Bayesian Determination of the Number of Replications in Crop Variety Trials

Siraj Osman Omer[1], Ashutosh Sarker[2] and Murari Singh[3]

Second Bayesian Young Statisticians Meeting (BAYSM 2014)
Vienna, September 18–19, 2014

[1].Experimental Design and Analysis Unit, Agricultural Research Corporation (ARC), P.O. Box 126 , Wad Medani, Sudan, Tell: + 249908246491, E-mail: sirajstat@yahoo.com
[2]. ICARDA, South Asia and China Regional Program CGIAR Block, NASC Complex, India
[3]. International Center for Agricultural Research in the Dry Areas (ICARDA), Amman, Jordan

Abstract

The number of replications required in a particular experiment depends on the magnitude of difference intended for detection and the inherent variability in the response. R and R2WinBUGS codes were used for Bayesian approaches based on prior information on heterogeneity in experimental fields in terms of coefficient of variation (CV) on lentil seed yield. For such a distribution, the Bayesian estimate was obtained as a function of an assumed ratio of variances of prior distributions of the means. The replication increased with the observed level of heterogeneity. For CV less than 16%, the frequentist estimates for number of replications were less than those under the Bayesian while above that level the trend was reversed.

Key words: Bayesian approach, replication, coefficient of variation, R2WinBUGS

1 Introduction

In experimental studies, replication is one of the three cornerstones of statistical inference as per Fisher's 3Rs (Michael, 2001). Bayesian sample size has been discussed by Rahme et al. (2000); Dendukuri et al. (2004); Sahu and and Smith (2006); Joseph and Bélisle (2013) among others. This paper describes Bayesian approach for determining the number of replication, based on behavior of data sets in terms of coefficient of variation from crop variety trials. An analytical approach and simulation using R2WinBUGS software have been used.

2 Bayesian Approach

Bayesian approach for estimation number of replication (NR) is reflected in setting prior information (O'Hagan and Stevens, 2001). The Bayesian inference is based on a prior probability distribution of parameters such as $(\mu \ and \ \sigma^2)$ for the CVs data. The conditional distribution of unknown parameter CV=θ , say given the observed CV = y may be expressed as $f(\theta|y) \propto g(\theta)f(y|\theta)$, called the a posteriori or simply *a posterior* density function of θ , which is obtainable from the famous Bayes' Theorem available in standard texts (Gelman et al., 2003). This a *posteriori* density is used to obtain the expected value of θ as an estimate ofθ , mean, standard error and its credible interval. Based on an experience of fitting distribution to the CVs, let CV follow a shifted log-normal distribution, i.e. log (CV-a) follows a normal distribution. Take $y_i^* = \log(y_i - a)$ and assume $y_i^* \sim N(\tau_i , \sigma^2)$ for a series of i=1,2,…,k values of CV, arising from k trials. Assuming a prior for τ_i , $\tau_i \sim N(\tau , \sigma_0^2)$, the posterior expectation of τ_i given information of y_i can be computed as in the following (Stroup, 1996).

$$E(\tau_i|y_i^*) = \frac{\frac{y_i^*}{\sigma^2}+\frac{y^*}{\sigma_0^2}}{\frac{1}{\sigma^2}+\frac{1}{\sigma_0^2}} \qquad (1)$$

where $y^* = mean(y_i^*) = \frac{1}{n}\sum y_i^* \ = \frac{1}{n}\sum \log(y_i - a)$. Let $\sigma_0^2 = \sigma^2/m$, $\sigma^2 = s^2$, where s^2 is an estimate of σ^2 from fitting the distribution to the observed series of CVs. The above formula can be expressed as

$$E(\tau_i|y_i^*) = \frac{\dfrac{y_i^*}{s^2}+\dfrac{y^*}{s^2/m}}{\dfrac{1}{s^2}+\dfrac{1}{s^2/m}}$$

