BENIN JOURNAL OF
STATISTICS
ISSN 2682-5767
Vol. 4, pp. 53– 74 (2021)

On the Use of Bayes Factor and p-Value in Hypothesis Testing

A. H. Ekong[1*], **O. E. Asiribo**[2] **and G. A. Dawodu**[3]

[1,2,3]*Department of Statistics, Federal University of Agriculture Abeokuta, Ogun State, Nigeria*

(*Received: 18 December 2019; accepted: 03 June 2020*)

Abstract. The focus in experimental data analysis is to assess treatments-mean structure and such decision is taken by the reported p-values from the corresponding F-tests. However, there are concerns about the use of p-value as a means of making decision in tests of hypotheses, hence the motivation for this study which proffers Bayes factor as a Bayesian alternative. Constraints were placed on effect parameters of a one-way model by assuming that the standardized random effects followed a normal distribution with mean zero and level-variance, g. Jefferys prior was placed on the mean and the error variance, while the inverse chi-squared with one degree of freedom was chosen for g. Simulated data were analysed to contrast Bayes factor with p-value, using different sizes of samples and effects. A one-way design data set from an experiment on the effect of genetic strain on fecundity of fruit fly Drosophila melanogaster was also analysed for multiple comparisons of treatment means using Bayes factors, and the posterior sample estimates using Markov chain Monte Carlo simulation were used to validate Bayes factors results.

Keywords: p-value, Bayes factor, ANOVA, multiple comparison, Bayesian.

Published by: Department of Statistics, University of Benin, Nigeria

1. Introduction

Publications on Bayesian alternative to the classical hypotheses tests dates back to the work of Jeffreys (1935, 1961) who developed the Bayesian approach for scientific inference. Jefferys was concerned with the comparison of predictions made by two competing scientific theories or hypotheses. In his approach, statistical models were introduced to represent the probability of data according to each of the two theories and Bayes' theorem was used to compute the posterior

*Corresponding author. Email: anieekong@outlook.com

probability that one of the theories is correct using the quantity Bayes factor. Later on, Kass et al., (1995) studied some practical aspects of the Bayesian technique such as *"how conclusion may be drawn from them, how they can provide answers when non-Bayesian methods are hard to construct, what their strengths and limitation are", as well as "accounting for uncertainty in model-building process, which is closely connected to the methodology for hypothesis testing"* (Kass et al., 1995). The work of Raftery (1995) was a follow up to Kass et al., (1995) with focus on the Bayesian Information Criterion (BIC) or Schwarz Criterion, which Kass et al., (1995) noted was an asymptotic approximation to a Bayes factor with uniform priors. Weakliem (1999) criticized Raftery (1995) on the choice of prior. One of the criticisms of Weakliem (1999) was that the uniform prior set on variance of effect sizes was more spread out priors making it conservative in the sense that they tend to favor the null hypothesis more and hence find less evidence for an "effect" of interest in a study. The BIC approximation was for regression settings and this led Wagenmakers (2007) to show the relationship between the BIC and the ANOVA. The relationship according to Wagenmakers (2007) stemmed from the point that BIC can be estimated from the proportion of unexplained variance which is readily obtained from ANOVA table's sums of squares. Yet the criticisms about the uniform prior and the sample size parameter in regression which are not clearly defined in designed experiments were still untreated. Rouder et al. (2012) dealt with the issue of choice of prior for the Bayes factor by looking at default priors for ANOVA designs and regression models on effects for designs in experimental psychology.

1.1 *P-value in tests of hypotheses*

Analysis of variance is a standard method for describing and estimating heterogeneity among the means of a response variable across levels of multiple factors. In most experimental settings, ANOVA is used to test the presence of treatment effects (Koech et al., 2014). In other words, the interest is to know if the levels-means are all the same, or some of them differ. ANOVA partitions the variability in a data set into parts associated with different mean structures plus an error structure. When in addition, the error structure for the data is independent normal with constant variance, the information provided by an ANOVA can also be used to construct tests for comparing the different models of different mean structures that are represented in the ANOVA (Oehlert, 2010). The ANOVA investigates the variation in measurements taken on the observations as explained by the grouping introduced by the classification factor(s).

The mean square for error (MSE) in ANOVA is a random variable which depends on the random errors in the data. If an experiment were repeated, there would be different random errors and thus a different mean square for error. However, the expected value of the mean square for error, averaged over all the possible outcomes of the random errors, is the variance of the random errors,

σ^2. Thus, the mean square error estimates the error variance, no matter the values of the treatment effects, $\alpha_i's$, assuming all possible outcomes of the errors have been gotten through a hypothetical replication (all exhaustive cases) of the experiment.

