APPLICATIONS OF BAYESIAN METHODS IN ANALYSIS OF VARIANCE

BY

KOECH KIBITOK MILTON
REG NO: SC/PGM/014/10

A THESIS SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE AWARD OF THE DEGREEE OF MASTER OF SCIENCE IN BIOSTATISTICS OF THE DEPARTMENT OF MATHEMATICS AND COMPUTER SCIENCE, SCHOOL OF SCIENCE, UNIVERSITY OF ELDORET, KENYA

NOVEMBER, 2013

DECLARATION

Declaration by the candidate

I declare that this thesis is my original work and has not been presented for the award of any degree or otherwise in any other university. No part of this document may be reproduced without prior permission of the author and/or University of Eldoret.

SIGN…………………………………….DATE…………………………………………

Koech Kibitok Milton

SC/PGM/014/10

Declaration by the supervisors

This thesis has been submitted for examination with our approval as University supervisors**.**

Dr. Argwings Otieno..…………………….DATE……………………………………….
Department of Mathematics and Computer Science,
University of Eldoret,
P.O BOX 1125-30100,
ELDORET,
KENYA.

Dr. Victor Kimeli……………………………..DATE………………………………….
Department of Mathematics and Computer Science,
University of Eldoret,
P.O BOX 1125-30100,
ELDORET,
KENYA.

DEDICATION

This research is dedicated to my family and the whole scientific community.

ABSTRACT

Analysis of variance (ANOVA) is a standard method for describing and estimating heterogeneity among the means of a response variable across the levels of multiple categorical factors. In most experimental settings, ANOVA is used to test the presence of treatment effects. Bayesian hypothesis testing literature on ANOVA is scant; the dominant treatment is still classical or frequentist. One impediment to adoption of Bayesian approach is lack of practical development, particularly lack of ready-to-use formulae and algorithms. The aim of this research was to construct a Bayesian hierarchical model for hypothesis test in ANOVA designs using non-informative priors, conditionally conjugate priors as well as the Zellner-g priors. First, the posterior distributions were obtained. Then the effects of various hyper parameters on variance parameters in ANOVA were illustrated. Markov Chain Monte Carlo (MCMC) and Gibbs sampling were then used to obtain posterior point estimates from these posterior distributions. The 95% credible intervals were also obtained and then used to draw inferences. Posterior F-values were obtained for the different priors and finally compared with those obtained using classical approach. Conditional conjugate Normal posterior distribution for means was obtained while conditional conjugate inverse gamma posterior distributions for the variances were also obtained. An F-Value of 4.598 was obtained using the classical approach while posterior F-value of 4.56 was obtained for normal priors for means and conjugate inverse Gamma for the variances. Posterior F-value of 4.62 was obtained using Zellner-g prior (g=n=30) whereas Posterior F-value of 4.52 was obtained using Zellner-g prior ($g=k^2=30$).The results indicated that the F-Values obtained using the classical and the Bayesian approach are similar. The Bayesian test for ANOVA designs is useful to both researchers and students; both groups will get to appreciate the importance of Bayesian approach when applied to practical statistical problems.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

BUGS	Bayesian inference Using Gibbs Sampling
CI	Credible Interval
DIC	Deviance Information Criteria
DF	Degrees of freedom
MCMC	Markov Chain Monte Carlo
HDR	Highest Density Range
ML	Maximum Likelihood
ANOVA	Analysis of Variance

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENT

My heartfelt gratitude goes to the almighty God for his care throughout this study.

My sincere gratitude also goes to my supervisors, Dr. A.R. Otieno and Dr. V. Kimeli for their dedication, guidance, patience, positive criticism and concern during their supervision of this study. With profound humility they have provided me with academic and moral support during the entire course of my graduate program.

I am also greatly indebted to the late Dr A. Koross, Dr. J. Mutiso, Dr. J. Kinyanjui, Dr. K. Nyongesa, Dr. M. Kosgey and Dr. G. Kerich for their assistance and encouragement. I wish to also express my appreciation to all members of Mathematics and Computer Science Department, University of Eldoret, for their inspiration and support.

My appreciation also goes to my parents, brother and sisters for their love and support in the course of my study.

I am also humbled by the great support given to me by my postgraduate colleagues during the write up of this thesis.

Thank you so much and may God bless you

CHAPTER ONE

INTRODUCTION

1.1 Background information

In statistical inference, there are two broad categories of interpretations of probability: Bayesian inference and frequentist (classical or traditional) inference. These views often differ with each other on the fundamental nature of probability. Frequentist inference loosely defines probability as the limit of an event's relative frequency in a large number of trials, and only in the context of experiments that are random and well-defined. Bayesian inference, on the other hand, is able to assign probabilities to any statement, even when a random process is not involved. In Bayesian inference, probability is a way to represent an individual's degree of belief in a statement, or given evidence. Within Bayesian inference, there are also different approaches and interpretations of probability. The most popular interpretations and approaches are objective Bayesian inference (Berger, 2006) and subjective Bayesian inference (Anscombe and Aumann, 1963). Objective Bayesian inference is often associated with (Jeffreys, 1961). Subjective Bayesian inference is often associated with (Brooks, 2003; Gilks, 1994; and De-Finetti, 1937). The first major event to bring about the rebirth of Bayesian inference was by (De-Finetti, 1937).

In many social science settings, the data available for analysis span multiple groups. In these settings it is often plausible that any statistical model that might fit to the data need to be flexible, so as to capture variation across the groups, typically accomplished by letting some or all of the parameters vary across the groups. Examples include survey data gathered over a set of locations (e.g., states, districts, countries); experimental studies deployed in multiple locations; studies of educational outcomes where the subjects are students, who are grouped in classes or schools, which are in school districts, which in turn are in states etc.

In analysis of data of this type, the researcher is interested with the parameters that vary at each group level. These group level parameters go by different names, in different contexts, in different disciplines, and depending on the estimation method being used. Examples include “contextual effects”, “fixed effects”, “random effects”, and “varying” or “stochastic coefficients”. This between-group parameter variation is potentially of great substantive interest, since it speaks to a fundamental issue in empirical social science. Moreover, group by-group analysis is often an important preliminary step in data analysis: a useful and easily-implemented method for assessing parameter heterogeneity, but one that is often overlooked (Berger, 2006). Indeed, one of the most vocal proponents of Bayesian modeling in the social sciences, Andrew Gelman, refers to group-by-group analysis as “secret weapon”: A “weapon” in that group-by-group analysis can be enormously helpful, but “secret” in that in the rush to implement various panel data estimators or Bayesian models and the like, analysts often neglect to take advantage of the insights available from group-by-group analysis. But the general point is that breaking a large data set into group specific pieces will generally result in a better fit to the group-specific data than from a pooled analysis (Gelman, 2007)

Bayesian methods have become increasingly popular in almost all scientific disciplines (Poirier, 2006). One important reason for this gain in popularity is the ease with which Bayesian methods can be applied to relatively complex problems involving, for instance, hierarchical modeling or the comparison between non-nested models. These tests are the cornerstone of data analysis in fields such as biology, engineering, economics, sociology, and psychology. Most researchers report p-values from t-test and F-tests as evidence favoring certain theoretical positions and disfavoring others, but Bayesian approach offers advantages even when the analysis to be run is not complex. For, instances, a traditional frequestist approach to a t-test or one-way analysis of variance ANOVA; two or more group design with one outcome variable would result in a p-value which would be interpreted as the probability of the data (result) assuming the null hypothesis is true. Often the p-values interpretation is abbreviated and it is interpreted as indicating empirical support for or against a null hypothesis. Thus, the first goal of this research was to show how the Bayesian framework of hypothesis testing can be carried out in ANOVA designs. ANOVA is one of the most popular statistical methods used to assess whether or not two or more population means are equal in most experimental settings.

1.2 Statement of the problem

This study entailed constructing Bayesian hierarchical model by applying simulation based techniques; Markov Chain Monte Carlo (MCMC) and Gibbs sampling for carrying out hypothesis tests in ANOVA designs using non-informative priors, conditionally conjugate inverse-Gamma priors and the Zellner-g priors.

1.3 Objectives of the research

I. To obtain the posterior means using non-informative priors.

II. To obtain posterior variance parameters in ANOVA using Zellner g-priors and normally distributed data.

III. To obtain posterior variance parameters using conditionally conjugate inverse-Gamma priors and normal data.

IV. To develop posterior estimates for the unknown parameters i.e the posterior mean, median, "between", and "within", variances,2.5th and 97.5th quartiles.

V. To compare the results under Bayesian to those under the classical approach.

1.4 Significance of the study

Bayesian approach give better results than frequentist approach by accommodating uncertainty in the estimation of parameters in the models, and lead to more appropriate inferences. In classical statistics, computing the uncertainty of functions of random variables such as parameters is not straightforward and involves approximations such as the delta method (Williams *et al.,* 2002). In a Bayesian analysis with Markov Chain Monte Carlo, estimating such and much more complex models and there derived quantities including their uncertainty is trivial once we have a random sample from the posterior distribution of their constituent parts.

Moreover, the greatest impediment to the large-scale adoption of the Bayesian approach to statistical analysis is the lack of ready and easy-to-use formulas, algorithms and tests for statistical models that can be used in practice. Therefore, there is a dire need to conduct this research to help both researchers and students appreciate Bayesian inference when applied to statistical models.

1.5 Outline of the thesis

The organization of this thesis is as follows. Chapter one covers basic concepts of standard methods that are both widely taught and employed, as well as recent shifts in the practice of ANOVA. Chapter two presents an alternative framework of ANOVA along with modifications to the standard ANOVA table summary. Chapter three illustrates Bayesian method and compares it to the classical approaches. In particular, Chapter three it presents an example in which the classical ANOVA yields identical F-values as those of classical approach. Chapter five and six presents the discussion of results obtained in Chapter 4. Recommendations are finally made at the end of this Chapter.

CHAPTER TWO

LITERATURE REVIWIEW

2.1 Introduction

Bayesian statistics became applicable from the beginning of the 21st century. Until late years of the 1980's, Bayesian statistics had been considered as an interesting alternative to the 'classical' theory. The main tools of Bayesian theory were probability theory since all unknown quantities included in the model are considered as random variables. Hence Bayesians were considered as a kind of heretic scientists for several reasons.

Many statisticians accused Bayesian theory to be subjective since a specification of a prior distribution was needed in order to set up inference. But, as history had improved, the main reason for preventing Bayesian theory to expand and establish an accepted quantitative approach for data analysis was the intractabilities involved in the calculation of the posterior distribution except for some simple cases. Asymptotic methods had given solutions to specific problems, but no generalization was possible. Until the early 1990's, two groups of Bayesians; (Gelfand, *et al.,* 1992), (re)discovered Markov Chain Monte Carlo methods (MCMC). Physicists were familiar with MCMC methodology from 1950's. Nick Metropolis and his associates had evolved one of the first electronic super-computers (for those days) and had been testing their theories in Physics using Monte Carlo techniques. The implementation of the MCMC methods in combination with the fast evolution of personal computers had made the new computational tool popular within a few years. Bayesian statistics suddenly became fashionable opening new avenues for statistical research. Using MCMC, complicated models that describe and solve problems that could not be solved with traditional methods can now be set-up and estimated. Since 1990, when MCMC firstly appeared in statistical science, a lot of important related papers had appeared in the bibliography. During 1990-95, MCMC related research focused on the implementation of the new methods in various popular models (for example Gelman & Rubin, 2004; Dellaportas & Smith, 1993; Gelfand *et al*, 1992). The development of MCMC methodology has also promoted the implementation of random effects and hierarchical models.

Bayesian models deal with the possibility of parameter variation across groups by positioning a model for the parameters above the model for the data. The "hierarchy" then arises because the model for the parameters sits "above" the model for the data. Indeed, in this sense all Bayesian models are hierarchical, in that a prior for θ sits above the model for y, the latter indexed by the parameter θ. This notion of a statistical model as a nested hierarchy of stochastic relations permeates all hierarchical modeling, highlighting why hierarchical models are very amenable to Bayesian analysis. Generically, Bayesian hierarchical statistical models have the form:

$y_j|\theta \sim f(y_j|\theta)$ (model for the data in group j = 1, . . . , J)

$\theta|\upsilon \sim f(\theta|\upsilon)$ (between-group model or "prior" for the parameters θ)

$\upsilon \sim P(\upsilon)$ (prior for the hyper parameters υ),

Writing the hierarchy from "bottom" to "top" i.e, the model for the parameters is above that of the data. The inferential challenge is to compute the posterior density of all the parameters, $\theta = (\theta_1, . . . ,\theta_J, \upsilon)'$ and any marginal posterior densities for specific elements of θ that are of interest. Markov chain Monte Carlo and Gibbs sampling are extremely well-suited to this task.

2.2 Advantages of the Bayesian approach to statistics

Key advantages of the Bayesian approach and of the associated computational methods include the following:

2.2.1 Numerical Tractability

Many statistical models are currently too complex to be fitted using classical statistical methods, but they can be fitted using Bayesian computational methods (Link *et al.*, 2002). However, it is reassuring that, in many cases, Bayesian inference gives answers that numerically closely match those obtained by classical methods.

2.2.2 Absence of asymptotics

Asymptotically, that is, for a large sample, classical inference based on maximum likelihood (ML) is unbiased, i.e., in the long run right on target. However, for finite sample sizes, i.e., for a data set, ML may well be biased (Le Cam, 1990). Similarly, standard errors and confidence intervals are valid only for large samples. In contrast, Bayesian inference is exact for any sample size.

2.2.3 Ease of Error Propagation

In classical statistics, computing the uncertainty of functions of random variables such as parameters is not straightforward and involves approximations such as the delta method (Williams *et al.,* 2002). In a Bayesian analysis with Markov Chain Monte Carlo, estimating such and much more complex models and their derived quantities including their uncertainty is trivial once we have a random sample from the posterior distribution of their constituent parts.

2.2.4 Formal Framework for Combining Information

By basing inference on both what we knew before (the prior) and what we see now (the data at hand), and using solely the laws of probability for this combination, Bayesian statistics provides a formal mechanism for introducing external knowledge into an analysis. This may greatly increase the precision of the estimates (McCarthy and Masters, 2007); some parameters may only become estimable through this precise combination of information.

In classical statistics, we always feign total ignorance about the system under study when analyzed.

However, within some limits, it is also possible to specify ignorance in a Bayesian analysis. That is, under the Bayesian paradigm, one can base the inference on the observed data alone and thereby obtain inferences that are typically very similar numerically to those obtained in a classical analysis.

2.2.5 Intuitive Appeal

The interpretation of probability in the Bayesian paradigm is much more intuitive than in the classical statistical framework; in particular, we directly calculate the probability that a parameter has a certain value rather than the probability of obtaining a certain kind of data set, given some Null hypothesis.

Hence, popular statements such as “I am 99% sure that …” are only possible in a Bayesian mode of inference, but they are impossible in principle under the classical mode of inference. This is because, in the Bayesian approach, a probability statement is made about a parameter, whereas in the classical approach, it is about a data set.

Furthermore, by drawing conclusions based on a combination of what we knew before (the prior, or the “experience” part of learning) and what we see now (the likelihood, or the “current observation” part of learning), Bayesian statistics represent a mathematical formalization of the learning process, i.e., of how we all deal with and process information in science as well as in our daily life

2.3 Comparison of Bayesian approach with frequentist approach.

Bayesian inference considers the data to be fixed (which it is), and parameters to be random because they are unknowns. Frequentist inference considers the unknown parameters to be fixed, and the data to be random and estimation is not based on the data at hand only, but the data at hand plus hypothetical repeated sampling in the future with similar data. “The Bayesian approach delivers the answer to the right question in the sense that Bayesian inference provides answers conditional on the observed data and not based on the distribution of estimators or test statistics over imaginary samples not observed" (Link *et al.*, 2010).

Bayesian inference estimates a full probability model. Frequentist inference does not.

There is no frequentist probability distribution associated with parameters or hypotheses. Therefore Bayesian inference estimates are given as P(hypothesis│data). In contrast, frequentist inference estimates are P(data│hypothesis). Even the term 'hypothesis testing' suggests that it should be the hypothesis that is tested, given the data, not the other way around.

Bayesian inference has an axiomatic foundation (Cox, 1946) that is uncontested by frequentists. Therefore, Bayesian inference is coherent to a frequentist, but frequentist inference is incoherent to a Bayesian. Bayesian inference has a decision theoretic foundation (Roberts, 2007; Bernardo and Smith, 2000). The purpose of most of statistical inference is to facilitate decision making (Roberts, 2007). Therefore the optimal decision is the Bayesian decision.