Or,

$$E(\tau_i|y_i^*) = \frac{y_i^*+my^*}{1+m} \qquad (2)$$

where m is the ratio of variance of the shifted log-normal distribution to variance of its mean parameter, and may be assigned a fixed value. Consider the expression for replication (r)

$$r \geq \left[\frac{\sigma}{\delta^*}\right]^2 \left[z_{1-\alpha/2} + z_{1-\beta}\right]^2 = \theta^2 c \qquad (3)$$

where $\theta = \sigma/\mu$ the coefficient of variation, $\delta^* = (\mu - \mu_0)/\mu$. Here $\mu_0$is a known value yielding the difference $\mu - \mu_0$ for detection using r- replications, μ is the unknown population mean and $c = \left(z_{1-\alpha/2} + z_{1-\beta}\ \right)^2/\delta^{*2}$. Thus, $\theta = \sqrt{\frac{r}{c}}$. Having observed a value of CV, $\theta_i = y_i^*$ we can use equation (2) and (3) to produce $E(\tau_i|y_i^*) = \frac{y_i^*+my^*}{1+m} = \log(y_i^* - a) = \log(\sqrt{\frac{r}{c}} - a)$

or,

$$r = c\left[exp\left(\frac{y_i^*+my^*}{1+m}\right) + a\right]^2 \qquad (4)$$

The expressions from (3) and (4) will be called number of replications (NR) from frequentist and Bayesian analytical approaches, respectively.

3 Dataset and Priors

The data were a set of CVs on seed yield from 226 lentil trials conducted by ICARDA during 1999–2005. Prior information on mean and variance of the shifted log- normal distribution of (CV) was used. For Bayesian- simulation method on CV for trial ’i’, say cv_i, compute a transformed data value $y_{si} = log(cv_i - a)$. The model used was $y_{si}{\sim}N(\mu_i, \sigma_e^2)$. We further assume $\mu_i{\sim}N(\mu, \sigma_\mu^2)$. The priors for $\sigma_e$and σ_μ were assumed as uniform (0, 100), half-normal (0, 0.05) and gamma (0.05, 0.5). The choice of priors for Bayesian analysis was made on lowest value of deviance information criterion (DIC). The DIC values were -85.63, -6519.19 and -347.27 for three priors, respectively. The best prior set based on half normal (0, 0.05) was selected. The posterior mean of r for equation (4) based on the half normal prior and m=1 will be called NR from Bayesian simulation approach. The Bayesian codes can be obtained from the first author.

4 Determination of NR using frequentist and Bayesian analytical methods

The Bayesian analytical value were tabulated for 10 values of m, m = 0, 0.1, 0.2, 0.25, 0.5,1,2, 4, 5, 10.

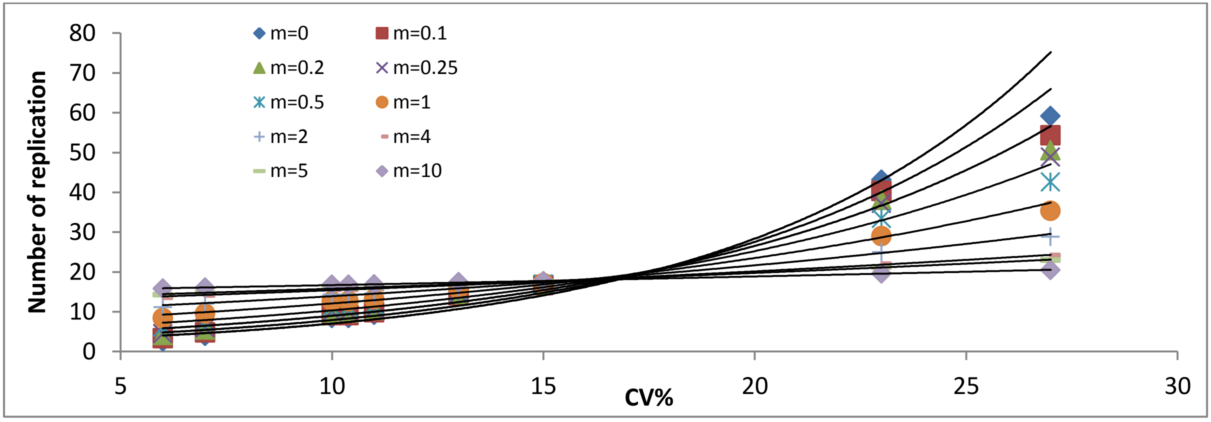

Fig 1. Sample size (r) of number of replication in Bayesian analytical approach for different values of m=σ^2/σ_0^2 for each observed values of CV.