When a null hypothesis is true, both MSE and MSTrt (mean square for treatment) vary around σ^2, so their ratio (the F-statistic) is about one. When a null hypothesis is false, MSTrt tends to be bigger than σ^2, and the F-statistic tends to be bigger than one. Thus the null hypothesis is rejected for sufficiently large values of the F-statistic. This probability is called p-value or observed significance level of the test (Oehlert, 2010).

Conventional ANOVA allows decisions about experiments to be made by reporting p-values from the corresponding F-tests as evidence for or against certain theoretical positions about the system or process under experimentation. There are concerns about the use of p-value as a means of making decision in hypothesis testing and model selection; for instance, that its computation is based on hypothetical data used in calculating the expected values of MSTrt and MSE, and hence the test statistics (Wagenmakers, 2007).

In what follows the concerns with p-value in hypotheses testing are presented:

(1) Since small p-value indicates evidence of inconsistency of data with the null, one is often uncertain about the real implication about the hypothesis or the assumptions made. For instance, while a low p-value implies that the data are unlikely assuming a true null, it cannot evaluate which competing case is more likely: that is the null is true, but the selected sample was unusual, or the null is false.

(2) Another fundamental concern with p-value is that it is based on data never observed but hypothetically replicated. These hypothetical data are data expected under the null, without which it is impossible to construct the sampling distribution of the test statistic from which the p-value is calculated (from the definition of p-value) (Wagenmakers, 2007).

(3) When a test is significant, the null hypothesis is rejected but never affirmed. If a null is true the best case outcome is a statement about a lack of significant evidence against the null (i.e., there is no significant evidence for an effect), even though it is desirable to state a positive evidence for a lack of an effect. (Rouder et. al., 2012).

(4) The evidence of effect or validity of a model in an experiment can be assessed by the probability that such effect or model is true given that some result has been observed. This is in contrast to the Frequentists p-value, which gives the probability of observing a result conditioned on the assumption of the truth of a hypothesis.

(5) It is prone to misinterpretation. For instance, given a test with a p-value of 0.04, this means that assuming there is no effect; you would get the observed difference or more in 4 % of studies due to random sampling

error. The p-value is not an error rate, neither is it the probability of an effect.

Siegfried (2016) noted that criticism about p-values-statistical measures widely used to analyse experimental data in most scientific disciplines, has finally reverberated loudly enough for the scientific community to listen, with the publication by the American Statistical Association (ASA) in 2016 on the concerns with p-values and their widespread misuse, as well as few approaches to supplement or even replace p-values like credibility, or prediction intervals; Bayesian measures of evidence, such as likelihood ratios or Bayes factors; among others. Siegfried (2016) noted that p-values may be valid and useful under certain specific circumstances, but those circumstances are rarely relevant in most experimental contexts.

In view of the aforementioned, this study focuses on Bayes factor as a plausible Bayesian alternative to the Frequentists p-value, particularly for the balanced one-way analysis of variance designs and its application in multiple comparisons of means, and its veracity by posterior distributions.

2. Materials and Method

2.1 *Re-parameterisation of model parameters*

Consider the re-parameterisation of the one-way ANOVA model (Rouder et. al., 2012):

$$\mathbf{y} = \mu\mathbf{1} + \sigma\mathbf{X}\beta + \epsilon \tag{1}$$

where μ is grand mean parameter, $\mathbf{1}$ is a column vector of 1's with length N, $\mathbf{X}$ is an Nxb design matrix, ϵ is a column vector of error terms of length N : $\epsilon \sim N(0, \sigma^2\mathbf{I})$ and β is the b-levels re-parameterised vector of effect size or standardized effects relative to the standard deviation of the error, σ. In equation (1), there are a total of $b + 1$ parameters that describe the b-cell means, and this implies that the $b + 1$ parameters are not uniquely determined; hence the model may not be identifiable. In classical statistics, constraints are added for uniqueness which in turn reflects whether effects are treated as fixed or random, depending on what additional constraint is added. Re-parameterizing the model in terms of effect size does not change the basic nature of the role of the prior on effects. The impact of effect-size parameterization is that we have a basic scale about the ranges of effect sizes that applies broadly across different tasks and populations.

In the effect-size parameterization, the noninformative prior for μ and σ^2 is called Jeffreys prior, with equal weights on all values. According to Jeffreys

(1961), it is given as

$$\psi(\mu, \sigma^2) \propto \frac{1}{\sigma^2} \tag{2}$$

The proper prior on the effect size parameter β_i is assumed to be $\beta_i \propto N(0, g)$. The important question is how g, the variance of effect sizes, should be set. We note that for realistic values of g, the β_i cannot be unpredictable, that is to say, for realistic variation among the levels of a factor, the difference between the effect of one level from that of the other levels of the same factor cannot be inconsistent. For instance, we may disagree strongly that the effect of a fertilizer (with three levels) on the yield in height of a crop would result in two levels having yield of height in centimetres unit, while the third yields height in metres unit. So we can agree with Rouder, et al. (2016) that as long as the variance of the effects is finite the model is identifiable and estimable.