Bayesian inference includes uncertainty in the probability model, yielding more realistic predictions. Frequentist inference does not include uncertainty of the parameter estimates, yielding less realistic predictions. Bayesian inference is consistent with much of philosophy of, where knowledge cannot be built entirely through experimentation, but requires prior knowledge (Roberts, 2007)

Bayesian inference may use Deviance Information Criteria (DIC) to compare models with different methods including hierarchical models, where frequentist model fit statistics cannot compare different methods or hierarchical models.

Bayesian inference safeguards against over fitting by integrating over model parameters. While Bayesian inference is not immune to over fitting, over fitting is largely a frequentist problem. Bayesian inference uses observed data only. Frequentist inference uses both observed data and future data that are unobserved and hypothetical.

Bayesian inference uses prior distributions, so more information is used and 95% probability intervals of posterior distributions should be narrower than 95% confidence intervals of frequentist point-estimates.

Finally, Bayesian inference via MCMC algorithms allows more complicated models that frequentist are unable to estimate.

2.4 Reasons why Bayesian approach has not been widely adopted.

Given all the advantages of the Bayesian approach to statistics mentioned above, it may come as a surprise that currently most statisticians still use classical statistics.

The resistance to the Bayesian philosophy is widely due to its perceived subjectivity of prior choice and the challenge of avoiding to, unknowingly; inject information into an analysis via the priors. However, the lack of a much more widespread adoption of Bayesian methods in statistics mostly has practical reasons. First, a Bayesian treatment shines most in complex models, which may not even be fit in a frequentist mode of inference (Link *et al.*, 2002). Hence, until very recently, most applications of Bayesian statistics featured rather complex statistical models. These are neither the easiest to understand in the first place, nor may they be relevant to the majority of scientists.

Secondly, typical introductory books on Bayesian statistics are written in what is fairly heavy mathematics to most researchers. Hence, getting to the entry point of the Bayesian world of statistics has been very difficult for many researchers. Thirdly, Bayesian philosophy and computational methods are not usually taught at universities. Finally, and perhaps most important, the practical implementation of a Bayesian analysis has typically involved custom-written code in general-purpose computer languages such as FORTRAN or C++. Therefore, for someone lacking a solid knowledge in statistics and computing, Bayesian analyses are essentially out of reach.

CHAPTER THREE

STATISTICAL MODELLING AND BAYESIAN ESTIMATION

3.1 Introduction

This study sought to give a summary to Bayesian statistics and how it is conducted in practice by applying simulation-based methods (MCMC and Gibbs sampling**).** First, Bayesian estimation and model selection were explained in general. Non-informative priors, conditionally conjugate priors and Zellner-g prior were then used to generate posterior distributions. MCMC and Gibbs sampling techniques were then used to obtain posterior point estimates for the parameters from these posterior distributions. Credible bounds analog to Classical confidence intervals were also obtained.

3.2 General Modeling Principles.

Statistical models are used to describe real life problems observed under uncertainty. A statistical model is a collection of probabilistic statements (and equations) that describe and interpret present or predict future performance. It consists of three important components:

1. The response variable (or variables) Y

2. The explanatory variables X_1, $X_{2,}$ . . . , X_J and

3. A linking mechanism between the two set of variables.

The response variables Y are the main study variables and they compose the stochastic part of the model. Concerning these variables, we are frequently interested in describing the mechanism underlying or leading to the appearance of a certain outcome of Y and predict a future outcome of Y. Since the response variable is the stochastic component of the model, we can write:

$Y \sim$Distribution (θ)

Where, θ, is the parameter vector of the distribution used. For example, for a normal regression model, the response-stochastic component of the model is written as

$Y \sim$Normal(μ,σ^2).

Parameter vector θ is expressed as a function of the explanatory variables and a new alternative set of parameters, (α,σ^2) which substitutes the original ones in terms of estimation and inference. Concerning the new set of parameters, the vector θ summarizes the association between the response and the explanatory variables while the rest refer to other characteristics of the distribution such as the variance or the shape. Usually, the mean of the response model is associated with the response variables, but, in more complicated models, the variance or other moment functions can be also estimated via the explanatory variables. The function used to connect the stochastic and the deterministic part of the model (variables Y and X_i's) can be called the 'generalized linking' function. The above terminology and principles were originally introduced in the definition of generalized linear models (McCullagh and Nelder, 1989) but they can be adopted for a wide range of models.

3.3 Definition of statistical models

The most important phenomena in statistical science is the construction of probability models that represent, or sufficiently approximate, the true generating mechanism of a phenomenon under study. The construction of such models is usually based on probabilistic and logical arguments concerning the way that phenomenon works.

Assume a random variable Y, called response, which follows a probabilistic rule with density or probability function f(y|θ), where θ is the parameter vector. Then consider an independent and identically distributed (i.i.d) sample,$\mathbf{y}^T$= [y_1, . . . , y_n] of size n for this variable.

The joint distribution f(y|θ) =$\prod_{i=1}^{n}\left(y_i \middle| \theta\right)$is the likelihood function for the model which contains all the available information provided by the sample. Usually models are constructed in order to assess or interpret causal relationships between the response variable Y and various characteristics expressed as variables X_j, j ∈ V, called covariates or explanatory variables; j indicates a covariate or model term and V the set of all terms under consideration. In such cases, the explanatory variables are linked with the response variables via a deterministic function and part of the original parameter vector is substituted by an alternative set of parameters (denoted by **β**) that usually encapsulate the effect of each covariate on the response variable. For example, in a normal regression model with $\mathbf{y} \sim N(\mathbf{X\beta}, \sigma^2\mathbf{I})$ the parameter vector is given by $\boldsymbol{\theta}^T$= [$\boldsymbol{\beta}^T$,σ^2].

3.4 Bayes theorem

According to Bayesian paradigm, the unobservable parameters in a statistical model are treated as random. When no data is available, a prior distribution is used to quantify our knowledge about the parameter. When the data is available, we can update our prior knowledge using conditional distribution of parameters, given the data. The transition from the prior to posterior is possible via Bayes theorem.

Let us consider two possible outcomes A and B. Moreover assume that $A = A_1 \cup A_2 \cup . . . \cup A_n$ for which $A_i \cap A_j = \emptyset$ for every $i \neq j$. Then Bayes' theorem provides an expression for the conditional probability of A_i given B which is equal to

$$f(A_i|B) = \frac{f(B|A_i)p(A_i)}{f(B)} = \frac{f(B|A_i)f(A_i)}{\prod_{i=1}^{n} f(B|A_i)f(A_i)} \dots\dots\dots\dots (3.1)$$

In a simpler and more general form, for any outcome A and B, we can write

$$f(A|B) = \frac{f(B|A)P(A)}{f(B)} \dots\dots\dots\dots (3.2)$$

The above equation is also called *Bayes' rule.* Bayesian inference is based on this rationale. The above equation, offers a probabilistic mechanism of learning from data (Bernardo and Smith, 1994).The denominator in Bayes' rule, f(B), contains high-dimensional integrals which are analytically intractable. Historically, they had to be solved by more or less adequate numerical approximations. Often, they could not be solved at all. Ironically therefore, for a long time Bayesians thought that they had better solutions in principle than classical statisticians but unfortunately could not practically apply them to any except very simple problems. These were solved with the introduction of MCMC and Gibbs sampling methods.

Hence, after observing data $(y_1, y_2, . . . ,y_n)$ we calculate the posterior distribution $f(\boldsymbol{\theta}|y_1, . . . , y_n)$ which combines prior and data information(likelihood).The posterior distribution is therefore proportional to the product of the prior density and the likelihood function as follows.

$$f(\theta \ y_1, y_2, \dots y_n) = \frac{\prod_{i=1}^{n} f(y_i|\theta)g(\theta)}{\int_{i=1}^{n} \prod_{i=1}^{n} f(y_i|\theta)g(\theta)d(\theta)} \propto \prod_{i=1}^{n} f(y_i|\theta)g(\theta) \dots\dots\dots\dots (3.3)$$

In case there are k unknown parameters then the posterior distribution takes the following form.

$f(\theta_1,\theta_2,\ldots,\theta_k \,|\, y_1, y_2, \ldots . y_n) = \prod_{i=1}^{n} f(y_i \,|\, \theta_1, \theta_2, \ldots . \theta_k) g(\theta_1, \theta_2, \ldots . \theta_k)$…………..……..(3.4)

Further, if the random variables are independent, then for one unknown parameter

$\prod_{i=1}^{n} f(y_i \,|\, \theta) = f(y_1 \;\; \theta)\, f(y_2 \;\; \theta), \ldots\ldots f(y_n \;\; \theta)$……………………………………..……..(3.5)

And for k independent unknown parameters,

$\prod_{i=1}^{n} f(y_i \,|\, \theta_1, \theta_2, \ldots, \theta_k) = f(y_1 \,|\, \theta_1,\theta_2, \ldots, \theta_k)\, f(y_2 \,|\, \theta_1,\theta_2, . . , \theta_k), \ldots, f(y_n \,|\, \theta_1,\theta_2, \ldots . \theta_k)$

…………………………………………………………………………………………(3.6)

The posterior distribution obtained may be highly complex and hence difficult to integrate or use it to compute summary statistics such as the posterior mean, variance or even the posterior probabilities. A dramatic change of this situation came with the advent of simulation based approaches like MCMC and related techniques that draw samples from the posterior distribution (Link and Barker, 2010; Swain *et al.,* 2009; McCarthy and Masters, 2007; Mazzetta *et al.*, 2007 and Gelman, 2006)). These techniques circumvent the need for actually computing the normalizing constant in Bayes rule. Therefore, it is imperative to compute the necessary quantities of interest using Monte Carlo approach. However, simulating from an arbitrary high dimensional distribution is always difficult and often impossible to do directly. Markov Chain Monte Carlo (MCMC) simulates in a Markov Chain from posterior distribution as the stationary or limiting distribution.

This, along with the ever-increasing computer power which is required for these highly iterative techniques, has made the Bayesian revolution in statistics possible (Brooks, 2003). The ease with which difficult computational problems are solved by MCMC algorithms is one of the main reasons for the recent upsurge of Bayesian statistics, rather than the ability to conduct an inference without assuming that one is completely ignorant (i.e., has no prior knowledge about the analyzed system). Therefore this posterior distribution is the key element in Bayesian inference.

3.5 Bayesian estimation

In Bayesian estimation (O'Hagan and Forster, 2004; Lindley, 2000; Bernardo and Smith, 1994), uncertainty about parameters is quantified by probability distributions.

Suppose we have a model M and we wish to estimate the model parameters, θ. Then, we have to define a *prior distribution* over these parameters; f(θ|M). When data Y come in, this prior distribution f(θ|M) is updated to yield the *posterior distribution* f(θ|Y,M). According to Bayes' rule:

$$f(\theta \mid Y, M) = \frac{f(Y \mid \theta, M) f(\theta \mid M)}{f(Y \mid M)}$$

$$= \frac{f(Y \mid \theta, M) f(\theta \mid M)}{\int f(Y \mid \theta, M) f(\theta \mid M) d\theta} \qquad (3.7)$$

$$\propto f(Y \mid \theta, M) f(\theta \mid M) \qquad (3.8)$$

= Likelihood x prior

Hence, the posterior distribution for θ is proportional to the likelihood times the prior. In Bayesian parameter estimation, the researcher is interested in the posterior distribution of the model parameters f(θ|Y,M). However, in Bayesian model selection the focus is on f(Y | M), the marginal likelihood of the data under model M.

3.6 Bayesian model selection

In Bayesian model selection, competing statistical models or hypotheses are assigned prior probabilities. Consider two competing models, M_1 and M_2 with prior probabilities $f(M_1)$ and $f(M_2)$. After observing the data, the relative plausibility of M_1 and M_2 is given by the ratio of posterior model probabilities, that is, the posterior

$$\frac{f(M_1 \mid Y)}{f(M_2 \mid Y)} = \frac{f(M_1 \mid Y) f(Y \mid M_1)}{f(M_2 \mid Y) f(Y \mid M_2)} \text{……………………………………..……………………} (3.9)$$

Hence, the posterior odds are given by the product of the prior odds and the ratio of marginal likelihoods. The latter component is known as the *Bayes factor* (Kass and Raftery, 1995; Berger and Sellke, 1987; Dickey, 1971 and Jeffreys, 1961) and quantifies the change from prior to posterior odds; therefore, the Bayes factor does not depend on the prior model probabilities $f(M_1)$ and $f(M_2)$ and quantifies the evidence that the data provide for M_1 versus M_2.

In linear regression and analysis of variance (ANOVA), two models of special interest are the null model, M_N, that does not include any of the predictors (but does include the intercept) and the full model, M_F, which includes all relevant predictors. In this scenario, the main difficulty with the Bayes factor is its sensitivity to the prior distribution for the model parameters under test (Gelman, 2008; Berger, 2006; Press *et al.*, 2003 and Gelman *et al.*, 1996).

When there is limited knowledge about the phenomenon under study, the prior distribution for the parameters should be relatively uninformative. However, in order to avoid paradoxical results, the prior distribution cannot be *too* uninformative.

In particular, the Jeffreys-Lindley-Bartlett paradox (Robert, 1993; Berger and Delampady, 1987; Lindley, 1980 and Jeffreys, 1961) shows that with vague uninformative priors on the parameters under test, the Bayes factor will strongly support the null model. The reason is that the marginal likelihood f(Y |M) is obtained by averaging the likelihood over the prior; when the prior is very spread out relative to the data, a large part of the prior distribution is associated with very low likelihoods, decreasing the average.

3.7 Prior distributions

3.7.1 Conjugate Prior distributions

A prior distribution that is a member of the distributional family D with parameters $\boldsymbol{\alpha}$ is said to be conjugate to the distribution f(y|θ) if the resulting posterior distribution f(θ|y) is also member of the same distributional family. Therefore

If $\theta \sim D(\alpha)$ then $D(\bar{\alpha})$;

Usually the target posterior distribution is not analytically tractable. In the past, intractability was avoided via the use of ‘conjugate’ prior distributions. These prior distributions have the nice property of resulting to posteriors of the same distributional family as the priors. Extensive illustration of conjugate priors is provided by (Bernardo and Smith, 1994). Where $\boldsymbol{\alpha}$ and $\bar{\alpha}$ are the prior and posterior parameters of D respectively. In many simple cases, the posterior parameters are expressed as weighted means of the prior parameters and maximum likelihood estimators.

3.7.2 Non-informative and weakly-informative prior distributions

A prior distribution is characterized as weakly informative if it is proper but is set up so that the information it does provide is intentionally weaker than whatever actual prior knowledge is available. Non-informative prior distributions are intended to allow

Bayesian inference for parameters about which not much is known beyond the data included in the analysis at hand.

In many occasions we are interested in expressing our prior beliefs in a simpler and more straightforward manner. Usually such prior information is extracted by experts who are not familiar with simply probability notions such as dependence and correlation. Therefore, we need to simplify the prior structure using independent distributions for μ and ω^2 (or equivalently σ^2) and directly specify the prior precision of μ, instead of setting it proportional to σ^2. For example we may consider

$f(\mu,\sigma^2) = f(\mu)f(\sigma^2)$ with $f(\mu) = \text{Normal}(\mu_0,\sigma_0^2)$ and $f(\sigma^2) = \text{Inverse Gamma}(\alpha, \beta)$.

In this case, the resulting posterior distribution is of an unknown form. Consequently, it is difficult to evaluate the posterior summaries and their corresponding marginal densities.

In cases that conjugate priors are considered to be unrealistic or are unavailable, either asymptotic approximations such as Laplace approximation (Kass and Raftery, 1995) or numerical integration techniques can be used. Another appealing alternative is the usage of simulation based techniques. These methods generate samples from the posterior distribution.

3.7.3 Conditionally conjugate prior distributions

A family of prior distributions $f(\theta)$ is conditionally conjugate for θ if the conditional posterior distribution, $f(\theta|y)$ is also in that class. In computational terms, conditional conjugacy means that, if it is possible to draw θ from a class of prior distributions, then it is also possible to perform a Gibbs sampler draw of θ in the posterior distribution. Perhaps more important for understanding the model, conditional conjugacy allows a prior distribution to be interpreted in terms of equivalent data.

In this research we used, the normal prior distributions for the means across the samples, i.e α_j's which are conditionally conjugate given the other parameters, that is, the priors for α_j's give normal posterior distributions, conditional on all other parameters in the model.

The Inverse gamma priors on variance parameters were used in this study. The inverse-gamma family is conditionally conjugate, in the sense that if has an inverse-gamma prior distribution, then the conditional posterior distribution is also inverse-gamma.