The number of replication (NR) based on CV under Bayesian analytical approach has been presented in Fig 1. When the CV was 5% and 7%, the NR in frequentist approach was 3 and 4 respectively. The NR for Bayesian analytical approach increases with m for CV≤ 15% and decrease for CV ≥ 23%. Fig 1. indicates the decrease in replication with m for CV≤ 17% approximately, and increase for CV> 17%. The correlation between frequentist and Bayesian analytical approach was 0.99% (p<0.001).

Table 1: Number of the Replications (NR) using frequentist and Bayesian simulation approach (m= 1).

CV%	Frequentis method (NR)	Bayesian analytical (NR)	Bayesian simulation, m=1					
			NR	SE	MC error	2.50%	median	97.50%
5.7	2.6	6.1	11.0	10.5	0.17	0.3	8.1	36.3
7.1	3.9	7.4	11.8	10.3	0.16	0.5	9.0	37.1
10.8	9.2	11.8	15.0	10.6	0.14	1.4	12.4	40.4
13.3	13.8	15.1	17.4	10.5	0.16	2.2	15.2	43.0
14.6	16.6	17.0	18.8	11.2	0.16	2.6	16.9	45.5
23.5	43.2	33.4	31.0	15.8	0.26	5.5	30.1	63.8
27.5	59.1	42.6	38.2	20.6	0.35	6.9	36.1	77.0

Where NR=number of replication, SE=standard error, MC error= Monte Carlo error.

In Table 1, the larger NR is due to unusually high CV% observed and the variance estimation in the model. In Bayesian simulation approach, posterior means of replications are higher than their respective median throughout, which indicates that the posterior distribution of NR is skewed on the right (longer tail). The Bayesian analytical values of NR are in between median and mean under Bayesian simulation approach. NR values under frequetist were less than that under Bayesian approaches when the CV level is up to 15% while for higher CVs, frequentist values are higher than the Bayesian values. The correlation coefficient between the number of replications and CV% was (r=+ 0.99) for both Bayesian and frequentist approaches. Generally, an increase in the CV% will increase the number of replications. In crop variety trials, one normally considers number of replications between 2-4, which are found relatively too low when field heterogeneity exceeds 10% in term of CV values in the present evaluation using Bayesian approaches.

Acknowledgements

First author is grateful to ICARDA and Arab Fund for Economic and Social Development (AFESD) for granting a fellowship for carrying out the research study.

References

[1] Dendukuri, A., Elham, R., Belisle, P., Joseph, L. (2004). “Bayesian Sample Size Determination for Prevalence and Diagnostic Test Studies in the Absence of a Gold Standard Test Biometrics” Statistics in Medicine, **60**: 388–397.

[2] O’Hagan, A., and Stevens, J. W. (2001). “Bayesian assessment of sample size for clinical trials of cost-effectiveness.” Medical decision making, **21**:219-230.

[4] Joseph, L. Bélisle, P. (2013). “Bayesian Sample Size Determination for Case-Control Studies When Exposure May be Misclassified. American Journal of Epidemiology Advance Access published September **12**, 2013.

[5] Gelman, A., Carlin, J. B., Stern, H. S., Rubin, D. B. (2003). “Bayesian Data Analysis.”Boca Raton: Chapman & Hall/CRC, 2nd edition.

[6] Michael, F.W. (2001). Guidelines for the Design and Statistical Analysis of Experiments inPapers Submitted to *ATLA*, ATLA. **29**:427.446.

[7] Rahme, E., Joseph, L. and Gyorkos, T. (2000). “Bayesian sample size determination for estimating binomial parameters from data subject to misclassification.”, applied Statistics **49**,119-228.

[8] Sahu, S. K. and Smith,T. M. F. (2006). “A Bayesian method of samplSize determination with practical applications”, J. R. Statist. Soc. A, **169** (2) 235–253.

[9] Stroup, W.W. (1996). “Mixed model procures to assess power, precision and sample size in the design of experiment.” Cary, NC: SAS Institute, Inc.