It is therefore more realistic to allow g to range over a distribution of values and hence a prior is needed for it. Zellner et al., (1980) recommended the inverse-χ^2 with one degree of freedom in equation (3), as a convenient and flexible choice of prior for g, which is less informative than the unit-information prior:

$$\theta(g) = (2\pi)^{-1/2} g^{-3/2} \exp^{-1/2g} \tag{3}$$

with a marginal prior on β_i as a univariate Cauchy with probability density function given as:

$$\psi(x) = \frac{1}{(1 + x^2)\pi} \tag{4}$$

These priors are designed to minimize assumptions about the range of effect size, i.e., reflect a minimum degree of information, and in this sense they are objective priors.

With $\beta = (\beta_1, ..., \beta_b)'$, the vector of b standardised effects, the multivariate Cauchy is given with joint pdf as (Kotz, et. al., 2004):

$$\psi(\beta) = \frac{\Gamma[(1 + b)/2]}{\Gamma(1/2)\pi^{b/2}[1 + \Sigma_{i=1}^{b} \beta_i^2]^{(1+b)/2}} \tag{5}$$

The main characteristic of this multivariate Cauchy form is the dependence among the effects such that they vary on a similar scale, and are not arbitrarily different, which is the case for the random effects with constant variance g.

2.2 Bayes factor for one-way balanced ANOVA

Here we look at the Bayes factor that quantifies the evidence coming from the observed data for competing hypotheses. Assuming two competing models M_0 and M_1 defined to be single mean and separate means respectively, for the balanced one-way design with b factor-levels. Let the two models be defined as Bayarri et al., (2007)

$$M_0 : \mathbf{y} = \mu\mathbf{1} + \epsilon \equiv f_0(\mathbf{y}|\beta, \sigma) = N(\mathbf{y}|\mathbf{X}\beta, \sigma^2); H_0 : \mathbf{C}\beta = 0 \tag{6}$$

$$M_1 : \mathbf{y} = \mu\mathbf{1} + \mathbf{X}\beta + \epsilon \equiv f_1(\mathbf{y}|\beta, \sigma) = N(\mathbf{y}|\mathbf{X}\beta, \sigma^2); H_1 : \mathbf{C}\beta \neq 0 \tag{7}$$

where $\mathbf{X} : Nb$ matrix of rank r with rlb and $\mathbf{C} : db$ matrix of rank d.

Given the priors for the balanced one-way ANOVA designs; standardized b random effects $\beta \propto N(\mathbf{0}, g\mathbf{I})$, where the variance of effects, $g : g \sim Inverse - \chi^2(1)$ (equation 3) under M_1, we have

$$f(\mathbf{y}) = \frac{1}{(2\pi)^{N/2}(\sigma^2)^{N/2}} exp\left\{ -\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{1}\mu)'(\mathbf{y} - \mathbf{1}\mu) \right\} \tag{8}$$

$$\pi(\beta) = \frac{1}{(2\pi)^{b/2}(\sigma^2)^{b/2}(g\mathbf{I})^{b/2}} exp\left\{ -\frac{1}{2\sigma^2}(g\beta'\beta) \right\} \tag{9}$$

$$\theta(g) = (2\pi)^{-1/2} g^{-3/2} \exp^{-1/2g} \tag{10}$$

Recall the Jeffreys prior on μ and σ^2,

$$\psi(\mu, \sigma^2) = \frac{1}{\sigma^2} \tag{11}$$

Then the likelihood function of β under M_1 is

$$P_{M_1}(D|\beta) = \int\int\int f(\mathbf{y})\pi(\beta)\psi(\mu, \sigma^2) d\beta d\mu d\sigma^2 \tag{12}$$

$$P_{M_1}(D|\beta) = \int\int\int \frac{1}{(2\pi)^{(N+b)/2}(\sigma^2)^{(N+b+2)/2}(g\mathbf{I})^b/2} exp\left\{ -\frac{1}{2\sigma^2}(g\beta'\beta \right.$$

$$\left. + (\mathbf{y} - \mathbf{1}\mu)'(\mathbf{y} - \mathbf{1}\mu)) \right\} d\beta d\mu d\sigma^2 \tag{13}$$

$-\infty < \mu < \infty,\ 0 \leq \sigma^2 < \infty$. The marginal density of the data given M_1 after integrating with respect to μ and σ^2 then becomes

$$P(D|M_1) = \int \frac{1}{(2\pi)^{(N+b)/2}(\sigma^2)^{(N+b+2)/2}(g\mathbf{I})^b/2} exp\Bigg\{ -\frac{1}{2\sigma^2}(g\beta'\beta$$

$$+ (\mathbf{y} - \mathbf{1}\mu)'(\mathbf{y} - \mathbf{1}\mu)) \Bigg\} \theta(g)dg \tag{14}$$

