Available online at www.pelagiaresearchlibrary.com

Pelagia Research Library

European Journal of Experimental Biology, 2014, 4(6):129-133

ISSN: 2248 –9215
CODEN (USA): EJEBAU

Bayesian determination of the number of replications in crop trials

Siraj Osman Omer[1]*, Abdel Wahab Hassan Abdalla[2], Ashutosh Sarker[3] and Murari Singh[4]

[1]*Experimental Design and Analysis Unit, Agricultural Research Corporation (ARC), Wad Medani, Sudan, email:*
[2]*Department of Agronomy, Faculty of Agriculture, University of Khartoum, Sudan*
[3]*International Center for Agricultural Research in the Dry Areas (ICARDA) South Asia and China Regional Program CGIAR Block, NASC Complex, India*
[4]*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 error variability in the response. This paper describes Bayesian approach for determination of the number of replications, based on prior information on heterogeneity in experimental fields in terms of coefficient of variation (CV) on lentil seed yield. The distribution of observed CV was found to be a shifted log-normal distribution taken as prior information. For such a distribution, the Bayesian value was obtained as a function of an assumed ratio of variances of prior distributions of the means. The number of replications, under the Bayesian methods, declined with the ratio. Further, the replication increased with the observed level of heterogeneity. For CV less than 16%, the frequentist value for number of replications were less than those under the Bayesian, while above that level the trend was reversed.

Keywords: Bayesian approach, Replication, Coefficient of variation, R2WinBUGS.

INTRODUCTION

In experimental studies, replication is one of the three cornerstones of statistical inference as per Fisher's 3Rs [9]. A valid estimate of experimental error variance can be obtained only when there are replications and the precision of an experiment can always be increased by additional replications. Several authors, including Goldstein (1981); Pham-Gia and Turkkan (1992); Joseph *et al*. (1997); Sahu and Smith (2006) and M'Lan *et al.* (2008) have proposed Bayesian criteria for sample size determination based on posterior variances [7,13, 10, 17 and 12] have proposed Bayesian criteria for sample size determination based on posterior variances. Gittins and Pezeshk (2000) have proposed a criterion for Bayesian sample size for studies where the goal was to maximize an expected utility function [6]. Bayesian sample size method for prevalence study using a single non-gold standard dialogistic test has been discussed by Adcock (1995); Rahme *et al.* (2000) among others [2 and 14]. Wang and Gelfand (2002) considered Monte Carlo methods for the Bayesian sample size determination (SSD) problem [19]. Antony (2003) emphasized that within practical limits; any desired degree of precision ordinarily may be achieved by replication [3]. Replication are needed for 1) providing an estimate of experimental error, 2) improving the precision of an experiment by reducing standard deviation of the mean (i.e. standard error), 3) increasing the scope of the experiment, and 4) controlling an error variance by grouping similar experimental units together into replicates [8]. In variety trials, the effect size is the standardized mean difference among treatments (genotypes or cultivars) an investigator is interested to discover [15]. Therefore, to determine the number of replications, it is required to give a reliable evaluation of the significance of differences associated with differential treatments. Bayesian approach provides a procedure for making decision under uncertainty by integrating with prior information on parameters such as coefficient of variation. This paper describes Bayesian approach for determination of the number of

replications, based on behavior of data sets in terms of coefficient of variation from crop variety trials. Analytical approach and R2WinBUGS software have been used for the number of replications (sample size) calculation.

MATERIALS AND METHODS

Replication with power consideration

In a normal population, the determining sample size is based on detecting significant difference between the population mean (μ) and a given value (μ_0) intended for it with a given Type I error rate (α) and a power (1- β) where β is Type-II error rate. Type I error rate is probability of rejecting the value μ_0 for the population mean when it is the true value, and is based on the test statistic:

$$Z = \frac{\overline{y} - \mu_0}{\sqrt{\frac{\sigma^2}{r}}} \quad (1)$$

Also, Type II error rate = β is probability of not rejecting (i.e. accepting) and μ_0 as true mean when an alternative value is true. In above statistic, r is the number of replications, and σ^2 is population variance [11]. The distribution of Z is standard normal distribution. The number of replications, associated with Type I error and power 1- β is given by