3.7.4 Zellner g-prior

This is a conjugate prior which is considered when a Normal-inverse-gamma prior distribution is assigned to the parameters under consideration. These prior takes the form

$\alpha_j \quad \sigma^2 \sim \mathrm{Normal}(\mu, gV\sigma^2)$ and

$\sigma^2 \sim$Inverse Gamma(v_0/2,$\sigma_0{}^2 k_0$/2)

Where g is the parameter controlling the overall magnitude of the prior variance and $V=(X^TX)^{-1}$.

The default choice of g=n is usually adopted since it has an interpretation of adding prior information equivalent to one data point (Kass and Wasswerman, 1995). Another choice of g is to set it equal to the square of the number of predictors of the regression model: $g=k^2{}_,$ where k is the number predictors in the model, i.e, the Risk Inflation Criterion, (Foster and George, 1994). Furthermore, (Fernandez *et al.,* 2001) suggested to take $g=\max\{n,k^2\}$ as a "benchmark prior" This prior has been widely used because it considerably simplifies posterior computations and reduces the number of prior variance parameters that remain to be specified down to one. It also allows for comparison between different values of g (Liang *et al.,* 2008) for discussion and extensions concerning the g-priors.

When no information is available, the above prior set up is simplified by letting the matrix

V= g**I**$_j$ with j=J+1 and g set large to express prior ignorance (for example g=100). This means that the components of the vector α_j will be a priori independent. Hence this prior can be simply written as

$\alpha_j \quad \sigma^2 \sim \text{Normal}(\mu, gV\sigma^2),$ for j=1,…..,J.

Where, μ are components of the prior mean vector, μ_0.

Another alternative is to consider a case where all parameters are a priori independent. It is not conjugate, and hence MCMC methods need to be implemented in order to estimate posterior distribution. However, this prior set up is conditionally conjugate resulting in conditional posterior distributions for α_j's and σ^2 allowing for construction of an efficient Gibbs sampler.

3.8 Modeling data and parameters that vary by groups

Let, y_{ij} be the response variable, i = 1, . . . ,n_j indexes observations within groups j = 1, . . . , J and let n =$\sum_j n_j$be the total number of observations. Within each group the mean of y is α_j ; for simplicity, we assume homoskedasticity across groups such that V($\mathbf{y}_j$) = σ^2 for all j . The means $\bar{y}_1$, . . . ,$\bar{y}_j$ and an estimate of the common variance $\tilde{\sigma}^2$ are sufficient statistics for the data if we assume a normal model for the data. We are interested in the possibility that the means vary across groups. The following hierarchical model operationalizes this possibility.

$y_{ij} = \mu + \alpha_j + \varepsilon_{ij}$

$i=1,2,\ldots,n_j$

$j=1,2,\ldots.,J$

$V(\varepsilon_{ij})= \sigma^2$

$y_{ij}\ |\alpha_j,\sigma^2 \sim \text{Normal}(\alpha_j,\ \sigma^2)$……………………………………………………………..(3.10)

$\alpha_j\ \ \mu_0\ ,\omega^2 \sim \text{Normal}(\mu_0, \omega_0{}^2)$…………………………………………………………..(3.11)

Equation (3.10) is a normal model for the data, with parameters α_j and σ^2, while equation (3.11) is a model for how α_j(means), vary across the groups. The parameter μ_0 is the mean of the distribution of the group means, and this group-level distribution has variance, $\omega_0{}^2$, also known as the between variance;,σ^2 is known as the within variance for groups j. The parameters in the group-level model, μ_0 and $\omega_0{}^2$ are known as hyper parameters. Prior densities for these Parameters, along with a priors for the σ^2 "within variance", are necessary to complete the specification of these models.

3.9 Derivation of posterior distributions

Proposition 1

Assume the model $y_{ij}\sim$ Normal(α_j,σ^2) and $\alpha_j \sim$Normal($\mu_0, \omega_0{}^2$) where i = 1, . . . , n_j indexes observation with group j , j = 1, . . . , J. Then

$\alpha_{j|}y_{j,}\ \sigma^2,\ \mu,{}_{,}\omega^2) \sim \text{Normal}(\bar{\mu}_j,{}_{,}\nu_j)$

Or $\mu_j=\left(\dfrac{\left(\frac{\mu_0}{\omega_0{}^2}-\frac{n\bar{y}}{\sigma^2}\right)}{\left(\frac{1}{\omega_0{}^2}+\frac{n}{\sigma^2}\right)}\right)$ and $v_j = \left(\frac{1}{\omega_0{}^2}+\frac{n}{\sigma^2}\right)^{-1}$

And where $\bar{y}_j = n_j^{-1}\sum_{i=1}^{n} y_{ij}$ is the maximum likelihood estimate of α_j (i.e., the group mean).

3.9.1 Posterior distribution from Normal prior and normal likelihood case

This is when continuous data are available and we are interested in making inference on how the means vary across the groups assuming that the variance σ^2 is known and that the data y_{ij} follow a normal distribution. Observing the data $y_{1,\ldots}y_n$ from the groups j=1,…J, we consider the case where the parameters in the model are $\theta=(\alpha_{1,\ldots}\alpha_J,\mu_o,\omega_o^2,\sigma^2)$.The Bayesian Model is therefore,

$y_{ij}\sim\text{Normal}(\alpha_j,\sigma^2)$, model for the data.

$\alpha_j\sim\text{Normal}(\mu_o,\omega_o^2)$, model for how the means vary across the groups, where $i=1,\ldots n_j$ are observations in the samples j, j=1,…,J

The posterior distribution for the model was estimated where the likelihood of the data is normal with mean α_j and variance σ^2. The conjugate normal prior on μ was used, with mean μ_o variance ω_0^2

$f(\alpha_j \mid y, \sigma^2) \propto f(y \mid \alpha_j, \sigma^2)f(\alpha_j)$

$\propto$ likelihood × prior

Then the parameter of interest is denoted by θ, in this case α_j,

$\alpha_j \mid y_j,\sigma^2,\mu_o,\omega_o^2 \sim \text{Normal}(\hat{u},\tilde{V})$

Posterior distribution $\propto$ likelihood × prior

$f(\alpha_j \mid y, \sigma^2) \propto f(y \mid \alpha_j, \sigma^2)f(\alpha_j)$

$\propto \text{Normal}(\alpha_j,\sigma^2)\times \text{Normal prior}(\mu_0, \omega_0^2)$

Let θ represent the parameter of interest, ie, $\alpha_{j,}$ then

$$f(\alpha_j \mid y) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i-\alpha_j)^2}{2\sigma^2}\right) \times \frac{1}{\sqrt{2\pi\omega_0{}^2}} \exp\left(-\frac{(\alpha_j-\mu_0)^2}{2\omega_0{}^2}\right)$$

$$\propto \exp\left(\sum_{i=1}^{n} \frac{(y_i-\alpha_j)^2}{2\sigma^2} - \frac{(\alpha_j-\mu_0)^2}{2\omega_0{}^2}\right)$$

$$= \exp\left[\frac{1}{2}\left(\sum_{i=1}^{n} \frac{(y_i-\alpha_j)^2}{\sigma^2} + \frac{(\alpha_j-\mu_0)^2}{\omega_0{}^2}\right)\right]$$

$$= \exp\left[-\frac{1}{2\sigma^2\omega_0{}^2}\left(\omega_0{}^2 \sum_{i=1}^{n}(y_i-\alpha_j)^2 + \sigma^2(\alpha_j-\mu_0)^2\right)\right]$$

$$= \exp\left[-\frac{1}{2\sigma^2\omega_0{}^2}(\omega_0{}^2 \sum_{i=1}^{n}(y_i{}^2 - 2\alpha_j y_i + \alpha_j{}^2) + \sigma^2(\alpha_j{}^2 - 2\alpha_j\mu_0 + \mu_0{}^2)\right]$$

The brackets are then opened in order to get the equation in terms of sufficient statistic, $\bar{y}$

$$f(\alpha_j \mid y) \propto \exp\left[-\frac{1}{2\sigma^2\omega_0{}^2}\left(\omega_0{}^2 \sum_{i=1}^{n}(y_i{}^2 - 2\alpha_j \frac{n}{n} y_i + \alpha_j{}^2\right) + \sigma^2(\alpha_j{}^2 - 2\alpha_j\mu_0 + \mu_0{}^2)\right]$$

$$= \exp\left[-\frac{1}{2\sigma^2\omega_0{}^2}\omega_0{}^2 \sum_{i=1}^{n}(y_i{}^2) - \omega_0{}^2 2\alpha_j n\bar{y} + n\omega_0{}^2\alpha_j{}^2 + \sigma^2\alpha_j{}^2 - 2\sigma^2\alpha_j\mu_0 + \sigma^2\mu_0{}^2)\right]$$

The terms are then factored into several parts. Since $\sigma^2\mu_0{}^2$ and $\omega_0{}^2 \sum_{i=1}^{n}(y_i{}^2)$ do not contain, α_j, then they are represented by a constant T, which will drop into the normalizing constant.

$$f(\alpha_j \mid y) \propto \exp\left[-\frac{1}{2\sigma^2\omega_0{}^2}(\alpha_j{}^2(\sigma^2 + n\omega_0{}^2) - 2\alpha_j(\mu_0\sigma^2 + \omega_0{}^2 n\bar{y} + T)\right]$$

$$= \exp\left[-\frac{1}{2}\left(\alpha_j{}^2\left(\frac{\sigma^2+\omega_0{}^2 n}{\sigma^2\omega_0{}^2}\right) - 2\alpha_j\left(\frac{\mu_0\sigma^2+\omega_0{}^2 n\bar{y}}{\sigma^2\omega_0{}^2}\right) + T\right)\right]$$

$$= \exp\left[-\frac{1}{2}\alpha_j{}^2\left(\frac{1}{\omega_0{}^2} + \frac{n}{\sigma^2}\right) - 2\alpha_j\left(\frac{\mu_0}{\omega_0{}^2} + \frac{n\bar{y}}{\sigma^2}\right) + T\right]$$

$\alpha_j{}^2$, is simplified by multiplying the above by $\dfrac{\left(\frac{1}{\omega_0{}^2}+\frac{n}{\sigma^2}\right)}{\left(\frac{1}{\omega_0{}^2}+\frac{n}{\sigma^2}\right)}$.

$$f(\alpha_j \mid y) \propto \exp\left[-\frac{1}{2}\left(\frac{1}{{\omega_0}^2}+\frac{n}{\sigma^2}\right)\left(\alpha_j{}^2\left(\frac{\left(\frac{1}{{\omega_0}^2}+\frac{n}{\sigma^2}\right)}{\left(\frac{1}{{\omega_0}^2}+\frac{n}{\sigma^2}\right)}\right)-2\alpha_j\left(\frac{\frac{\mu_0}{{\omega_0}^2}+\frac{n\bar{y}}{\sigma^2}}{\left(\frac{1}{{\omega_0}^2}+\frac{n}{\sigma^2}\right)}\right)\right)+T\right]$$

$$=\exp\left[-\frac{1}{2}\left(\frac{1}{{\omega_0}^2}+\frac{n}{\sigma^2}\right)\left(\alpha_j{}^2-2\alpha_j\left(\frac{\frac{\mu_0}{{\omega_0}^2}+\frac{n\bar{y}}{\sigma^2}}{\left(\frac{1}{{\omega_0}^2}+\frac{n}{\sigma^2}\right)}\right)\right)+T\right]$$

$$= \exp\left[-\frac{1}{2}\left(\frac{1}{{\omega_0}^2}+\frac{n}{\sigma^2}\right)\left(\alpha_j-\left(\frac{\left(\frac{\mu_0}{{\omega_0}^2}+\frac{n\bar{y}}{\sigma^2}\right)}{\left(\frac{1}{{\omega_0}^2}+\frac{n}{\sigma^2}\right)}\right)\right)^2\right]$$

This is a density function of a normal distribution i.e

$$f(\theta \mid y) \propto \exp\left[-\frac{1}{2}\left(\frac{1}{{\omega_0}^2}+\frac{n}{\sigma^2}\right)\left(\alpha_j-\left(\frac{\left(\frac{\mu_0}{{\omega_0}^2}+\frac{n\bar{y}}{\sigma^2}\right)}{\left(\frac{1}{{\omega_0}^2}+\frac{n}{\sigma^2}\right)}\right)\right)^2\right] \quad \text{,with}$$

Posterior mean: $\mu_1 = \left(\frac{\left(\frac{\mu_0}{{\omega_0}^2}+\frac{n\bar{y}}{\sigma^2}\right)}{\left(\frac{1}{{\omega_0}^2}+\frac{n}{\sigma^2}\right)}\right)$

Posterior variance: $v_j = \left(\frac{1}{{\omega_0}^2}+\frac{n}{\sigma^2}\right)^{-1}$

Posterior precision: $\frac{1}{v_j} = \frac{1}{{\omega_0}^2}+\frac{n}{\sigma^2}$

Posterior precision is therefore the sum prior precision and data precision.

We also look more closely at how the prior mean μ_0 and the posterior mean μ_1 relate to each other;

Posterior mean: $\mu_j = \left(\frac{\left(\frac{\mu_0}{{\omega_0}^2}+\frac{n\bar{y}}{\sigma^2}\right)}{\left(\frac{1}{{\omega_0}^2}+\frac{n}{\sigma^2}\right)}\right)$

$$= \frac{\frac{{\omega_0}^2\sigma^2+n\bar{y}{\omega_0}^2}{{\omega_0}^2\sigma^2}}{\frac{\sigma^2+n{\omega_0}^2}{{\omega_0}^2\sigma^2}}$$

$$= \frac{{\omega_0}^2\sigma^2+n\bar{y}{\omega_0}^2}{\sigma^2+n{\omega_0}^2}$$

$$=\frac{{\omega_0}^2\sigma^2}{\sigma^2+n{\omega_0}^2}+\frac{n\bar{y}{\omega_0}^2}{\sigma^2+n{\omega_0}^2}$$

As n increases, the data mean dominates the prior mean and as ${\omega_0}^2$ decreases(less prior variance, greater prior precision) the prior mean become more important.

3.9.2 Posterior distribution from inverse gamma prior with normal likelihood case

This was to estimate posterior distribution of a model whose likelihood is from a normal distribution with a conjugate inverse Gamma prior with shape parameter α_0 and scale parameter β_0

$f(\sigma^2 \mid y,\mu) \propto f(y \mid \mu, \sigma^2)f(\sigma^2)$

Posterior distribution is therefore, Normal(μ, σ^2) × Inverse Gamma(α_0, β_0), where$\alpha_0=\frac{v_0}{2}$ and $\beta_0 = \frac{v_0{\sigma_0}^2}{2}$

Letting θ to be the parameter of interest, in this case σ^2, then

$$f(\sigma^2 \mid y,\mu) \propto \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\theta}} \exp\left(\frac{(y_i-\mu)^2}{2\theta} \times \frac{{\beta_0}^{\alpha_0}}{(\alpha_0)}\theta^{-(\alpha_0-1)}\exp-\left(\frac{\beta_0}{\theta}\right)\right)$$

$$\propto \prod_{i=1}^{n} \theta^{-1/2}\exp\left(-\frac{(y_i-\mu)^2}{2\theta} \times -\theta^{-(\alpha_0-1)}\exp-\left(\frac{\beta_0}{\theta}\right)\right)$$

$$= \theta^{-n/2}\exp\left(-\frac{\sum_{i=1}^{n}(y_i-\mu)^2}{2\theta} \times \theta^{-(\alpha_0-1)}\exp\left(\frac{\beta_0}{\theta}\right)\right)$$

$$= \theta^{-\left(\frac{n}{2}+\alpha_0+1\right)}\exp\left[-\left(\frac{\beta_0}{\theta}+\frac{\sum_{i=1}^{n}(y_i-\mu)^2}{2\theta}\right)\right]$$

$$=\theta^{-\left(\alpha_0-\frac{n}{2}+1\right)}\exp\left[-\left(\frac{2\beta_0+2\frac{\sum_{i=1}^{n}(y_i-\mu)^2}{2}}{2\theta}\right)\right]$$

$$= \theta^{-\left(\alpha_0-\frac{n}{2}+1\right)}\exp\left[-\left(\frac{\beta_0}{\theta}+\frac{\frac{\sum_{i=1}^{n}(y_i-\mu)^2}{2}}{2\theta}\right)\right]$$

$$= \theta^{-\left(\alpha_0-\frac{n}{2}+1\right)}\exp\left[-\left(\frac{\beta_0+\frac{\sum_{i=1}^{n}(y_i-\mu)^2}{2}}{\theta}\right)\right]$$

This is a density function of an Inverse Gamma distribution with parametersα_1 and β_1

Then the posterior distribution $f(\sigma^2 \mid y,\mu) \propto \theta^{-\left(\alpha_0-\frac{n}{2}+1\right)}\exp\left[-\left(\frac{\beta_0+\frac{\sum_{i=1}^{n}(y_i-\mu)^2}{2}}{\theta}\right)\right]$

$\alpha_1 = \alpha_0 + \frac{n}{2}$

$\beta_1 = \beta_0 + \frac{\sum_{i=1}^{n}(y_i-\mu)^2}{2}$

The posterior is then Inverse Gamma($\alpha_0 + \frac{n}{2}, \beta_0 + \frac{\sum_{i=1}^{n}(y_i-\mu)^2}{2}$) distribution.