For M_0, $g \sim 0$, with Jeffreys prior on μ and σ^2 and $f(\mathbf{y})$, we have the marginal density of the data given M_0 as

$$P(D|M_0) = \frac{1}{(2\pi)^{N/2}(\sigma^2)^{(N+2)/2}} exp\Bigg\{ -\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{1}\mu)'(\mathbf{y} - \mathbf{1}\mu) \Bigg\} \tag{15}$$

The extent to which the data D, support M_1 over M_0 is evaluated by the ratio of marginal densities of the data given M_1 and that M_0, given by the Bayes factor, B_{10}

$$B_{10} = \frac{P(D|M_1)}{P(D|M_0)} \tag{16}$$

Bayes factor is the resultant odds from dividing the likelihood of one model by the likelihood of another model. Bayes factors are computed by integrating the likelihood with respect to the priors on parameters. These intractable integrals have been implemented in R's *BayesFactor* package by Morey and Rouder (2014).

2.3 *Bayes factor interpretation*

The emphasis of Bayesian hypothesis testing is on the amount by which the data shift one's beliefs in favour of either of the competing hypotheses. For example, a Bayes factor of $B_{10} = 10$ means that the data show support by a factor of 10 in favor of model M_1; likewise, a Bayes factor of $B_{10} = 0.1$ means that the data give support by a factor of 10 in favor of model M_0.

Kass et al., (1995) modification of Jeffreys (1961) scale for interpretation of Bayes factor is used as shown in Table 1.

2.4 *Bayes factor in multiple comparisons of Means*

Refer to Andrade et al., (2017) for the problems with some statistical procedures for multiple comparisons, as well as some procedures with robustness.

Table 1: Grades of interpretation of Bayes factor

Bayes Factor B_{10}	*Evidence against* M_0
1 - 3	Not worth more than a bare mention
3 - 20	Positive
20 - 150	Strong
$>$ 150	Very Strong

Andrade et al., (2017) noted that a fair number of articles take into account the problem of multiple comparisons from the Bayesian point of view. Of note was the Bayesian alternative proposed by Andrade et al., (2010) using a methodology based on a posteriori t-multivariate distribution. The only impediment noted about the Bayesian alternatives was the difficulty to use them due to difficulty with computations, but with software like R, it is a becoming a thing of the past.

Suppose we want to see if $H^{(a,b)} : \mu_a = \mu_b$ for each (a, b) in the set of all possible hypotheses H^{Φ} for the data. The Bayesian approach requires a measure of belief in $H^{(a,b)}$ and the effect size $\delta^{(a,b)} = \frac{(\mu_a - \mu_b)}{\sigma}$, thus the distribution over $(\mu_i)_{i=1\,tot}$, where t is the number of treatment means, consists of two components to reflect the possibility of equal means. Since Bayes factor gives the odd ratio for support of competing hypotheses, the multiple comparisons problem is equivalent to a Bayesian model selection problem by Bayes factor.

Let us consider an example of three treatments means to compare, μ_1, μ_2 and μ_3, the entire spectrum of hypotheses, H^{Φ}, are

$$(NullModel)\;\; H^0 : \mu_1 = \mu_2 = \mu_3$$

$$H^{(1,2)} : \mu_1 = \mu_2 \neq \mu_3$$

$$H^{(1,3)} : \mu_1 = \mu_3 \neq \mu_2$$

$$H^{(2,3)} : \mu_2 = \mu_3 \neq \mu_1$$

$$(FullModel)\;\; H^{full} : \mu_1 \neq \mu_2 \neq \mu_3$$

Bayes factor is used to test any of these hypotheses with equality constraint, what the equality constraint means is that if equality holds, the prediction for the data would not change if the two treatment conditions that are supposed to be the same had exactly the same population distribution (Morey, 2015). With Bayes factor, comparisons are made against the null hypothesis and the resulting Bayes factors can be used to compare the other hypotheses. By so doing we can see, if the treatment means differ, which pair of means are not different.

Still on our example, suppose we want to test if $H^{(1,2)} : \mu_1 = \mu_2$ is true, all we need to do is to test $H^{full}\ vs\ H^0$, the full model against the null so that we can see the evidence (Bayes factor, $B_{f,0}$) for difference in means or equality. Then

we can test $H^{(1,2)}\ vs\ H^0$ by obtaining the Bayes factor $B_{(1,2),0}$. In the case if there is difference in means, that is the Bayes factor $B_{f,0}$ gave evidence from the data in support of the full model, we can then test very simply if $H^{(1,2)} : \mu_1 = \mu_2$ is true by obtaining the Bayes factor $B_{f,(1,2)}$ taking the ratio of their Bayes factor, $B_{f,0}/B_{(1,2),0} = B_{f,(1,2)}$. With this we can obtain the Bayes factors for all the possible models and see how the data supports the comparisons we are making and how each comparison fares with another.

