., 2005). This process involves specifying a Bayesian model for the historical data, including prior distributions for the means and common standard deviation, and then using BUGS to estimate the posterior distributions for the two means and the standard deviation. Again, values from these posterior distributions are used in conventional power calculations to yield an empirical estimate of the predictive power distribution. We implemented this process using BUGS and the BRugs package, which allows BUGS code to be run via the R system. BUGS and R code for this case study is available from the author.

7. Results

The Bayesian process, instead of producing a single point estimate of power for the new clinical trial, yields a predictive (or posterior) distribution which summarizes the uncertainty about the power of planned new trial, and shows plausible values for this power. The predictive power distribution derived using the Bayesian Bootstrap is summarised in Figure 1 – plotted as an empirical cumulative distribution function based on generation of 5000 samples.

The median of this predictive power distribution is 82%, which is close to the point estimate of power provided by the simple nQuery calculation.

The predictive power distribution generated using the BUGS package is shown as Figure 2. The median of this distribution is 81%.

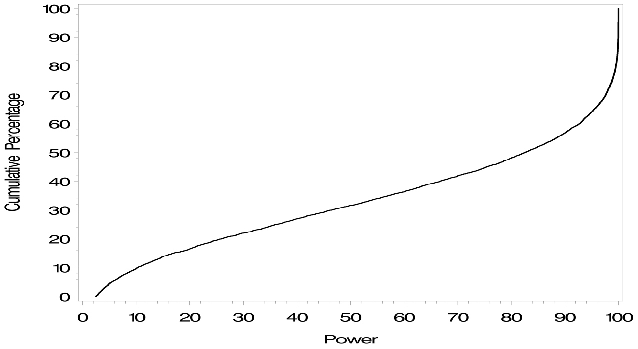

Figure 1. Bayesian Bootstrap Predictive Power [as CDF] [n=150]

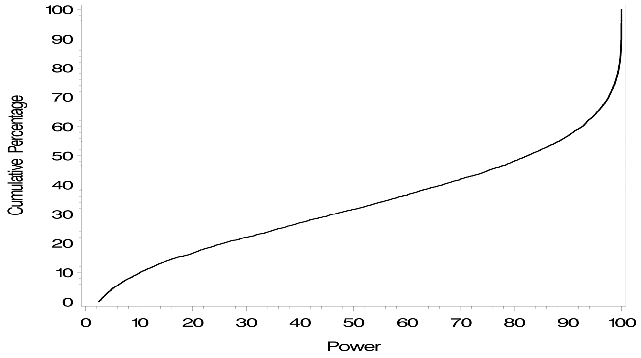

Figure 2. BUGS Predictive Power [as CDF] [n=150]

The similarity between the predictive distributions generated by the Bayesian Bootstrap and by BUGS is illustrated in Figure 3, which overlays the plots of the two simulated predictive distributions as empirical cumulative distribution functions. The two predictive distributions are clearly very similar.

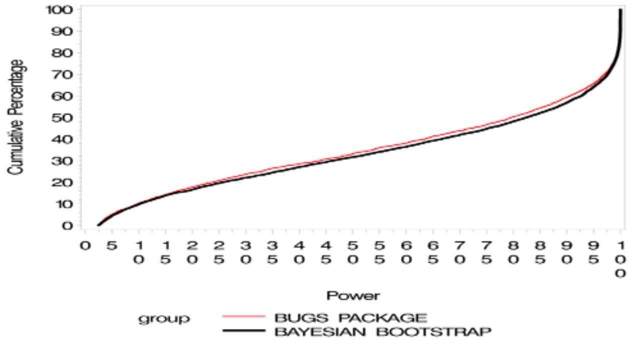

Figure 3. Overlay of Plots of Predictive Power Distributions [n=150]

An alternative presentation is in the form of empirical density functions, and this representation is shown in Figure 4; again the estimates from BUGS and from Bayesian Bootstrapping appear very similar. This representation is the one illustrated by Spiegelhalter et. al. (2004, see Section 6.5.5.), and the forms of the density functions in their example and in the present one are similar and typical of predictive power distributions.