3.9.3 Hierarchical One-Way ANOVA

Analysis of variance (ANOVA) is the generalization of a t-test to more than two groups. There are different kinds of ANOVA: one-way, with just a single factor, and two- or multi-way, with two or more factors, and main- and interaction-effects models. Here, we presented a one-way ANOVA and introduce the concept of random effects. In random-effects models, a set of effects (e.g., group means) are constrained to come from some distribution, which is most often a normal. We will first generate and analyze fixed-effects and then random-effects ANOVA.

A full specification of the normal, one-way ANOVA model as a Bayesian hierarchical model is:

$$y_{ij}|\alpha_j, \sigma^2 \sim \text{Normal}(\alpha_j,\sigma^2) \tag{3.12}$$

$$\alpha_j \,|\, \mu_{o,}\omega_o^2 \sim \text{Normal}(\mu_{o,}\omega_o^2) \tag{3.13}$$

$$\mu_0 \sim \text{Normal}(b_0, B_0) \tag{3.14}$$

$$\sigma^2 \sim \text{inverse-Gamma}(v_0/2,\sigma^2\ v_0/2) \tag{3.15}$$

$$\omega^2 \sim \text{inverse-Gamma}(k_0/2,\ k_0\omega_0^2/2) \tag{3.16}$$

A model with unit-wise heteroskedasticity results when we let the "within-unit" variance parameter σ^2 vary over units (i.e., instead of σ^2, we would have the parameters; σ_1^2, $\sigma_2^2,\ldots,\sigma_J^2$)The hyper parameters of the normal prior for μ_0 (the mean b_0 and the variance B_0) and the hyper parameters of the priors for the model parameters are in the vector,θ = $(\alpha_1, . . \alpha_J,\mu_0,\sigma^2,\omega_0^2)$.

The hierarchical structure of the model implies that the prior density for θ can be factored as follows:

$f(\theta) = f(\alpha_1, \ldots, \alpha_J, \mu_o, \sigma^2, \omega_0{}^2)$

$= f((\alpha_1, \ldots, \alpha_J \mid \mu_0, \omega_0{}^2)\ f(\mu_0) f(\ \sigma^2) f(\omega_0{}^2)$

$= \prod_{i=1}^{n} f(\alpha_j \mid \mu_0, \omega_0{}^2)\ f(\mu_0) f(\ \sigma^2) f(\omega_0{}^2)$

3.10 Markov Chain Monte Carlo (MCMC) and Gibbs Sampling

MCMC is a set of techniques to simulate draws from the posterior distribution f(θ | Y) given a model, a likelihood f(Y | θ), and data Y, using dependent sequences of random variables. That is, MCMC yields a sample from the posterior distribution of a parameter. MCMC was developed in 1953 by the physicists Metropolis, and later generalized by Hastings (1970), and so one of the main MCMC algorithms is called the Metropolis Hastings algorithm. Many different algorithms of MCMC are available now. One of the most widely used MCMC techniques is Gibbs sampling

(Geman, 1984). It is based on the idea that to solve a large problem, instead of trying to do all at once, it is more efficient to break the problem down into smaller sub units and solve each one in turn. Here is a sketch of how Gibbs sampling works:

Let the data be Y andθ be the vector of unknowns parameters to be investigated, ie $\theta = (\alpha_1, \ldots \alpha_J, \mu_0, \omega_0{}^2, \sigma^2)$, hence the Gibbs sampler algorithm works as follows for the unknown parameters.

1. Choose starting (initial) values $\alpha_1{}^{(0)}, \ldots, \alpha_J{}^{(0)}, \mu_0{}^{(0)}\ \omega_0{}^{2(0)}, \sigma^{2(0)}$

2. Simulate $\alpha_1{}^{(1)}$ from the distribution Normal$\left(\left(\frac{\left(\frac{\mu_0}{\omega_0{}^2} + \frac{n\bar{y}}{\sigma^2}\right)}{\left(\frac{1}{\omega_0{}^2} + \frac{n}{\sigma^2}\right)}\right), \left(\frac{1}{\omega_0{}^2} + \frac{n}{\sigma^2}\right)^{-1}\right)$

   Simulate$\alpha_2{}^{(1)}$,from $f(\alpha_2 \mid \alpha_1{}^{(1)}, \ldots \alpha_J{}^{(0)}, \mu_0{}^{(0)}, \omega_0{}^{2(0)}, \sigma^{2(0)}, Y)$

   Simulate $\alpha_3{}^{(1)}$, from $f(\alpha_3 \mid \alpha_1{}^{(1)}, \alpha_2{}^{(1)}, \alpha_4{}^{(0)}, \ldots \alpha_J{}^{(0)}, \mu_0{}^{(0)}, \omega_0{}^{2(0)}, \sigma^{2(0)}, Y)$

.

.

.Simulate $\omega^{2(1)}$ from distribution Inverse Gamma $\left(\frac{k_0+J}{2}, \frac{k_0\omega_0^{2}}{2}\right)$

-Simulate $\sigma^{2(1)}$ from distribution Inverse Gamma $\left(\frac{v_0+n}{2}, \frac{v_0\sigma^{2}}{2}\right)$

- Iterate this procedure

3. Repeat step 2 many times (e.g. 100s, 1000s,100000s, etc) to eventually obtain a sample from f(θ│Y), i.e target density or the limiting distribution

Step 2 is called an update or iteration of the Gibbs Sampler and after convergence is reached, it leads to one draw (Posterior sample) consisting of k values from the joint posterior distribution f(θ│Y). The conditional distributions in this step are called “full conditionals” as they condition on all other parameters. The sequence of random draws for each of k parameter resulting from step 3 forms a Markov Chain. Therefore, a simple summary of a Bayesian statistical analysis is as follows:

1. We use a degree-of-belief definition of probability rather than a definition of probability based on the frequency of events among hypothetical replicates.

2. We use probability distributions to summarize our beliefs or our knowledge (or lack thereof) about each model parameter and apply Bayes rule to update that knowledge with observed data to obtain the posterior distribution of every unknown parameter in the model. The posterior distribution quantifies all our knowledge about these unknowns given the data, the model, and prior assumptions. All statistical inferences are based on the posterior distributions.

3. However, posterior distributions are virtually impossible to compute analytically in all but the simplest cases; hence, we use simulation methods (MCMC) to draw a series of dependent samples from the posterior distribution and base our inference on that sample. WinBUGS (the MS Windows operating system version of BUGS: Bayesian Analysis Using Gibbs Sampling) is a versatile package that has been designed to carry out Markov chain Monte Carlo (MCMC) computations for a wide variety of Bayesian models) applies a MCMC algorithm for the model specified and includes the data set and conducts the iterative simulations for the target distributions, (Link *et al.*, 2006).

3.10.1Output from MCMC Algorithm

Once the iterations of MCMC are completed, a series of random numbers from the joint posterior distribution $f(\theta \mid Y)$are obtained. Essentially, it is important to make sure that these numbers come from a stationary distribution, i.e., that the Markov Chain that produced them was at an equilibrium. If that is the case, then this becomes the estimate of the posterior distribution. Also, these numbers should not be influenced by the choice of initial parameter values supplied to start the Markov Chains (the initial values); and these successive values are correlated. This is called convergence monitoring. Once convergence is attained, the posterior samples are summarized to estimate any desired feature of the posterior distribution, for instance, the mean, median, or mode as a measure of central tendency. This is a Bayesian point estimate or the standard deviation of the posterior distribution and is a Bayesian measure of the uncertainty of a parameter estimate.

3.10.2 Convergence Monitoring

This term refers to whether the algorithm has reached its equilibrium (target) distribution. If this is true, then the generated sample comes from the correct target distribution. Hence monitoring the convergence of the algorithm is essential for producing results from the posterior distribution of interest.

There are many ways to monitor convergence. The simplest way is to monitor the MC (Markov Chain) error since small values of it will indicate that the quantity of interest has been calculated with precision. Monitoring autocorrelations is also very useful since low or high values indicate fast or slow convergence respectively.

A second way is to monitor the trace plots, i.e. the plots of the iterations versus the generated values. If all values are within a zone without strong periodicities and (especially) tendencies then assumes convergence. After the burn-in period the generated sampled values are stabilized within a zone. Most methods use at least two parallel chains, but another possibility is to compare successive sections of a single long chain. The simplest method is just to inspect plots of the chains visually: they should look like nice oscillograms around a horizontal line without any trend. Visual checks are routinely used to confirm convergence.

The first step in making an inference from an MCMC analysis is to ensure that an equilibrium distribution has indeed been reached by the Markov Chain, i.e., that the chain has converged. For each parameter, the chain started at an arbitrary point (the initial value or init chosen for each parameter), and because successive draws are dependent on the previous values of each parameter, the actual values chosen for the initial values is noticeable for a while. Therefore, only after a while the chain is independent of the values with which it was started. These first draws are discarded as a burn-in as they are unrepresentative of the equilibrium distribution of the Markov Chain.

Another, more formal check for convergence is based on the Gelman Rubin (or Brooks Gelman Rubin) statistic (Gelman et *al*., 2004), called Rhat when using WinBUGS from R via R2WinBugs. Values near 1 indicate likely convergence, and 1.1 is considered by some as an acceptable threshold (Gelman and Hill, 2007; Gelman *et al.,* 2004 and Bernardo, 2003). With this approach, it is important to start the parallel chains at different selected or at random places.

3.10.3 Summarizing the Posterior for Inference

The aim of a Bayesian analysis is not the estimate of a single point, as the maximum of the likelihood function in classical statistics, but the estimate of an entire distribution. That means that every unknown (e.g. parameter, function of parameters, prediction, and residual) has an entire distribution. The posterior can be summarized graphically, e.g., using a histogram or a kernel-smoother.

Alternatively, the mean, median, or mode can be used as a measure of central tendency of a parameter (i.e., as a point estimate) and the standard deviation of the posterior as a measure of the uncertainty in the estimate, i.e., as the standard error of a parameter estimate. Finally, the Bayesian analog to a 95% confidence interval is called a Bayesian credible interval (CRI) and is any region of the posterior containing 95% of the area under the curve. There is more than one such region, and one particular CRI is the highest-posterior density interval (HPDI). However, in this research, we only considered 95% CRI’S, bounded by the 2.5th and the 97.5th percentile points of the posterior sample of a parameter.

3.10.4 Forming Predictions

Predictions are expected values of the response for future samples or of hypothetical values of the explanatory variables of a model, or more generally, of any unobserved quantity. Predictions are very important for:

(a) Presentation of results from an analysis and

(b) To understand what a model entails. For example, the biological meaning of an interaction or a polynomial term can be difficult to determine from a set of parameter estimates. Because predictions are functions of parameters and of data (values of covariate), their posterior distributions can again be used for inference with the mean and the 95% CRIs often used as the predicted values along with a 95% prediction interval.

3.11 Algorithm for Gibbs Sampler

The posterior density for this problem $f(\theta\ |Y) \propto f(\theta)f(Y|\theta)$ is high dimensional: θ contains Jα_j's parameters, plus μ_0, and the two variances σ^2 and $\omega_0{}^2$ making a total of J + 3 parameters. This is where the Gibbs Sampler becomes especially very useful. Characterizing the J+3 dimensional posterior density $f(\theta\ |\mathbf{Y})$ was done by successively sampling from J+3 conditional densities.

Figure 3.1, below provides a graphical representation of the hierarchical model provided from which it will be straight forward to deduce the forms of the conditional distributions required to implement a Gibbs Sampler. The conditional independence relations among the random quantities in the model are shown in the diagram: e.g., given that α_j, the data y_{ij} are conditionally independent of the hyper parameters μ_0 and the between-unit variance ω^2. Moreover, given other components (data or parameters) of the model, α_j and

μ_o are conditionally independent of one another, means that these group-specific parameters can be updated one at a time, in a series of iterative updating steps (Robert, 2007).

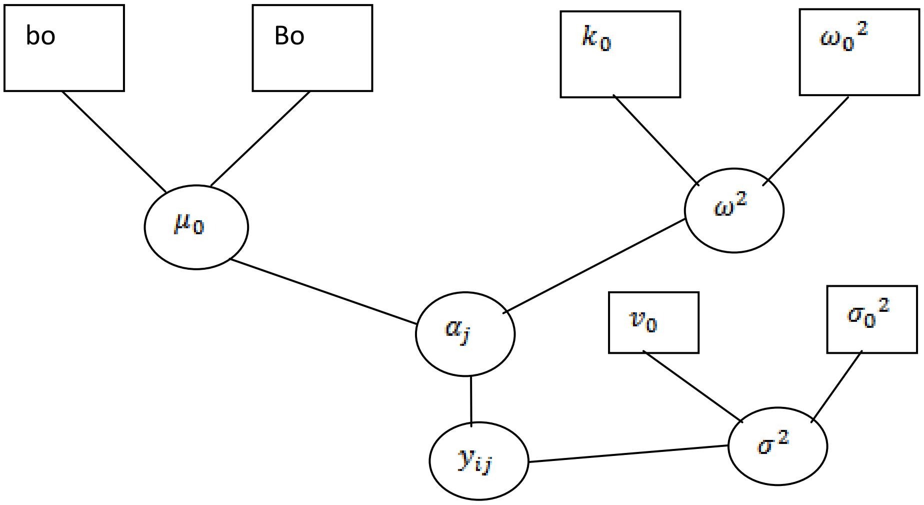

Figure 3.1: Graphical displays of data, priors and hyper parameters for the hierarchical model.

This then leads to the specific conditional distributions needed to implement the Gibbs sampler.

1. $f(\alpha_j|y, \mu_{0,}\omega_0{}^2, \sigma^2)$, j = 1, . . . , J .Each α_j has priors with parameters μ_0 and $\omega_0{}^2$ and the data in groups, j=1,2,…J, and this data $\mathbf{y}_{ij}$ has the priors with parameters α_j and σ^2.Therefore

   Since $f(\alpha_j \,|\, \mu_0, \omega^2) \equiv N(\mu_0,\omega_0{}^2)$ and $f(y_{ij} \,|\, \alpha_j,\sigma^2) \equiv N(\alpha_j,\sigma^2.)$. This conditional distribution then leads to

   $$\alpha_j|y, \mu_{0,}\omega_0{}^2, \sigma^2) \sim \text{Normal}\left(\left(\frac{\left(\frac{\mu_0}{\omega_0{}^2}+\frac{n\bar{y}}{\sigma^2}\right)}{\left(\frac{1}{\omega_0{}^2}+\frac{n}{\sigma^2}\right)}\right), \left(\frac{1}{\omega_0{}^2}+\frac{n}{\sigma^2}\right)^{-1}\right)$$ from proposition.

2. f(μ_0|bo,B_0), The prior for μ_0 are the hyper parameters, the prior mean b_0 and prior variance B_0 respectively. This again becomes Normal distribution (from proposition) and therefore leads to

$$\mu_0|y,bo,B_0\omega_0^{\;2}\sim\text{Normal}\left(\left(\frac{\frac{b_0}{B_0^{\;2}}+\frac{\mu_j J}{B_0}}{\frac{1}{B_0}+\frac{J}{\omega^2}}\right),\left(B_0^{\;-1}+\frac{J}{\omega_0^{\;2}}\right)^{-1}\right)$$

3. f(ω^2| k_0,$\omega_0^{\;2}$). The priors forω^2 are just its prior hyper-parameters, k_o and $\omega_0^{\;2}$. These priors are conditioned to the α_j; and this has its priorsω^2 and μ_0. Hence,

The prior density f(ω^2 | ko, $\omega_0^{\;2}$) is an inverse-Gamma density, while the α_j have normal densities, and so the results of Proposition 1 apply. That is, the inverse-Gamma prior over ω^2 is conjugate with respect to the normal “likelihood” over the α_j and so

f(ω^2 | ko, $\omega_0^{\;2}$)~Inverse Gamma$\left(\frac{k_0+J}{2},\frac{k_0\omega_0^{\;2}+S_u}{2}\right)$

Where $S_u=\sum_{i=1}^{J}\left(\alpha_j-\mu_0\right)^2$

4. f(σ^2|v_o, $\omega_0^{\;2}$.).The priors for σ^2 are just its prior hyper parameters, v_o and $\omega_0^{\;2}$. These priors are combined with the likelihood i.e the data y_{ij}; to give a posterior distribution.