3. Results and Discussion

3.1 *Simulated data analysis for p-values and Bayes factors with different sample sizes*

The following figures show the simulation of ANOVA data set for a one-way design and with each is a brief discussion explaining the resulting p-values and Bayes factors for different sample sizes simulations. The simulation was carried out in R software for a balanced one-way ANOVA with three treatment groups having total sample sizes 9, 15, 30, 75 and 150. 500 iterations for p-value and Bayes factor were computed for each of the sample size scenario simulated.

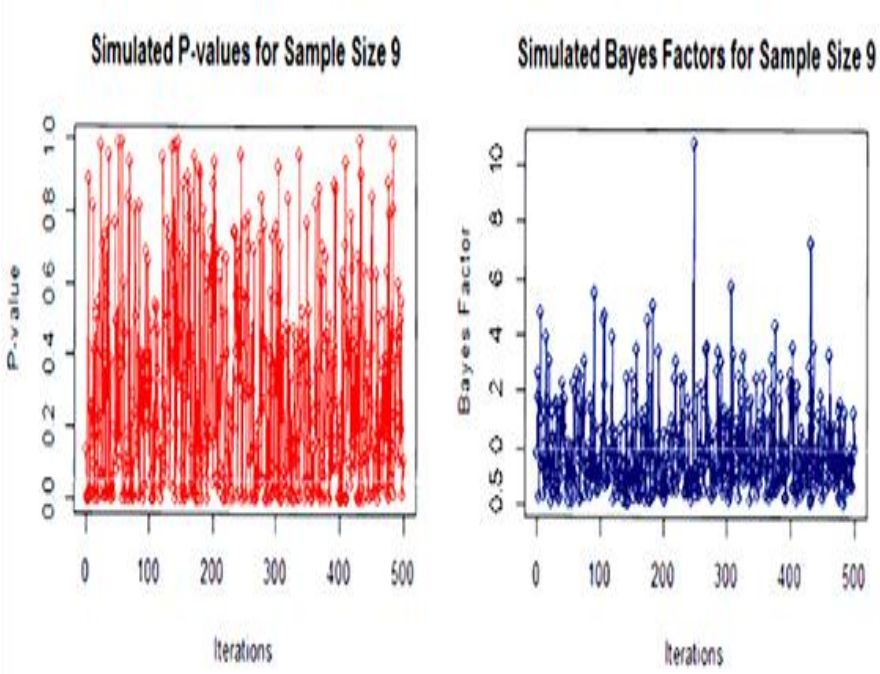

Figure 1: Plots of p-values and Bayes factors with sample size 9

3.1.1 Simulation results for sample sizes

The simulation result is summarized in Table 2.

We note from the forgoing simulation with small sample size that the p-values do not produce much significant result, that is, it fails to reject the null hypothesis and in such case we are left with nothing to say about the alternative