$$r \geq \left(\frac{\sigma}{\delta}\left\{z_{\left(1-\frac{\alpha}{2}\right)} + z_{(1-\beta)}\right\}\right)^2$$

or,

$$r \geq \left(\frac{CV}{\delta^*}\left\{z_{\left(1-\frac{\alpha}{2}\right)} + z_{(1-\beta)}\right\}\right)^2 \quad (2)$$

where z_α is the α-quantile of standard normal distribution, δ is the difference between the true mean and the given value, where δ = μ - μ_0. Also $\frac{\sigma}{\delta}$ can be expressed as $\frac{\sigma/\mu}{\delta/\mu} = \frac{CV}{\delta^*}$, where $\delta^* = \frac{\delta}{\mu}$. Similar expressions have been presented by [16] in context of design of experiment. We write the percentile from normal distribution (0, 1) as $\int_{-\infty}^{z_\alpha} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx = \alpha$ or $z_\alpha = \Phi^{-1}(\alpha)$, r is sample size, σ is standard deviation and, Φ^{-1} is the standard normal quantile function.

Dataset

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 σ_e and σ_μ were assumed as uniform (0, 100), half-normal (0, 0.05) and gamma (0.05, 0.5).

Bayesian approach to determine number of replication

Bayesian approach for determination of number of replication is reflected in setting prior information [1]. We present the case where records of prior estimates of mean and CVs values are available and the prior datasets allowed examining the distribution of these two parameters. The prior information on parameters might result from a series of already observed CVs datasets. The Bayesian inference is obtained in terms of the probability distribution of parameters such as $(\mu \text{ and } \sigma^2)$ for the CV 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 [5]. This a *posteriori* probability density is used to obtain the expected value of θ as an estimate of θ, mean, standard error and its confidence interval, called Bayesian confidence interval, or, credible interval. To determine the number of replications, CV data is very informative about the quantity being estimated, and then an uninformative prior is an easy choice. Posterior predictive model have assessed by Gelman *et al.* (1996), which reflect their roles in determination of sample size [4]. 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 with a mean and a standard deviation [18]. Take $y_i^* = \log(y_i - a)$ and $y_i^* \sim N(\tau_i, \sigma^2)$ for a series of i=1,2,…,k values of CV, say arising from k trials. Assuming that $\tau_i \sim N(\tau, \sigma_0^2)$, the posterior estimation of τ_i given information of y_i can be computed as in the following.

$$E(\tau_i|y_i^*) = \frac{\frac{y_i^*}{\sigma^2} + \frac{y^*}{\sigma_0^2}}{\frac{1}{\sigma^2} + \frac{1}{\sigma_0^2}} \quad (3)$$

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

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

or,

$$E(\tau_i|y_i^*) = \frac{y_i^* + my^*}{1+m} \quad (4)$$

As in above, m, the ratio of variance of the shifted log-normal distribution to variance of its mean parameter may be assigned a fixed value. Using equation (2), r can be computed as follows

$$r \geq \left[\frac{\sigma}{\delta}\right]^2 [z_{1-\alpha/2} + z_{1-\beta}]^2 = \theta^2 c \quad (5)$$

where $\theta = \sigma/\mu$ the coefficient of variation, $\delta^* = (\mu - \mu_0)/\mu$ 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 equations (4) and (5) to produce

$$E(\tau_i|y_i^*) = \frac{y_i^* + my^*}{1+m} = \log(y_i^* - \text{a}) = \log(\sqrt{\frac{r}{c}} - a)$$

or,

$$r = c\left[\exp\left(\frac{y_i^* + my^*}{1+m}\right) + a\right]^2 \quad (6)$$

RESULTS AND DISCUSSION

Selection of Priors

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 (2) 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.

Table 1: Discrepancy statistics values for selection of the priors for CVs datasets

Priors model	$\bar{D}$	$\hat{D}$	p_D	DIC
P_1	-85.24	-84.85	-0.39	-85.63
P_2	-84.94	6349.32	-6434.25	-6519.19
P_3	-217.67	-88.07	-129.60	-347.27
where:				
Priors model	Parameters			
	σ_e	σ_μ		
P_1	Uniform(0, 100)	uniform (0,100)		
P_2	Half normal (0, 0.05)	Half normal (0, 0.05)		
P_3	Gamma(0.5, 0.5)	Gamma(0.05, 0.5)		

where $\bar{D}$ = posterior mean of (- 2 × log-likelihood). $\hat{D}$ = - 2 × log-likelihood at posterior means of parameters. p_D = effective number of parameters, DIC= Deviance information criterion.