   The prior density f(σ^2|v_o,σ^2) is an inverse-Gamma density, while the y_{ij} have normal densities, and again, the results of Proposition 1 apply. That is

   σ^2|v_o, $\omega_0^{\;2}$~Inverse-Gamma ($\frac{v_0+n}{2},\frac{v_0\sigma^2+S_Y}{2}$)

   Where n=$\sum_j \mathbf{n}_j$is the total number of observations in group j and

   $\mathbf{S}_Y=\sum_{i=1}^{J}\sum_{i=1}^{n}(\mathbf{y}_{ij}-\boldsymbol{\alpha}_j)^2$is the total sum-of-squares of **Y**.

An iteration of the Gibbs Sampler consists of sampling from each of these conditional distributions, with the sampled values stored and available as conditioning arguments in subsequent steps. Formally, the Gibbs Sampler makes the transition from:

$\theta^{(t)}=(\alpha_1^{(t)},\ldots..\alpha_J^{(t)},$ų$_0^{(t)},\omega^{2(t)},\sigma^{2(t)})$ to $\theta^{(t+1)}$ as follows

1. Sample, $\alpha_j\,^{(t+1)}$ from the normal density given in equation 3.13, with the conditioning arguments ų$_0$, ω^2,σ^2 set to ų$_0^{(t)},\omega,^{2(t)},\sigma^{2(t)}$ respectively

2. Sample $\mu_o^{(t+1)0}$ from the normal density given in equation 3.14, with the conditioning arguments ų and ω^2 set to$\alpha_j^{2(t)}$ and $\omega^{2(t)}$ respectively (i.e $\alpha_{j's}$ the were “updated” in step 1).

3. Sample $\sigma^{2(t+1)}$ from the inverse-Gamma density in equation 3.15, with the conditioning argument S_Y updated to $S_Y^{(t+1)}$. That is, S**y** is a function of the α_j, which were updated to $\alpha_j^{(t+1)}$ in step 1

4. Sample $\omega^{2(t+1)}$ from the inverse-Gamma density in equation 3.16, with the conditioning argument S_u set to $S_u^{(t+1)}$. That is, $S\mu$ is a function of both α_j and μ_0; the α_j were updated to were updated to $\alpha_j^{(t+1)}$ in step 1 and μ_0 was updated to $\mu^{(t+1)}$ in step 2

After these four steps, a complete, $\theta^{(t+1)}$ has been sampled, and becomes conditioning arguments in the next iteration. The one-way ANOVA can be parameterized in various ways. We adopted a means parameterization of the linear model for the fixed-effects, one-way ANOVA:

$y_i = \alpha_j + \varepsilon_i$

$\varepsilon_i \sim$ Normal $(0,\sigma^2)$

Here, y_i is the observation i in group j, α_j is the mean group j, and residual ε_i is the random deviation of, i from its population mean α_j. It is assumed to be normally distributed around zero with constant variance σ^2.

Without any further assumption, the population means α_j's are simply some unknown constants that are estimated in a fixed-effect's ANOVA. If, however, a distributional assumption about the population means α_j's is added, we obtain a random-effects ANOVA:

$y_i = \alpha_j + \varepsilon_i$

$\varepsilon i \sim \text{Normal}\ (0,\sigma^2)$

$\alpha_j \sim \text{Normal}\ (\mu_0,\omega^2)$

The interpretation of α_j and ε_i as population mean and residual, respectively, is unchanged. But now, the α_j's parameters are no longer assumed to be independent; rather, they come from a second normal distribution with mean μ_0 and variance ω^2. The latter are also called hyper parameters, because they are one level higher than the parameters α_j's that they govern.

Thus, typical fixed-effects factors would be sex or cereal variety in an agricultural experiment etc. Typical random-effects factors might be time (e.g., year, month, or day) or location, such as experimental blocks or other spatial units on which repeated measurements are taken. This similarity is because of the common stochastic process that generated them and thus creates a stochastic relationship among the effects of the levels of a random-effects factor.

In contrast, when factor levels are modeled as fixed they are considered unrelated or independent. Therefore, the reasons for moving from fixed-effects ANOVA to the corresponding random-effects ANOVA include:

1. Extrapolation of inference to a wider population,
2. Improved accounting for system uncertainty, and
3. Efficiency of estimation.

First, viewing the studied effects as a random sample from some population enables one to extrapolate to that population. This generalization can only be achieved by modeling the process that generates the realized values of the random effects (i.e., by assuming a normal distribution for the α_j (above). Second, declaring factor effects as random acknowledges that when repeating the study, we obtain a different set of effects, so the resulting parameter estimates will differ from those under study. Random-effects modeling properly accounts for this added uncertainty in our inference about the analyzed system. Third, when making random-effects assumption about a factor, these effects are no longer estimated independently; instead, estimates are influenced by each other and therefore are dependent. Specifically, individual estimates are "pulled in" toward the common mean μ, i.e., they are closer to μ than the corresponding fixed-effects estimates. This is why random effects estimators are said to be "shrinkage estimators". Estimates that are more imprecise and are based on a smaller sample size are shrunk more. When effects are indeed exchangeable, shrinkage results in better estimates (e.g with smaller prediction error) than the estimates obtained from a fixed-effects analysis (Gelman, 2007) WinBUGS software has been developed to carry out MCMC computations on a broad range of statistical models within the Bayesian framework that treats all quantities as random variables. The model it assumes consists of a joint distribution over all unobserved quantities such as parameters (or nodes in WinBUGS terms), and observed quantities such as collected data. It then conditions on the data to obtain a posterior distribution over the parameters through Bayes theorem. To obtain inferences on the unknown quantities of interest from the model, it marginalizes the posterior distribution by using the MCMC simulation techniques.

CHAPTER FOUR

DATA ANALYSIS AND RESULTS

4.1 Introduction

This chapter gives a presentation of results in the form of tabulations and their accompanying graphs. A short discussion of these results is also made.

One-way ANOVA design is used to illustrate the differences between the various priors and the effects they have on variance parameters.

This followed the data from an experiment that was set up to investigate to what extent the yield of dyestuff differs between batches of the raw material. The experiment featured six batches with five observations each as shown in the table below.

Table 4.1: Data from a balanced experiment with five samples each with six randomly chosen bathes of raw material

Batch	Yield (in grams)				
1	1545	1440	1440	1520	1580
2	1540	1555	1490	1560	1495
3	1595	1550	1605	1510	1560
4	1445	1440	1595	1465	1545
5	1595	1630	1515	1635	1625
6	1520	1455	1450	1480	1445

Source: (Box and Tiao, 1973)

The data in table 4.1, arose from a balanced experiment in which the total product yield was determined for 5 samples from each of 6 randomly chosen batches of raw material. In order to illustrate the behavior of the various parameters when the null hypothesis is true, the difference between the batch mean and the overall mean was subtracted from the batch data. The objective was to determine the relative importance of between batch variation versus variation due to sampling and analytic errors. We assume that the batches and samples vary independently, and contribute additively to the total error variance.

First, a classical one-way ANOVA is carried out to compute the F statistic and the corresponding p value for the data set. We used the following model for the yield

$$y_{ij} = \mu + \alpha_j + \varepsilon_{ij}$$

Where; μ is a fixed effect

α_j's are fixed contrasts from μ (so they sum to zero)

$\varepsilon_{ij} \sim N(0, \sigma^2)$, independent and identically distributed, (iid), are random draws for y_{ij} deviations from$\mu + \alpha_j$

There are seven parameters to be estimated: 5α_j's (6 minus 1 constraint), μ and σ^2.

Where y_{ij} is the yield for group j, α_j is the mean yield of batch j, σ_{within} is the inverse of the within-sample variance σ^2_{within} ,(i.e. the variation due to sampling and analytic error), μ is the true average yield for all the batches and ώ-within is the inverse of the between-sample variance $ώ^2_{between}$.

4.2 Classical (Frequentist or traditional) approach

In general, one way (factor) ANOVA techniques can be used to study the effect of k (>2) levels of a single factor.

To determine if different levels of the factor affect measured observations differently, the following hypothesis was tested.

Ho: $\mu_1=\mu_2 = \mu_3 = \mu_4, \dots . = \mu_{J.}$

Versus

H_1: $\mu_2 \neq \mu_3 \neq \mu_4, \dots . \neq \mu_J$.

To compare the means of J different populations, we have j groups of sizes, n_j .This can also be written as:

$y_{ij}= \mu + \tau_j+ \varepsilon_{ij}$, where j = 1,…, J (groups or samples) and i = 1, 2, ..., n_j

That is, an observation is the sum of three components:

1. The grand mean μ of the combined populations.
2. A treatment effect τ_j associated with the particular population from which the observation is taken; put in another way, τ_j is the deviation of the group mean from the overall mean.
3. A random error term ε_{ij}. This reflects variability within each population

An alternative way to write the model is:

$$y_{ij}= \mu_j+ \varepsilon_{ij},$$

Where μ_j= mean of the j^{th} population = $\mu + \tau_j$.

Assumptions

When applying one way analysis of variance there are key assumptions that should be satisfied. They are essentially the same as those assumed for k = 2, levels, and they include:

1. The population at each factor level is (approximately) normally distributed

2. The observations are obtained independently from the populations defined by the factor levels.

3. These normal populations have a common variance, σ^2.

4. The random error terms are independent, i.e the ε's are independent and identically distributed, (iid)

5. $\varepsilon_{ij} \sim N(0,\sigma^2)$

Thus for factor level i, the population is assumed to have a distribution which is $N(\mu_i,\sigma^2)$.

The following computational value for F and notations cater for both equal and unequal sample sizes.

Then, if H_0 is true,

$$F = \frac{\text{MS between}}{\text{MS within}} \sim F(J\text{-}1,\ N\text{-}J)$$

That is, if H_0 is true, then the test statistic F has an **F** distribution with J - 1 and N - J degrees of Freedom.

Table 4.2: Coefficient values obtained using the classical approach

	Coefficients	Estimate Std. Error	t value	Pr(>|t|)
(Intercept)	1527.50	9.04	168.985	< 2e-16 ***
Batch 1	-22.50	20.21	-1.113	0.27666
Batch 2	0.50	20.21	0.025	0.98047
Batch 3	36.50	20.21	1.806	0.08351
Batch 4	-29.50	20.21	-1.459	0.15739
Batch 5	72.50	20.21	3.587	0.00149
s-within	42.00			
s-between	49.5			

Residual standard error: 49.51 on 24 degrees of freedom. Multiple R-Squared: 0.4893, Adjusted R-squared: 0.3829 .F-statistic: 4.598 on 5 and 24 DF, p-value: 0.004398

Most of the coefficients are non-significant, suggesting that the batch means do not differ significantly from the grand mean. The coefficients for batch 6 is –sum(the rest) = -57.5.

This model can also be given using the “mixed effects” linear model:

$y_{ij} = \mu + \alpha_j + \varepsilon_{ij}$

Where:μ is a fixed intercept.

$\alpha_j \sim \mathrm{Normal}(0, \sigma^2)$, iid, is a random draw for batch mean’s deviation from μ

$\varepsilon_{ij} \sim \mathrm{Normal}(0, \omega^2)$, iid, is a random draw for y_{ij} deviations from $\mu + \alpha_j$

This structure of this model remains the same as the one given above, but the number of parameters here are only three; μ, σ^2 andω^2.

The,μ,is a fixed is a ”fixed effect”, $\alpha_j{}'s$ together are one “random effect”, and $\sigma^2,$ andω^2 are called “variance components”. Therefore, this model is equivalent to

$y_{ij} \sim \mathrm{Normal}(\mu, \sigma^2 + \omega^2)$

However, in this model, y_{ij} are no longer independent of each other: y_{ij} in the same batch j depend on the same random draw α_j and so are dependent.

Linear mixed-effects model fit by REML

Data: dyes

AIC	BIC	LOGLIK
325.6543	329.7562	-159.8271

Random effects

	Intercept	Residual
Std Dev	42.0061	49.5101

Fixed effects

	Value	Std.error	DF	t-value	p-value
(Intercept)	1527.5	19.38342	24	78.80448	0.000

From these, it is evident that

s-within, i.eσ = 42.00

s-between, i.e ω = 49.5

Formulae and notations used in classical approach**.**

Formulae and notations:

Number of samples (or levels)	=	J
Number of observations in ithsample	=	n_j
Total number of observations	=	N =$\sum_i n_i$
Observation j in ith sample	=	y_{ij}, j= 1,2,….. n_i
Sum of n_i observations in the ith sample	=	H_i = $\sum_i y_{ij}$

Sum of all N observations = $H = \sum_i H_i = \sum_i \sum_j y_{ij}$

Total sum of squares = $SS_T = \sum_i \sum_j y_{ij}^2 - \frac{H^2}{N}$

Between sum of squares, = $SS_B = \sum_i \frac{H_i}{n_i} - \frac{H^2}{N}$

Within sum of squares, SS_W $= SS_T - SS_B$

Hence,

Total mean square, = $MS_T = \frac{SS_T}{N-1}$

Between samples mean square = $MS_B = \frac{SS_B}{J-1}$

Within samples mean square, = $MS_W = \frac{SS_W}{N-J}$

Where MS_T, MS_B, and MS_W are mean total sum of squares, mean between sum of squares and mean within sum of squares respectively whereas SS_T, SS_B and SS_W are defined in the formulae and notations above.

The degrees of freedom, (DF) is then given by: (J-1) + (N - J) = (N-1)

Sum of squares for between variation, (SS_B) captures the variability between the groups. If all groups had the same mean, SS_B would equal 0. The term $SS_{Explained}$ is also used because it reflects variability that is “explained” by group membership. Since there are J groups, and one grand mean, hence DF for Between variation = J - 1.

Sum of squares for within variation (SS_W) captures variability within each group. If all group members had the same score, SS_W would equal to 0. It is also called SS Errors or SS Residual, because it reflects variability that cannot be explained by group membership. There are n_j degrees of freedom associated with each individual group, so the total number of degrees of freedom within = $\Sigma(n_j - 1) = N - J$

From the data given above, the classical (frequentist) analysis is done as follows:

Total sum of squares $= SS_T = \sum_i \sum_j y_{ij}^{\ 2} - \frac{H^2}{N} = 70{,}112{,}875 - \frac{45{,}825^2}{30}$

$= 115{,}187.5$

$= \sum_i \frac{H_i^{\ 2}}{n_i} - \frac{H^2}{N} = 56{,}357.5$

$= (7525^2/5+7640^2/5+7820^2/5+7490^2/5+8000^2/5+7350^2/5) - \frac{45{,}825^2}{30} = 56{,}357.5$

Within samples sum of squares, $SS_W = SS_T - SS_B = \sum_i \sum_j (y_{ij} - \alpha_j)^2$

$= 58{,}830$

Therefore,

Total sum of squares = between samples sum of squares + within samples sum of square

Table 4.3: ANOVA- table using classical method

Source	DF	Sum of squares	Mean square	F
Between	5	56,357.5	11,271.5	F = $\frac{MS_B}{MS_W}$ = 4.598
Within	24	58,830	2,451.25	
Total	29	115,187.5	4,215.98	

At a level of α= 0.05, the classical approach gave an F value (calculated) of 4.598 which was then compared with table values.

F(J-1,N-J) = F(5,24) = 2.7763 < Fcalculated = 4.598.

We reject the null hypothesis i.e the means are not equal.

4.3 Bayesian approach in data analysis

Graphical display for the model

Figure 4.1, demonstrates how the Bayesian hierarchical model for ANOVA can be analyzed using WINBUGS. Theta refers to the posterior grand mean, mu[i]'s are the posterior means across the groups while sigma2.btw and sigma2.with are between and within variances respectively.