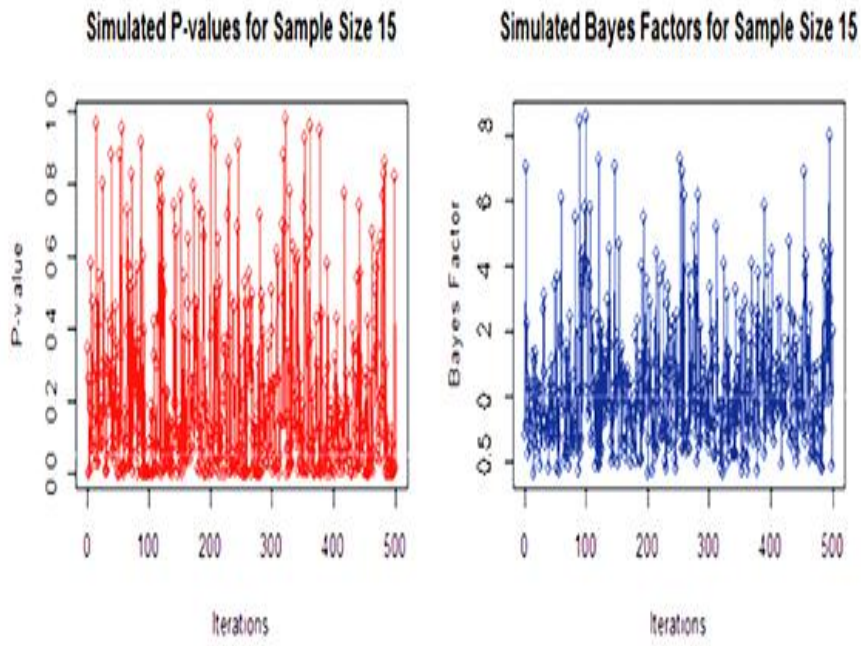

Figure 2: Plots of p-values and Bayes factors with sample size 15

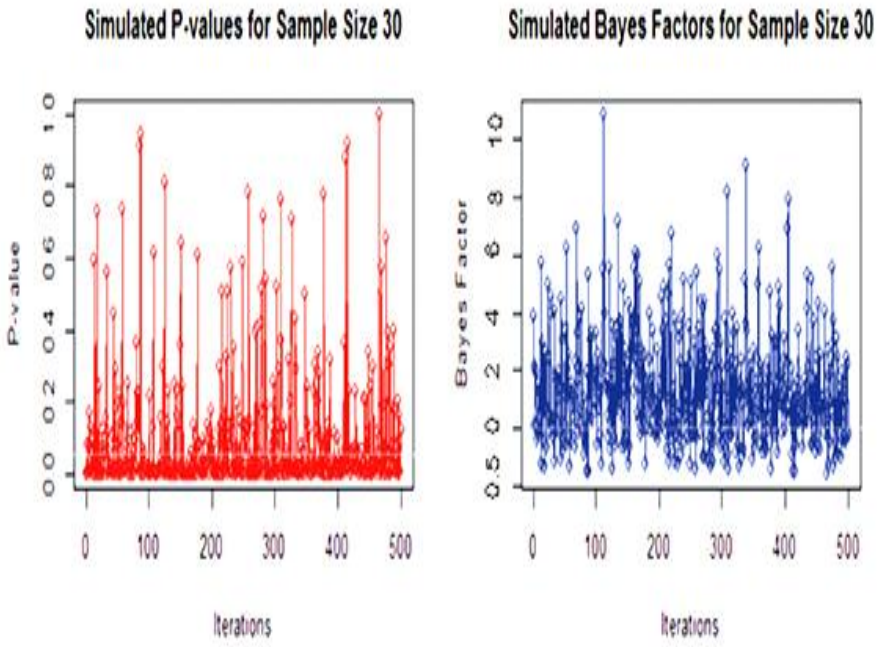

Figure 3: Plots of p-values and Bayes factors with sample size 30

hypothesis but conclude that there is no evidence against the null. This cannot tell if the null is true but the selected sample was unusual, or if the null is false. In contrast, the Bayes factor gave the probabilistic evidence for both the null hypothesis and the alternative given the data simulated with small sample sizes because it evaluates the ratio of likelihoods of both hypotheses using the data.

Secondly, we note that as the sample size increased, the p-value overwhelmingly gave significant results for the rejection if the null hypothesis. The Bayes factor also gave stronger evidence in favour of the alternative hypothesis as the