Determination of number of replications (nr) using frequentist and Bayesian analytical methods

From Table 2 give NR using frequentist and Bayesian analytical method. The Bayesian analytical value are tabulated for 10 values of m, m = 0, 0.1, 0.2, 0.25, 0.5,1,2, 4, 5, 10 covering a reasonably wide range for likely values for σ^2 and σ_0^2.

Table 2: Number of replication (NR) using frequentist and Bayesian analytical method

Trial No	CV%	Frequentist method	Bayesian analytical method (m=σ^2/σ_0^2)									
			0	0.1	0.2	0.25	0.5	1	2	4	5	10
5.7	2.6	2.6	3.4	4.1	4.5	6.1	8.4	11.2	13.6	14.3	15.8	5.7
7.1	3.9	3.9	4.8	5.5	5.9	7.4	9.6	12.0	14.2	14.8	16.1	7.1
10.4	8.5	8.5	9.2	9.8	10.1	11.2	12.7	14.3	15.6	16.0	16.8	10.4
10.5	8.6	8.6	9.3	9.9	10.1	11.3	12.8	14.3	15.7	16.0	16.8	10.5
10.8	9.2	9.2	9.9	10.5	10.7	11.8	13.1	14.6	15.8	16.1	16.9	10.8
13.3	13.8	13.8	14.2	14.4	14.6	15.1	15.7	16.4	16.9	17.1	17.4	13.3
14.6	16.6	16.6	16.7	16.8	16.8	17.0	17.2	17.4	17.5	17.6	17.7	14.6
23.5	43.2	43.2	40.4	38.1	37.2	33.4	29.1	25.0	21.9	21.2	19.6	23.5
27.5	59.1	59.1	54.4	50.5	48.9	42.6	35.4	28.8	24.1	23.0	20.5	27.5

CV: coefficient of variation, the parameters σ^2 and σ_0^2 are for $y_i^* \sim N(\tau_i, \sigma^2)$ and where $\tau_i \sim N(\tau, \sigma_0^2)$ $y_i^* = log(cv + 13.6)$.

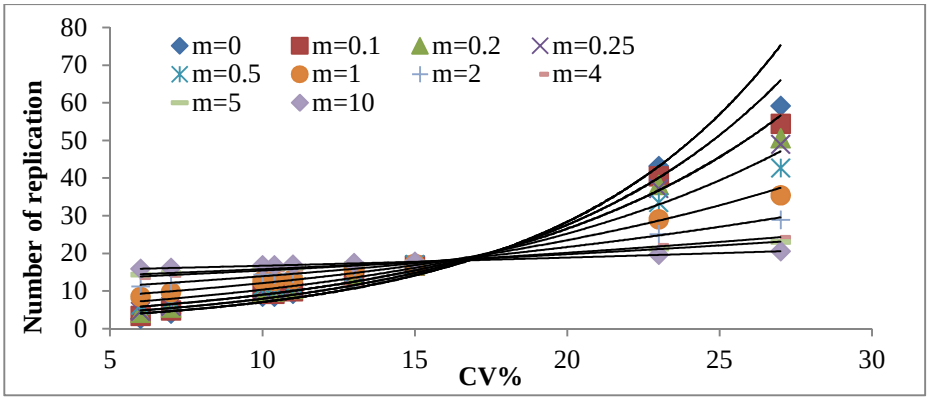

Figure 1: Sample size (r) of number of replication in Bayesian approach for different values of m=σ^2/σ_0^2 for each observed values of CV

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%. Bayesian approach of number of replication estimation may need more investigation, to consider use the CV% as a measure of variability or because of existence of outliers in CV% across heterogeneous environments.

Bayesian determination of number of replications using simulation

Table 3 shows the frequentist estimate and posterior means of Bayesian estimates for number of replication using simulations. The Bayesian estimate of sample size based on CV was considered at m=1. The Bayesian approach based on priors under model P_3 yielding higher number of replication that frequentist approach.