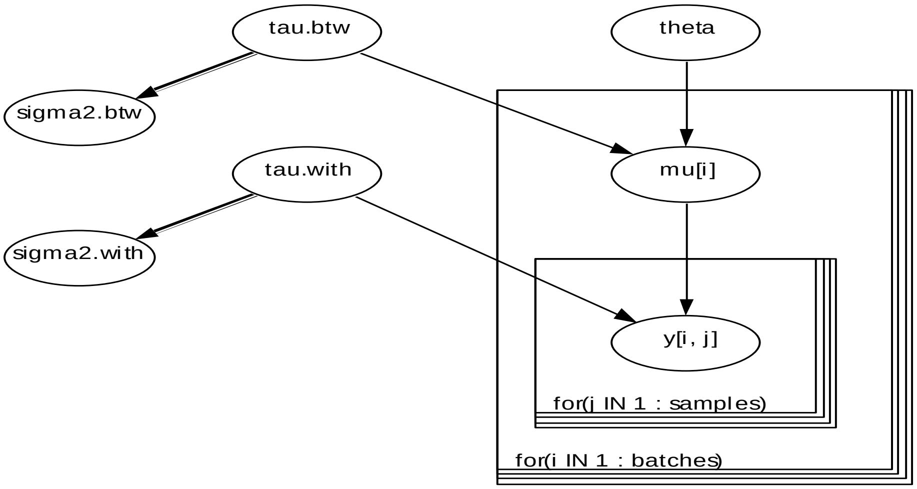

Figure 4.1: WinBugs Graphical display for the Bayesian model.

The posterior summaries after 100,002 iterations and additional discarded 5,000 burn-in iterations using Normal prior for the mean and Inverse-Gamma prior for the variance parameters will be produced. It also gives the posterior means for the batches, posterior between and within variances. Finally it also gives 95% credible set analog to confidence interval in frequentist approach

Table 4.4: Posterior point estimates from Normal posterior for means and inverse gamma posterior for variance.

	mean	sd	MC_error	val2.5pc	median	val97.5pc	start	sample
mu[1]	1514.0	20.47	0.1963	1471.0	1515.0	1552.0	5000	100002
mu[2]	1528.0	19.31	0.1142	1489.0	1528.0	1566.0	5000	100002
mu[3]	1550.0	22.19	0.2991	1510.0	1550.0	1595.0	5000	100002
mu[4]	1509.0	21.28	0.241	1466.0	1510.0	1549.0	5000	100002
mu[5]	1572.0	29.16	0.5545	1516.0	1575.0	1625.0	5000	100002
mu[6]	1492.0	25.92	0.4349	1443.0	1491.0	1541.0	5000	100002
s-with	49.74	9.24	0.1301	39.35	52.47	75.03	5000	100002
s-btw	41.65	27.15	0.4727	0.3101	37.34	102.4	5000	100002
sigma2.with	2474.54	4151.0	33.04	0.09619	1394.0	10490.0	5000	100002
sigma2.btw	1734.72	1069.0	15.35	1548.0	1753.0	5630.0	5000	100002
theta	1528.0	21.98	0.116	1483.0	1528.0	1572.0	5000	100002
F	4.56	8.355	0.06948	1.122E-4	2.589	21.19	5000	100002

The results in Table 4.4, above gives posterior numerical summaries from the model after 100,002 iterations and additional discarded 5,000 burn-in iterations using Normal prior for the mean and Inverse-Gamma prior for the variance parameters. MCMC algorithm gives the posterior means for the batches, posterior between and within variances. It also gives 95% credible set analog to confidence interval in frequentist approach. This gives a grand posterior mean of 1528.0, posterior within variance of 2474.54 and posterior between variance of 1734.72. These results closely agree with those obtained using frequentist approach. Posterior F-value was 4.56 which is similar to that obtained using Classical approach.

Table 4.5: Posterior summaries using Zellner's g-prior (g = n = 30)

	mean	sd	MC_error	val2.5pc	median	val97.5pc	start	sample
mu[1]	1514.0	20.25	0.2181	1472.0	1515.0	1552.0	5000	100002
mu[2]	1528.0	18.91	0.1426	1490.0	1528.0	1566.0	5000	100002
mu[3]	1549.0	22.08	0.3141	1509.0	1548.0	1593.0	5000	100002
mu[4]	1510.0	21.08	0.2589	1467.0	1511.0	1548.0	5000	100002
mu[5]	1570.0	29.62	0.5769	1515.0	1573.0	1624.0	5000	100002
mu[6]	1494.0	26.2	0.4547	1444.0	1493.0	1542.0	5000	100002
s-with	51.02	9.424	0.139	39.38	50.7	75.71	5000	100002
s-btw	42.89	26.06	0.4665	0.5191	35.07	95.89	5000	100002
sigma2.with	2603.40	3516.0	31.67	0.2695	1230.0	9194.0	5000	100002
sigma2.btw	1839.55	1097.0	16.24	1551.0	2777.0	5732.0	5000	100002
theta	1528.0	20.89	0.1496	1486.0	1527.0	1569.0	5000	100002
F	4.62	8.335	0.06848	1.122E-4	2.542	21.16	5000	100002

The results in Table 4.5, above gives posterior summaries for the data after 100,002 iterations and additional discarded 5,000 burn-in iterations using Zellner's g-prior with g = n = 30, where n is the total number of observations in the data set. When g=n=30, the posterior grand mean was 1,528.0, posterior within variance was 2,603 and posterior between variance was 1,839. Posterior F-value was 4.62 which is similar to that obtained using classical approach.

Table 4.6: Posterior summaries using Zellner's g-prior (g = k^2= 25).

	mean	sd	MC_error	val2.5pc	median	val97.5pc	start	sample
mu[1]	1514.0	20.38	0.2367	1471.0	1515.0	1551.0	5000	100002
mu[2]	1528.0	19.17	0.1186	1489.0	1528.0	1566.0	5000	100002
mu[3]	1550.0	22.15	0.4009	1511.0	1549.0	1595.0	5000	100002
mu[4]	1510.0	21.13	0.3012	1466.0	1511.0	1548.0	5000	100002
mu[5]	1572.0	29.34	0.7553	1517.0	1575.0	1625.0	5000	100002
mu[6]	1492.0	25.99	0.5722	1443.0	1491.0	1540.0	5000	100002
s-with	48.71	9.227	0.1691	39.39	52.57	74.94	5000	100002
s-btw	39.86	27.12	0.6395	0.3025	37.12	102.0	5000	100002
sigma2.with	2372.0	3894.0	44.48	0.09148	1378.0	10410.0	5000	100002
sigma2.btw	1541.34	1068.0	19.77	1552.0	2764.0	5615.0	5000	100002
theta	1534.0	20.89	0.1496	1486.0	1527.0	1569.0	5000	100002
F	4.52	8.255	0.06948	1.122E-4	2.592	21.26	5000	100002

The results in Table 4.6, above gives WinBUGS posterior summaries for the data after 100,002 iterations and additional discarded 5,000 burn-in iterations using Zellner's g-prior with g = k^2= 25, where k is the number of observations from each group. The effects of using different values of, g, is also demonstrated in Table 6 and Table 7 above. From Table 7, posterior grand mean is 1,528.0, whereas posterior within variance is 2,372.0 and posterior between variance is 1,584.34. Posterior F-value was 4.52 which is similar to that obtained using Classical approach.

4.3.1 Box Plots

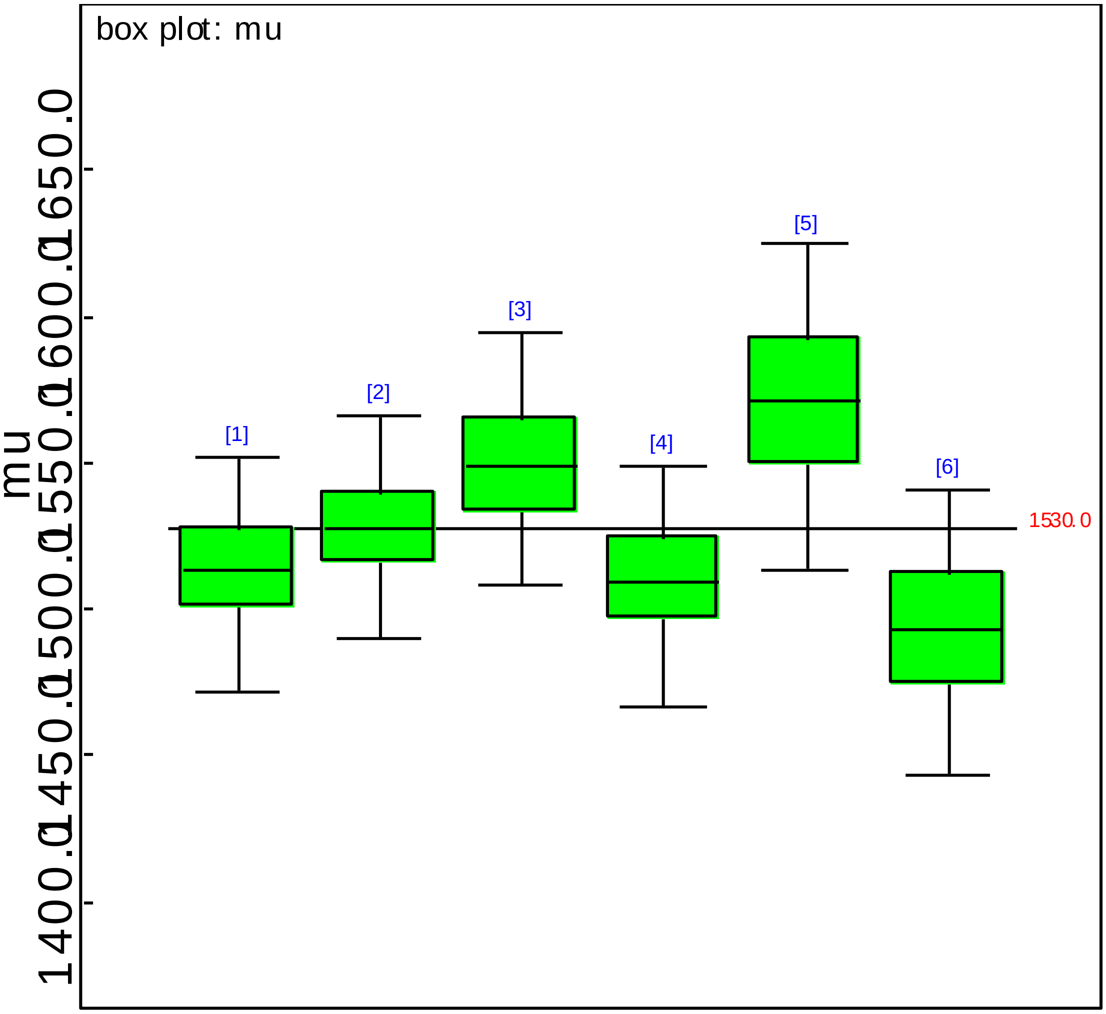

Figure 4.2: Box plots for within sample variance, posterior means and between sample variance respectively

Figure 4.2, above shows a graphical display of the posterior means across the groups. This is a posterior distribution of means which conditionally conjugate given the other parameters. These means are ranging from 1492 to 1572, with the posterior grand mean being 1527.5.

4.3.2 Posterior densities

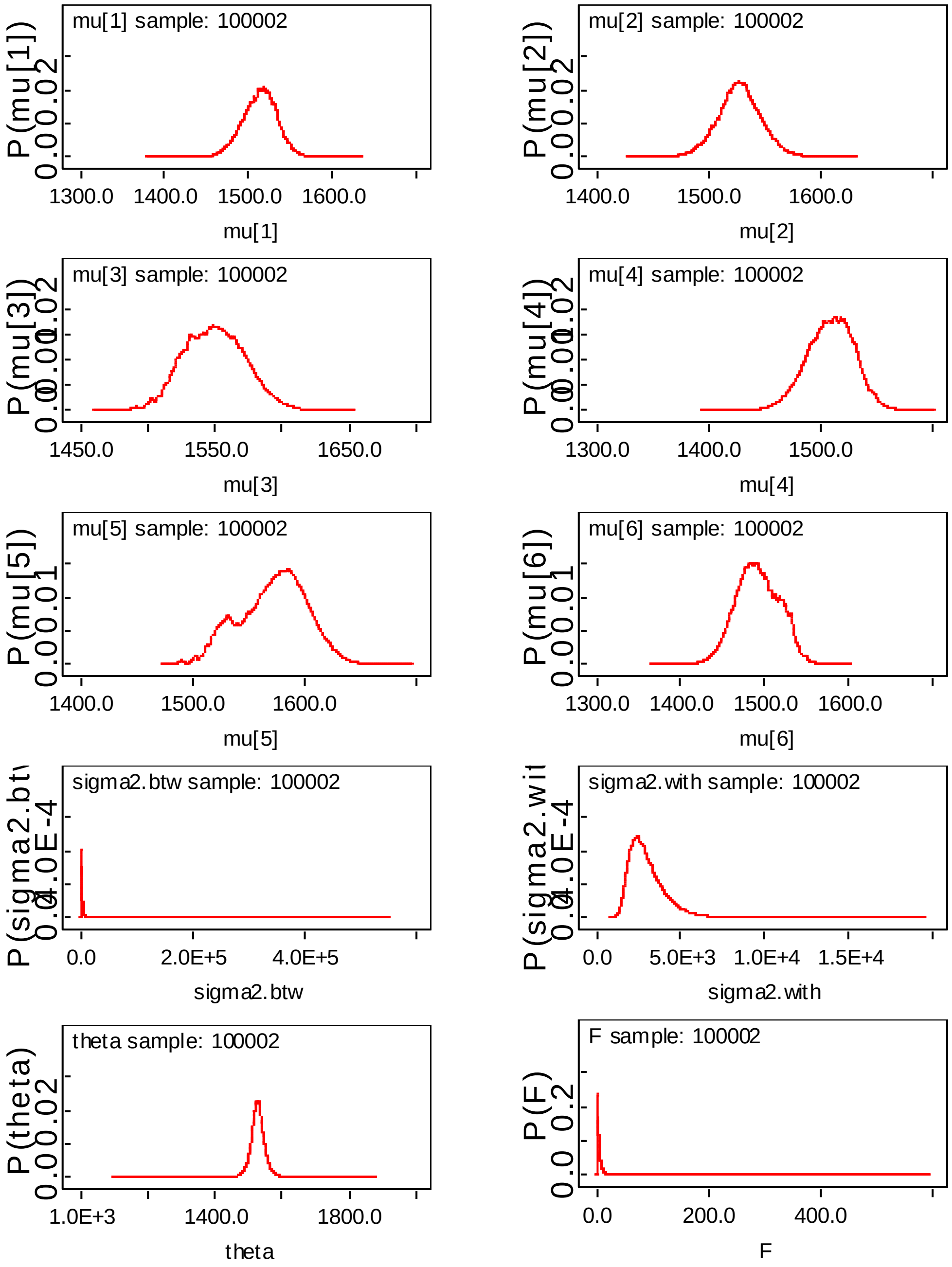

Figure 4.3: MCMC Posterior densities for the parameters.

These plots are like smoothed histograms. Instead of counting the estimates into bins of particular widths like a histogram, the effect of each iteration is spread around the estimate via a Kernel function e.g. a normal distribution. This means that at each point we get the sum of the Kernel function parts for each iteration.

4.4 Tests for Convergence

4.4.1 Time Series Trace Plots

First we checked whether the Markov chains have indeed reached a stable equilibrium distribution, i.e., have converged. Figures 4.4a-4.4j show the time series trace plots of the posterior means and the posterior variances.

TRACEPLOTS

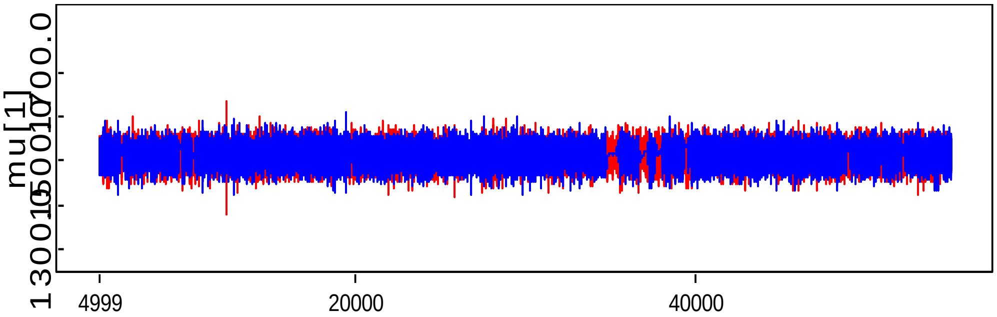

Figure 4.4a: Trace plot test for convergence for posterior mean of batch one.