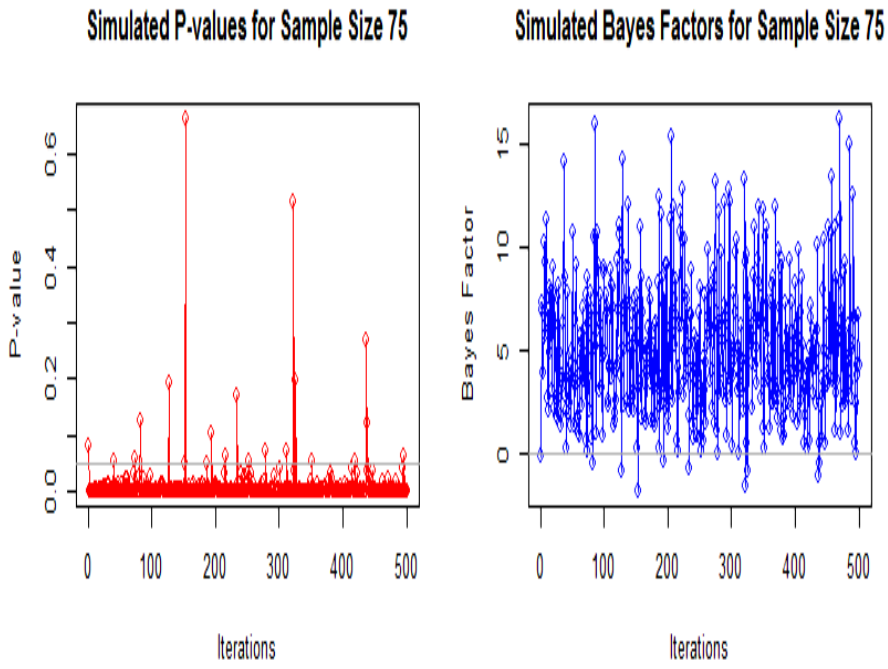

Figure 4: Plots of p-values and Bayes factors with sample size 75

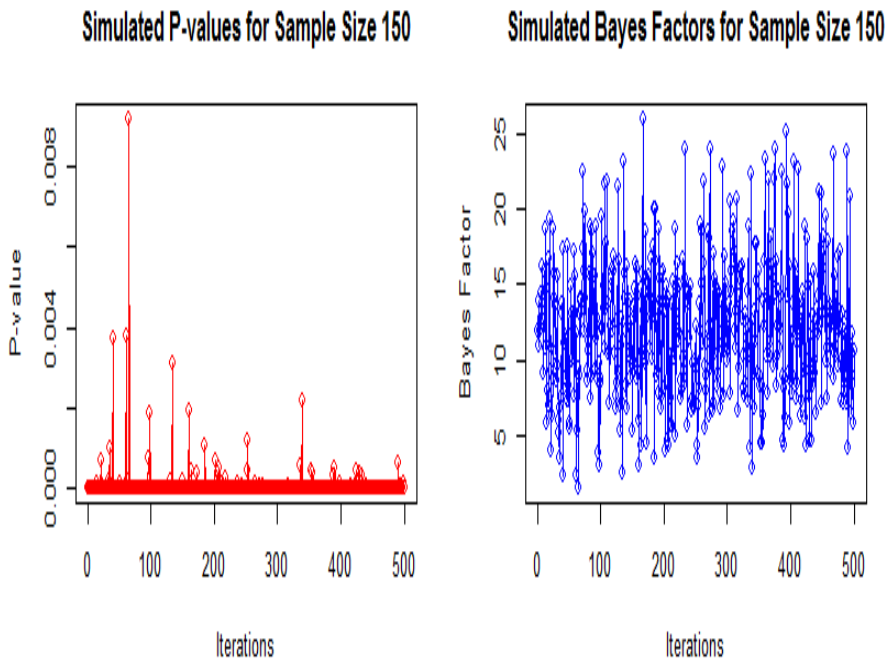

Figure 5: Plots of p-values and Bayes factors with sample size 150

sample size increased.

3.2 *Data simulation for p-values and Bayes factors with effect sizes*

The ensuing figures show the simulation of ANOVA data set for a one-way design with brief discussions explaining the resulting p-values and Bayes factor against possible effect sizes with different sample sizes. The simulation process follows the process taken in the previous section.

Table 2: Summary of simulation results of p-values and Bayes factors for sample sizes

	P-values		*Bayes factors*	
Sample Size	*Against* H_0	*Favour* H_0	*Favour* H_1	*Favour*H_0
9	28.4 %	71.6 %	36.4 %	63.3 %
15	34.2 %	65.8 %	47 %	53 %
30	64.4 %	35.6 %	72.2 %	27.8 %
75	96 %	14 %	97.6 %	2.4 %
150	100 %		100 %	

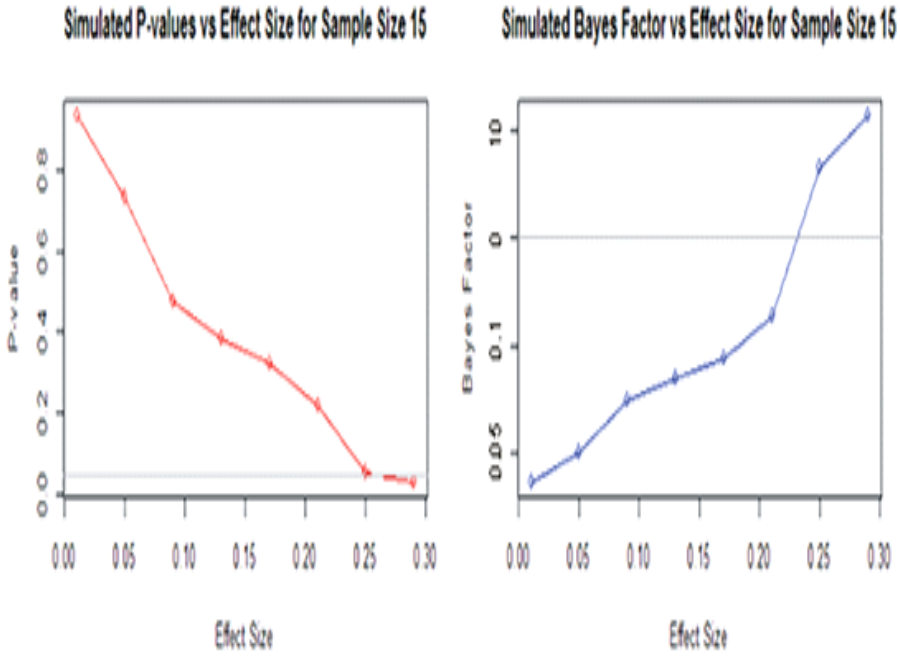

Figure 6: Plots of p-values and Bayes factors against effect size with sample size 15

3.2.1 Simulation results for effect sizes

We observed that the p-values were able to detect only large values of effect sizes when the sample size was small and when the sample size was large it could detect lower effect sizes. This showed that p-value is confounded or affected by both sample sizes and effect sizes. It was also observed that p-value almost never finds support for the null hypothesis that the effect size is zero; this affirms the bias of p-value against null hypotheses. The Bayes factor on the other hand was able to detect lower effect sizes even with small sample size and as sample size increased it still was able to detect small effect sizes as well as large effect sizes, though with larger values for larger effect sizes. With this, it can be said that the Bayes factor gave evidence for both hypotheses. This is expected because of the incorporation of effect size in the evaluation of Bayes factor. Thus only sample size does affect Bayes factor as evidence against the null in large sample size.