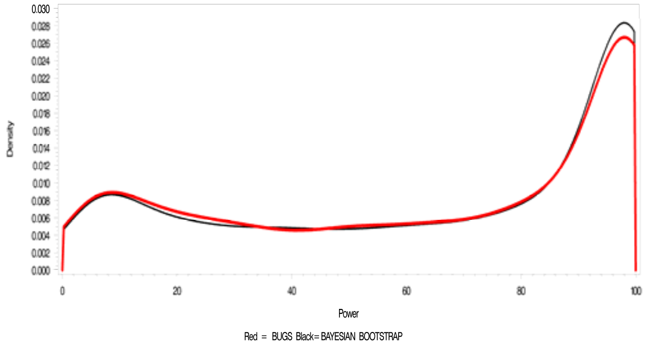

Figure 4. Predictive Power Distribution Plotted as PDFs.

8. Interpretation of the Predictive Distribution

The real value of the predictive power distribution is that it gives a comprehensive illustration of what it is reasonable to believe about the power of the planned future study. A single point estimate of study power may be very misleading, since it does not reflect the uncertainty that exists about the quantities that are used to calculate the power. Even a sensitivity analysis, in which a number of different values of the unknown quantities are used, gives an unsatisfactory summary, as this type of analysis does not indicate the probabilities of the different values - it usually simply illustrates that there is some uncertainty about the power estimate. The Bayesian predictive power distribution, on the other hand, completely summarizes what is known about the power, based on the historical data that are available.

As a single summary of the Bayesian estimate of study power, the median of the predictive power distribution is typically used – this gives a reasonable single estimate from what is often a skewed distribution. Alternatively, the predictive distribution can be used to determine a range of plausible power values and the probability attached to them. For example, in the results shown here, the Bayesian Bootstrap predictive power curve shows that there is 50% probability that the power of the planned study will be 82% or more. Sometimes, in undertaking predictive power analysis, we pre-specify a desirable probability of exceeding a conventional power of, say, 80%, and adjust the sample size until the predictive power curve suggests that we have achieved the desired probability for this power. Rubin & Stern (1998) provide a more detailed discussion of the use of predictive power distributions, and examples of applications in practice are provided by Hutton & Owens (1983) and Schmidli et. al. (2007).

9. Discussion

The use of Bayesian predictive power analysis is sometimes regarded as a curious mixture of Bayesian and Frequentist philosophies. The “uncertain quantity” which is being estimated is the study power, and the power of a study is a firmly frequentist concept. The Bayesian analysis tells us, in effect, what degree-of-belief to attach to any particular range of estimated power values.

Predictive power analysis appears still to be little used in practice, and one of the reasons for this may be that a conventional Bayesian analysis must either use analytical techniques, which in most cases, for reasons of tractability, severely restricts the distributional assumptions which can be made, or must make use of specialist Bayesian software such as the BUGS package.

The main advantage of the Bayesian Bootstrap technique is that it is computationally simple to implement and can easily be programmed in most packages and with most programming languages. This makes it an attractive option for commonly performed calculations such as power analyses.

The results reported here show that, for this example at least, the predictive power distribution from the Bayesian Bootstrap is very similar to that produced from a more conventional Bayesian analysis using BUGS.

Predictive power analysis is not the only Bayesian approach to determination of sample size and power. A convenient summary of some of the alternative Bayesian methods is given by Chow et. al. (2008). More detailed descriptions of other methods – some quite different from the one presented here - are also given by Pezeshk (2003), and Grouin et. al. (2007).

After its introduction by Rubin (1981) few applications of the Bayesian Bootstrap were published in the literature, though in recent years more have started to appear. Kim & Lee (2003) use the Bayesian Bootstrap to analyse proportional hazards models while Aldridge & Bowman (2005) discuss its use in the context of developmental toxicity studies. Douady et. al. (2003) utilise the Bayesian bootstrap in a genetic application, Price et. al. (2005) in the study of protein sequences, and a recent application in the context of estimation of an ROC curve has been provided by Gu et. al. (2008). Clearly the technique is becoming more popular, but, as Shao & Tu (1995) comment, the study of applications of the Bayesian bootstrap is still at an early stage.