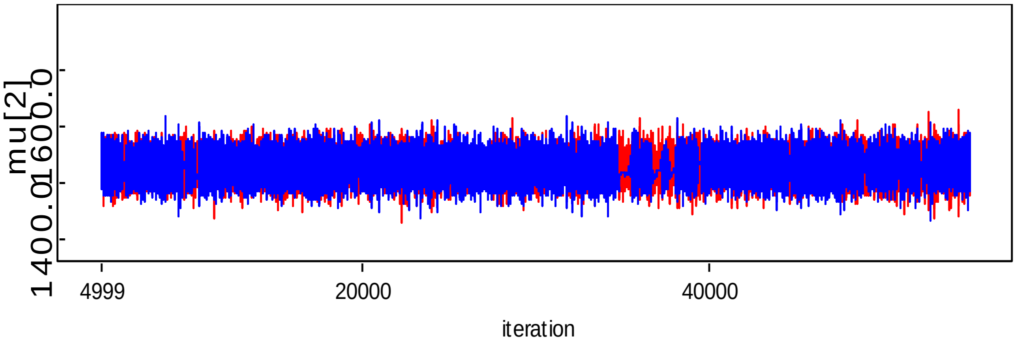

Figure 4.4b: Trace plot test for convergence for posterior mean of batch two.

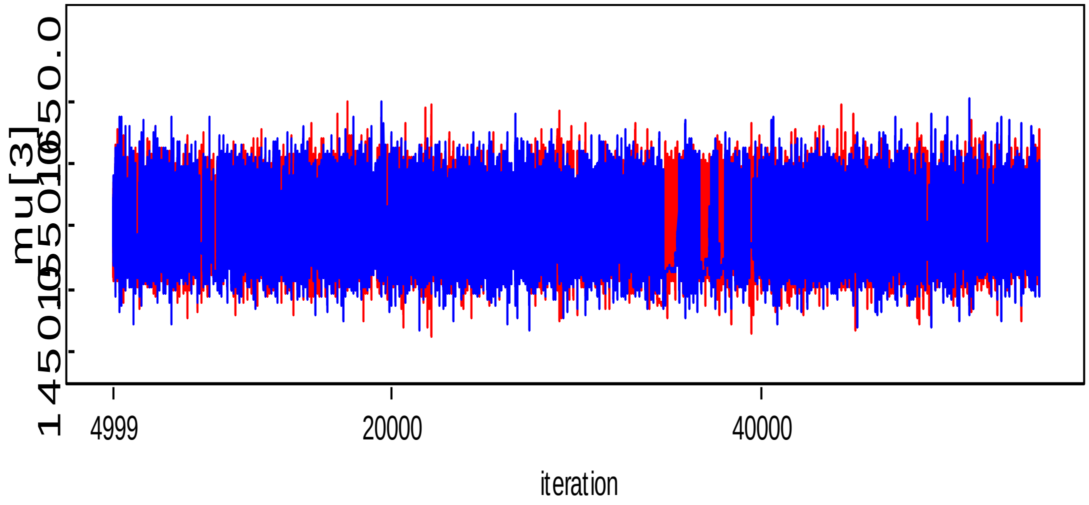

Figure 4.4c: Trace plot test for convergence for posterior mean of batch three.

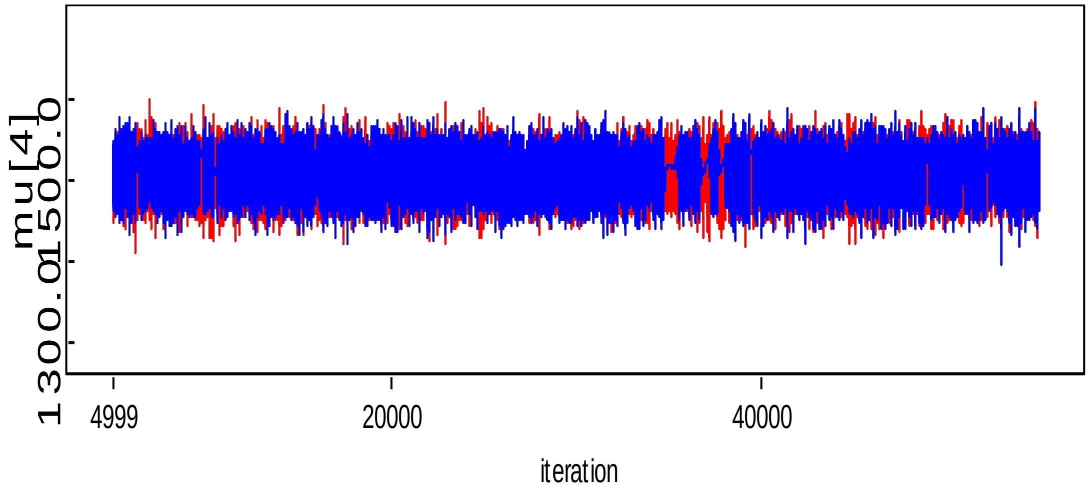

Figure 4.4d: Trace plot test for convergence for posterior mean of batch four.

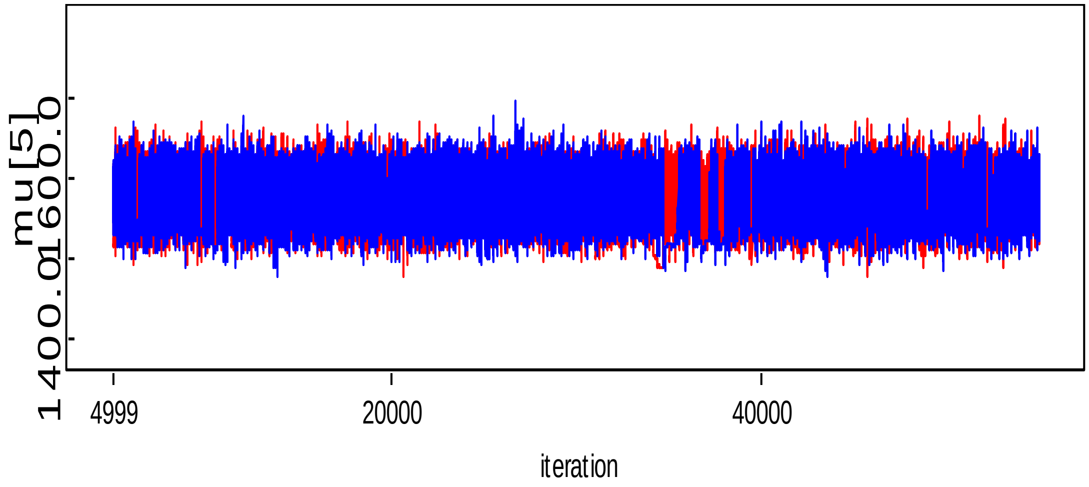

Figure 4.4e: Trace plot test for convergence for posterior mean of batch five.

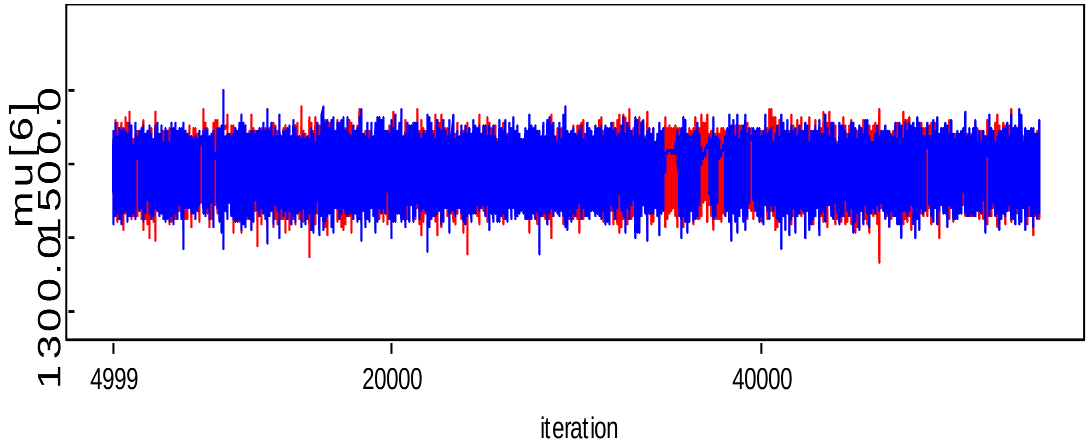

Figure 4.4f: Trace plot test for convergence for posterior mean of batch six.

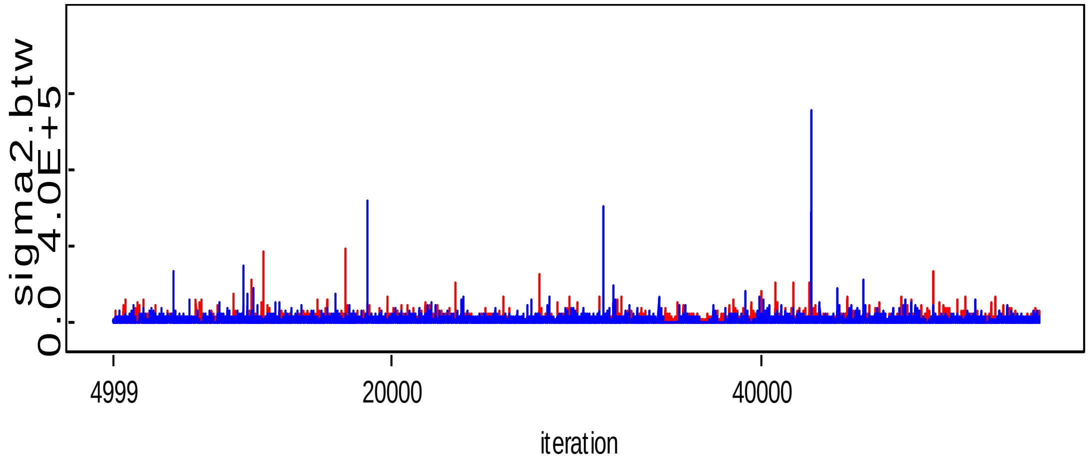

Figure 4.4g: Trace plot test for convergence for posterior between variance.

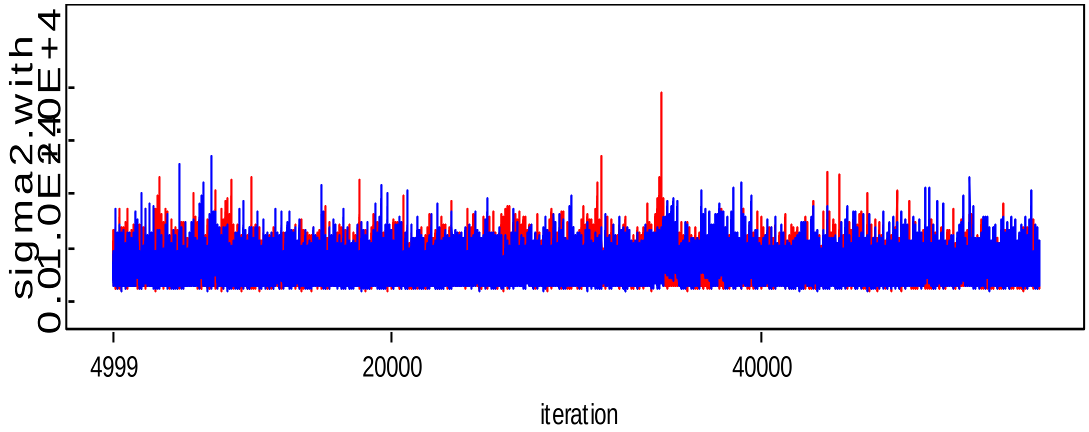

Figure 4.4h: Trace plot test for convergence for posterior within variance.

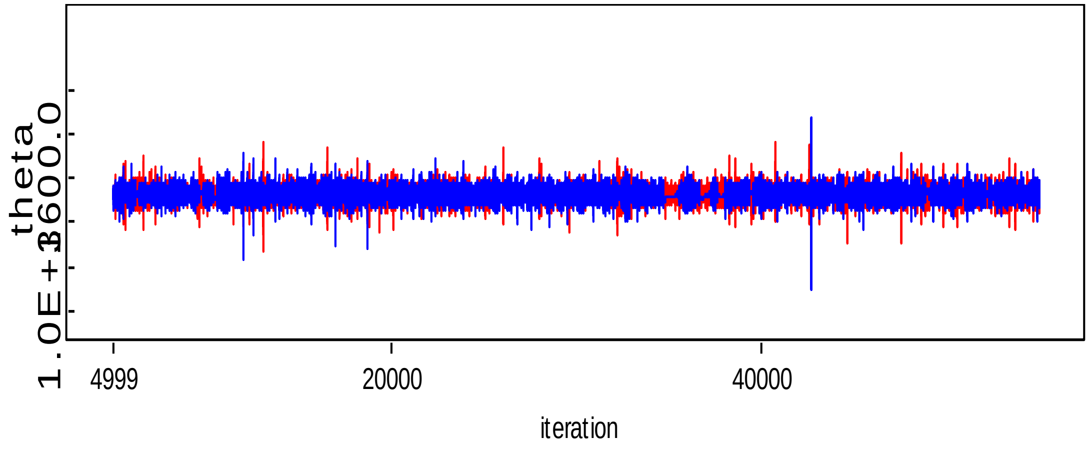

Figure 4.4i: Trace plot test for convergence for posterior grand mean

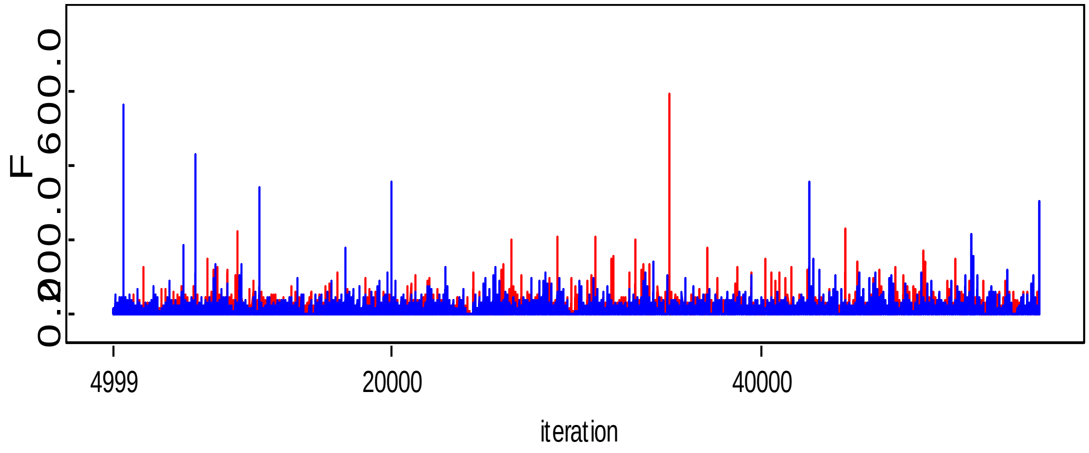

Figure 4.4j: Time series trace plot for convergence for posterior F-value

It was apparent from these plots that the effects of the initial values of the parameters and initial data took a while before the process begun to appear stationary. However, plots of the mean of the parameters against the number of iterations of the sampler produced smoother plots than did the raw sample values and could make it easier to identify and understand any non-stationarity.

4.4.2 Gelman–Rubin (BGR) diagnostic statistic

These were done visually or by inspecting the Brooks Gelman–Rubin (BGR) diagnostic statistic that WinBUGS displays. Values around 1 indicate convergence, with 1.1 considered as acceptable limit by (Gelman and Hill, 2007).