Table 3: Frequentist and Bayesian estimates of Number of the Replication (NR) corresponding m= 1

CV%	Frequentist 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 replications, SE=standard deviation, MC error= Monte Carlo error.

In Table 3, the results of this set of experiments confirmed that the trials with unusually high CV% required larger number of replication due to variance estimation. In Bayesian simulation approach, posterior means of replications are higher than their respective medians throughout, indicating their skewed distribution on the right (longer tail) NR

values under frequentist were less than that under Bayesian approaches when the CV level is up to 15% while for CVs higher than 15%, the frequentist values are higher than the Bayesian values. Generally, number of replications will increase with the increase in CV.

In general, in field crop research, decision on replications is required at the experimental design level. In practice, Bayesian estimation of number of replication for designed experiments in crop variety trials has not received much attention. In an ongoing variety trial, conducted over multiple locations and years, information is available on trial mean and heterogeneity in term of error variance and CV. In this paper, we have discussed simple methods of estimating the number of replication in crop trials integrating the prior information on CV with the current CV of field that can be used for determining the number of replications in future trials. Furthermore, the regression estimates show that CV is positively correlated with number of replications, indicating that the number of replications will increase with CV. The results of this chapter highlighted that Bayesian determination of number of replications very reasonable in comparison with the classical approach. In crop variety trials, with 15 or more genotypes one normally considers number of replications between 2 - 4, which are found too low when field heterogeneity exceeds 10% in term of CV values in the present evaluation using Bayesian approaches.

CONCLUSION

Replication is one way of increasing the precision of a statistical estimate, and this simple fact is the basis of sample size determination. The coefficient of variation is required to determine number of replications. Bayesian approach has advantage of incorporating prior information. Bayesian i.e. posterior means for replications were higher than their Bayesian medians throughout. Consistent differences were found in replication numbers obtained from the three methods. Therefore, the replications may be considered by rationalizing the three values, i.e., computed using frequentist, Bayesian- analytical and Bayesian- simulation methods presented here.

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. Authors acknowledge the comments from reviewers for improving the presentation.

REFERENCES

[1] Adcock CJ, *The statistician*, **1992**, 41, 399–404.
[2] Adcock CJ, *The statistician*, **1995**, 44, 155-161.
[3] Antony J, Butterworth Heinemann, Burlington, MA, First Edition, **1988**, pp, 7-10.
[4] Gelman A, Carlin , Stern H and Rubin D, London: Chapman & Hall/CRC, **2004**, pp 6-14.
[5] Gelman A, Meng, XL, Stern H, *Statistics Sinica* , **1996**, 6, 733–807.
[6] Gittins JC, Pezeshk, H, *Drug information journal*, **2000**, 34,355-363.
[7] Goldstein MA, *The Annals of Statistics*, **1981**,9, 670–672.
[8] Hedges SB, *Mol. Biol.Evol*, **1992**, 9, 366-369.
[9] Johnson DH, Crop Sci, **2006**, 46, 2486–2491.
[10] Joseph L, du Berger R, Belisle A, *Statistics in Medicine*, **1997**, 16(7):769.
[11] Lenth RVS, The American Statistician, **2001**, 55, 187-193.
[12] M'Lan CE, Joseph L, Wolfson DB, *Bayesian Analysis*, **2008**, 3,269–296.
[13] Pham-Gia TG, Turkkan N, *Statistician*, **1992**, 41, 389—397.
[14] Rahme E, Joseph L, Gyorkos T, *Applied Statistics*, **2000**, 49:119-228.
[15] Ralph G, O'Brien JC, Cary, NC: SAS Institute Inc., Chapter, **2007**, 10:237– 271.
[16] Taylor DJ, Muller KE, *The American Statistician*, **1995**, 49, 4347.
[17] Sahu SK, Smith TMF, *J. R. Statist. Soc. A*, **2006**, 169, 235–253.
[18] Stroup WW, experiment. Cary, NC: SAS Institute, Inc, **1996**.
[19] Wang F, Gelfand AE, *Statistical Science*, **2002**, 17,193–208.