REFERENCES

Aldridge G. and Bowman D. 2005. Bayesian bootstrap methods for developmental toxicity studies. *Journal of Statistical Computation and Simulation, 75(2):1 – 91.*

Banks D.L. 1988. Histospline smoothing the Bayesian Bootstrap. *Biometrika,* 75(4): 673-684.

Chow S.C., J. Shao and H. Wang. 2008. *Sample Size Calculations in Clinical Research* (see Chapter 13). Chapman & Hall/CRC.

Douady C.J., F. Delsuc, Y. Boucher, W.F. Doolittle and E.J.P. Douzery. 2003. Comparison of Bayesian and Maximum Likelihood Bootstrap Measures of Phylogenetic Reliability. *Mol. Biol. Evol.,* 20(2):248–254.

Efron B. And R.J. Tibshirani. 1993. *An Introduction To The Bootstrap.* Chapman & Hall Monographs on Statistics and Applied Probability. New York.

Grouin J.M., M. Coste, P. Bunouf and B. Lecoutre. 2007. Bayesian sample size determination in non-sequential clinical trials: Statistical aspects and some regulatory considerations. *Statistics in Medicine,* 26:4914–4924.

Gu J., S. Ghosal S. and A. Roy. 2008. Bayesian bootstrap estimation of ROC curve. *Statistics in Medicine,* published online DOI: 10.1002/sim.3366.

Huson L.W. and N. Kinnersley. 2008. Bayesian Fitting of a Logistic Dose-Response Curve with Numerically-Derived Priors. *Pharmaceutical Statistics* (in press).

Hutton J.L. and R.G. Owens. 1983. Bayesian sample size calculations and prior beliefs about child abuse. *The Statistician,* 42: 399-404.

ICH. 1998. International Conference on Harmonization: Harmonised Tripartite Guideline E9: Statistical Principles For Clinical Trials.

Lo A.Y. 1987. A Large Sample Study of the Bayesian Bootstrap. *Annals of Statistics,* 15(1): 360-375.

Pezeshk H. 2003. Bayesian techniques for sample size determination in clinical trials: a short review. *Statistical Methods in Medical Research,* 12: 489-504.

Price G.A.,G.E. Crooks, R.E Green and S.E. Brenner. 2005. Statistical evaluation of pairwise protein sequence comparison with the Bayesian bootstrap. *Bioinformatics,* 21 (20): 3824–3831.

Rubin D.B. 1981. The Bayesian Bootstrap. *Annals of Statistics,* 9(1): 130-134.

Rubin D.B., H.S. Stern. 1998. Sample size determination using posterior predictive distributions. *Sankhya : The Indian Journal of Statistics: Special Issue on Bayesian Analysis,* Vol spl, Series , Pt. 1: 161-175.

Schmidli H., F. Bretz and A. Racine-Poon. 2007. Bayesian predictive power for interim adaptation in seamless phase II/III trials where the endpoint is survival up to some specified timepoint. *Statistics in Medicine,* 26:4925–4938.

Shoa J. And T. Dongsheng. 1995. *The Jackknife and Bootstrap.* Springer Series in Statistics. Springer-Verlag, New York.

Spiegelhalter D., K.R. Abrams and J.P. Myles. 2004. *Bayesian Approaches to Clinical Trials and Health-Care Evaluation* (see Section 6.5.5). Wiley, Chichester.

Spiegelhalter D., A. Thomas, N. Best and D. Lunn. 2005. *WinBUGS user manual,* version 1.4', MRC BiostatisticsUnit, Institute of Public Health and Department of Epidemiology & Public Health, Imperial College School of Medicine. Downloadable from http://www.mrc-bsu.-cam.ac.uk/bugs.

Correspondence: l.huson@imperial.ac.uk