DIAGNOSTIC PLOTS

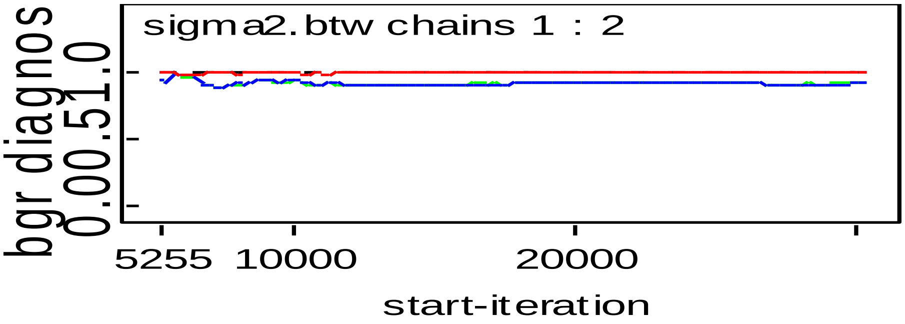

Figure 4.5a: Diagnostic plot for between variance

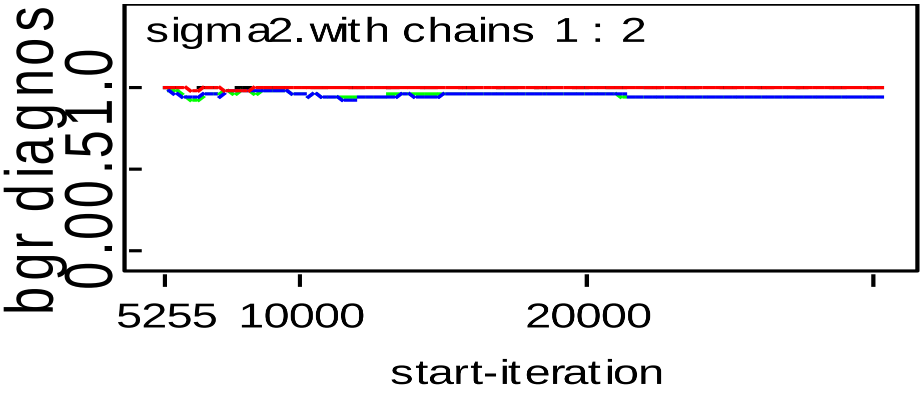

Figure 4.5b: Diagnostic plot for between variance

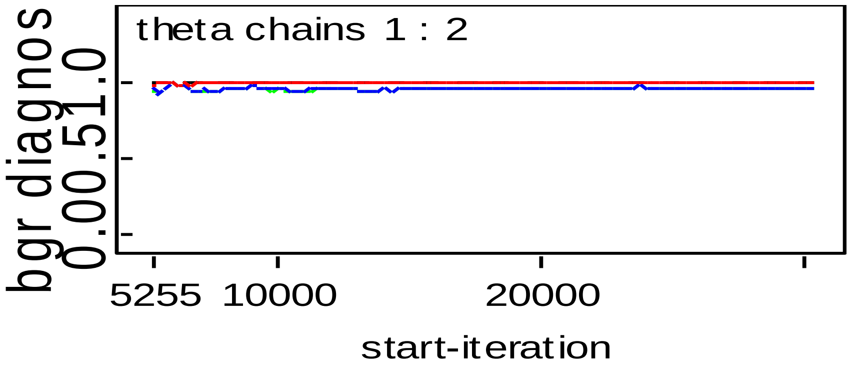

Figure 4.5c: Gelman-Rubin statistic diagnostic plot as a measure of convergence for the posterior grand mean.

All the plots suggested that convergence was achieved after 5,000 iterations of the sampler. Therefore the first 5,000 draws (the burn-in) were removed and the remaining 100,002draws were used to conduct the subsequent analysis.

These values are close to 1 indicating convergence. Visual inspection of the time series plot produced by the trace diagrams below again suggests that the Markov chains have converged

4.4.3Autocorrelation plots

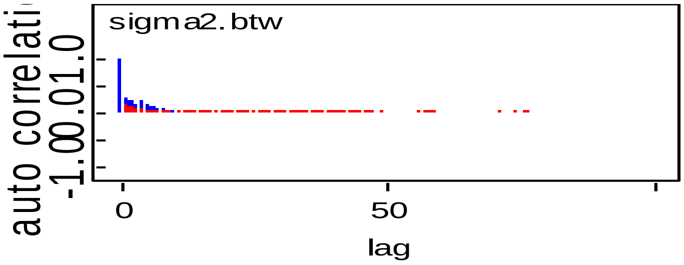

Fig 4.6a: Autocorrelation plot for posterior between variance

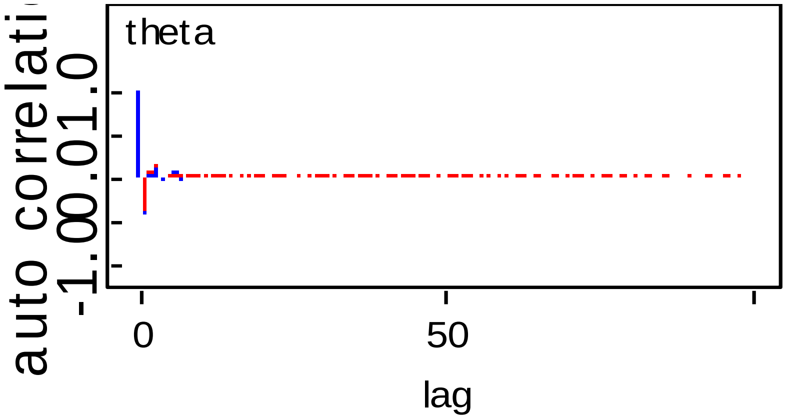

Figure 4.6b: Sampler Autocorrelation plot for posterior between variance

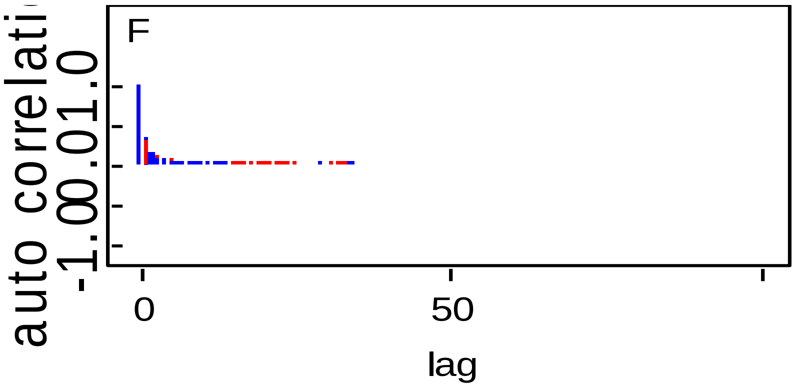

Figure 4.6c: Sampler Autocorrelation plots for the simulated parameters

The Autocorrelation function (ACF) measures how correlated the values in the chain are with their close neighbours. The lag is the distance between the two chains to be compared.

An independent chain will have approximately zero autocorrelation at each lag.

The plot of autocorrelations against the lag revealed that the correlations of the draws diminished as the chain progressed in length. The diminishing autocorrelations was a clear indication that convergence had been achieved and hence the draws were regarded as the draws from the target distribution.

CHAPTER FIVE

DISCUSSION OF RESULTS

5.1 Introduction

This study gives a summary to Bayesian testing analysis of variance and how it is conducted in practice using simulation-based methods, i.e MCMC and Gibbs sampling. WinBUGS (the MS Windows operating system version of BUGS: Bayesian Analysis Using Gibbs Sampling) is a versatile package that has been designed to carry out Markov chain Monte Carlo (MCMC) computations for a wide variety of Bayesian models.

After a burn-in of 5,000 draws (the first 5,000 draws from each Markov chain are discarded as not representative of the stationary distribution of the chain i.e the posterior distribution of the parameters in the model) and a further 100,002 iterations for each chain, the MCMC produced the summary statistics for the samples as shown in Table 4.4 and Table 4.5 and Table 4.6 respectively. As a Bayesian point estimate, typically the posterior means or the posterior medians (or sometimes also the mode), were reported in these tables, while the posterior standard deviation was used as a standard error of the parameter estimate. The range between the 2.5th and 97.5th percentiles represents a 95% Bayesian confidence interval and is called a credible interval.

Numerical summaries of the model using different priors appear in Table 4.4, Table 4.5 and Table 4.6, for the posterior grand mean μ, the “between” variance (ω^2) and the “within”, variance (σ^2). The left column summarizes the results of the WinBugs run, showing the mean of the MCMC output for each of the parameters, the standard deviation, and an estimate of the 95% HDR of the marginal posterior density of each parameter.

Assessing these plots indicates that the parameter traces look like straight hairy colorful caterpillars, with the two chains fluctuating rapidly around their equilibrium, and that there are no obvious upward or downward trends. Besides, the autocorrelation plots show little correlations, and kernel density plots show bell-like posterior distributions, and the Gelman-Rubin statistic show that the ratio of between to within variability is close to 1. All plots assume us that the model is converged.

These posterior point estimates give results similar to those obtained when using the classical or frequentist approach.

CHAPTER SIX

CONCLUSIONS AND RECOMMENDATIONS

6.1 Conclusion

In this thesis, Bayesian approach to hypothesis testing for ANOVA was investigated. Posterior means, within-variance, between-variance and Posterior F-values were also obtained. Specifically, inferences when following the Bayesian approach to analyzing this problem was based on 95% credible sets. This approach provides a more natural form of the inference for this problem than the likelihood testing in frequentist approach which relies on asymptotic theory. An F-Value of 4.598 was obtained using the classical approach. This is shown in ANOVA table 4.3. Posterior means were illustrated in tables 4, 5 and 6 for the different priors. Posterior F-value of 4.56 was obtained for normal priors for means and conjugate inverse Gamma for the variances. Posterior F-value of 4.62 was obtained using Zellner-g prior (g=n=30). Posterior F-value of 4.52 was obtained using Zellner-g prior (g=k^2=30). Posterior point estimates, i.e, means, modes and medians for posterior means mu's and posterior variances are also shown in tables 4.4,4.5 and 4.6.The results indicated that the results obtained using the classical and those of Bayesian approach are similar.

It was also shown that posterior values for the means, variances and F-values yielded values that closely agree with those obtained using the Classical Approach.

6.2 Recommendations

Application of Bayesian methods in analysis of variance should be used as it gives exact posterior point estimates. Estimation of parameters using Bayesian methods also allow us to include estimation of uncertainty that go along with these parameters. This work can be extentended to multi-facor ANOVA models. Such models can be fitted using dummy variables and constraints (corner constraints and sum-to-zero constraints) must be introduced in order to carry out analysis in Multifactor ANOVA designs.

More importantly, WinBUGS extends the SAS functionality. It can be used to fit a great variety of linear and nonlinear models, including Generalized Linear Models, categorical, and survival models, with or without random effects.

REFERENCES

Anscombe F, Aumann R. (1963). “A Definition of Subjective Probability. "*The Annals of Mathematical Statistics*, **34(1),** 199-205

Berger, J. O. (2006). “The Case for Objective Bayesian Analysis,” *Bayesian Analysis*, **1**, 385–402

Berger,J. O. and Malarpady, M. (1987). “Testing Precise Hypotheses,” *Statistical Science*, **2**, 317–352.

Berger, J. O. and Sellke, T. (1987). “Testing a Point Null Hypothesis: The Irreconcilabilty of p- Values and Evidence,” *Journal of the American Statistical Association*, **82**, 112-139.

Bernardo, J. M. (2003). Bayesian Statistics. Encyclopedia of Life Support Systems (EOLSS). *Probability and Statistics*. UNESCO, Oxford, UK. http://www.uv.es/~bernardo/BayesStat2.(accessed on 1st june 2012).

Bernardo, J. M. and Smith, A. F. M. (2000). *Bayesian Theory*, New York: Wiley.

Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian Theory*, New York: Wiley

Box, G. E. P. and Tiao, G. C. (1973). Imperical statistical analysis, 112-127.

Brooks, S.P. (2003). Bayesian computation: a statistical revolution. Philos. Trans. R. Soc. Lond. A. **361**, 2681- 2697.

Cox, R. T. (1946). Probability, frequency and reasonable expectation. *American Journal of Physics*, **14**, 1 -13.

De~ Finetti, B., 1937. *Probabilty induction and statistics: the art of guessing. New york Wiley.*

Dellaportas, P. and Smith, D. (1993). Bayesian analysis of Errors-in-Variables Regression Models. *Biometrics*, **51,** 1085–1095.

Dickey, J. M. (1971). “The Weighted Likelihood Ratio, Linear Hypothesis on Normal Location Parameters,”*The Annals of Mathematical Statistics,* **42**,204-223.

Fernandez, C., Ley, E., and Steel, M. (2001). “Benchmark Priors for Bayesian Model Averaging,” *Journal of Econometrics,* **100,** 381-427

Foster, D. and George, E. (1994). “The Risk Inflation Criterion for Multiple Regression,” *The Annals of Statistics,* **22,** 1947-1975.

Gelfand, A.E., Schmidt, A.E., Wu, S., Silander Jr., J.A., Latimer, A., Rebelo, A.G., (1992). Modelling species diversity through species level hierarchical modelling. Appl. Stat. **54**, 1- 20.

Gelman, A. (2008). “Objections to Bayesian Statistics,” *Bayesian Analysis***, 3,** 445–450.

Gelman, A.(2006). Prior distributions for variance parameters in hierarchical models. Bayesian Anal, **1,** 514 -534.

Gelman, A. (2005). Analysis of variance: why it is more important than ever (with discussion). Ann. Stat, **33**, 1- 53.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). Bayesian Data Analysis (2nd ed.), Boca Raton (FL): Chapman & Hall/CRC.

Gelman, A., Hill, J.(2007). Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press, Cambridge.

Gelman, A., Meng, X. L., Stern, H.S. (1996). Posterior predictive assessment of model fitness via realized discrepancies (with discussion). Stat. Sin. **6,** 733 -807.

Geman, S., (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Trans. Pattern. Anal. Mach. Intell**. 6,** 721 -741.

Gilks, W.R., Thomas, A., Spiegelhalter, D.J.(1994). A language and program for complex Bayesian modeling. Statistician **43**, 169 -178.

Hastings, W.K. (1970). Monte Carlo sampling methods using Markov chains and their applications. Biometrika **57,** 97-109.

Gilks, W.R., Thomas, A., Spiegelhalter, D.J. (1994). A language and program for complex Bayesian modeling. Statistician **43**, 169 178

Hastings, W.K. (1970). Monte Carlo sampling methods using Markov chains and their applications. Biometrika, **57,** 97- 109.

Jeffreys, H. (1961). *Theory of Probability*, Oxford, UK: Oxford University Press.

Kass, R. E. and Wasserman, L. (1995). “A Reference Bayesian Test for Nested Hypotheses and its Relationship to the Schwarz Criterion,” *Journal of the American Statistical Association*, **90**, 928–934.

Kass, R. E. and Raftery, A. E. (1995). “Bayes Factors,” *Journal of the American Statistical Association*, **90,** 377–395.

Kaufman, C. G. and Sain, S. R. (2010). “Bayesian Functional ANOVA Modeling Using Gaussian Process Prior Distributions,” Bayesian Analysis, **5,** 123–150.

Liang, F., Paulo, R., Molina, G., Clyde, M., and Berger, J. (2008). “Mixtures of g Priors for Bayesian Variable Selection,” *Journal of the American Statistical Association,* **103**,410-423.

Le Cam, L.(1990). Maximum likelihood: an introduction. Int. Stat. Rev. **58,** 153-171.

Lindley, D. V. (2006). Understanding Uncertainty. Wiley, Hoboken, NJ.

Lindley, D. V. (2000). “The Philosophy of Statistics,” The Statistician, **49,** 293-337.

Lindley, D. V. (1980). Theory and practice of Bayesian statistics. Statistician **32,** 1- 11.

Link, W.A., Barker, R.J. (2010). Bayesian Inference with Ecological Examples. Academic Press, San Diego, CA.

Link, W.A., Cam, E., Nichols, J.D., Cooch, E.G.(2006). On BUGS and birds: Markov chain Monte Carlo for hierarchical modeling in wildlife research. J. Wildlife Manage. **66,** 277- 291.

Link, W.A., Sauer, J.R., (2002). A hierarchical analysis of population change with application to Cerulean warblers. Ecology, **83**, 2832- 2840

Mazzetta, C., Brooks, S., Freeman, S.N. (2007).On smoothing trends in population index modeling. *Biometrics* ,**63**, 1007- 1014.

McCarthy, M.A., Masters, P. (2007). Profiting from prior information in Bayesian analyses of ecological data. *Journal on Applications to Ecology*.**42**, 1012-1019.

McCullagh, P., & Nelder, J. A. (1989). Generalised linear models (2nded.). London: Chapman & Hall.

O’Hagan, A. and Forster, J. (2004). *Kendall’s Advanced Theory of Statistics vol. 2B: Bayesian Inference (2nd ed.)*, London: Arnold.

Poirier, D. J. (2006).“The Growth of Bayesian Methods in Statistics and Economics Since 1970,” *Bayesian Analysis*, **1,** 969–980.

Press, S., Chib, S., Clyde, M., Woodworth, G., and Zaslavsky, A. (2003). *Subjective and Objective Bayesian Statistics: Principles, Models, and Applications*, Wiley-Interscience.

R Development Core Team. (2007). R: A language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.(accessed on 4th February 2013)

Robert, C. (1993). “A Note on Jeffreys-Lindley Paradox,” *StatisticaSinica*, **3**, 601–608.

Roberts, I. (2007). *Bayesian Modeling Using WinBUGS*, Hoboken, NJ: Wiley.

Swain, D.P., Jonsen, I.D., Simon, J.E., Myers, R.A.(2009). Assessing threats to species at risk using stage structured state space models: mortality trends in skate populations. *Ecol.Appl.* **19,** 1347- 1364.

Smith, A.F.M., Gelfand, A.E. (1992). Bayesian statistics without tears: A sampling resampling perspective. *Am. Stat.* **46**, 84- 88.

Williams, B.K., Nichols, J.D., Conroy, M.J. (2002). Analysis and Management of Animal Populations. *Academic Press*, *San Diego, CA*.