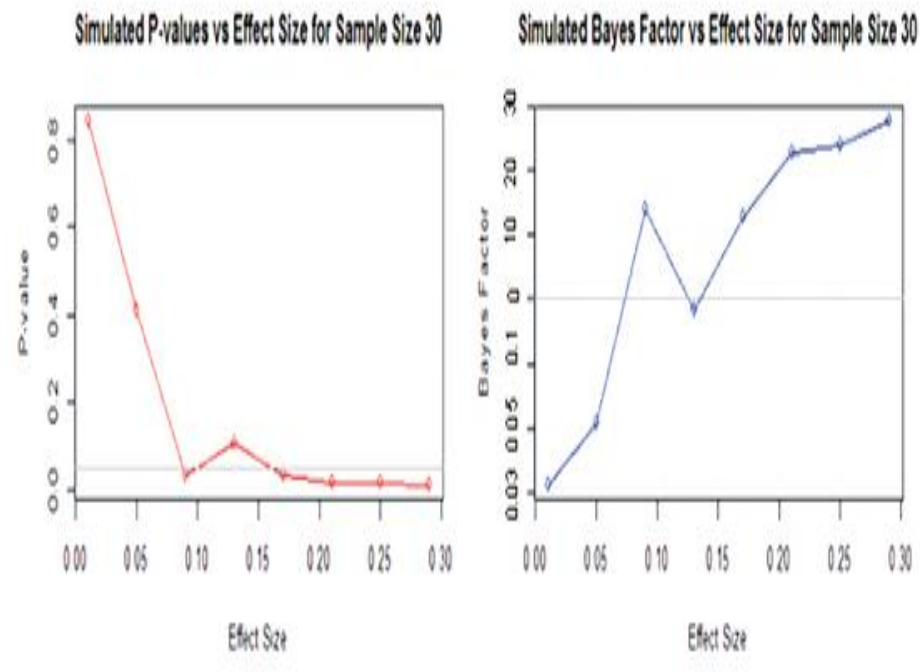

Figure 7: Plots of p-values and Bayes factors against effect size with sample size 30

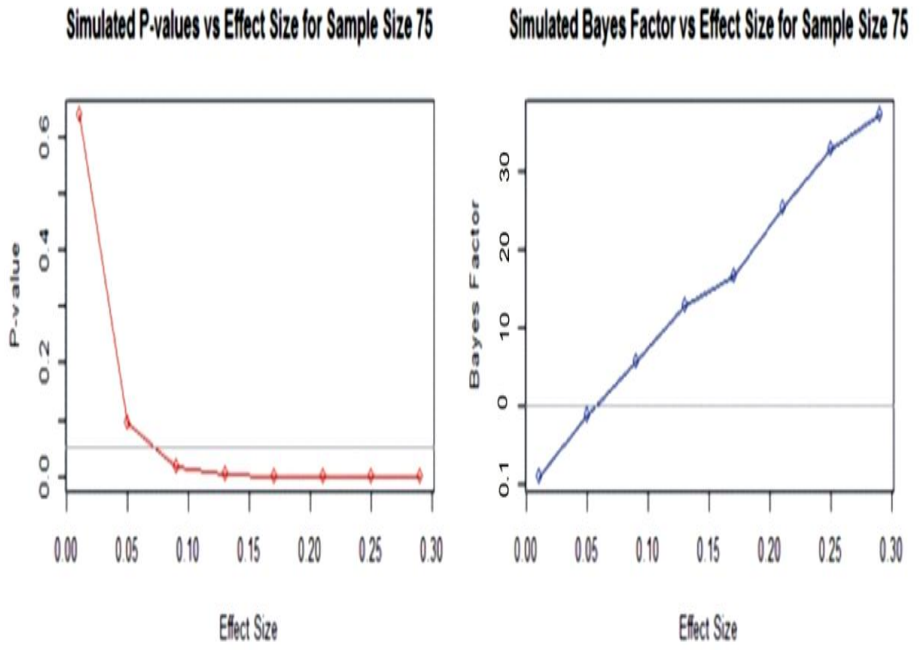

Figure 8: Plots of p-values and Bayes factors against effect size with sample size 75

3.3 *Bayes factor in one-way ANOVA and multiple comparisons*

The data set used is from a study of the fecundity of the fruit fly *Drosophila melanogaster* by Hand, et al. (1994). There are 25 female flies from each of three strains per diem fecundity (number of eggs laid per female per day for

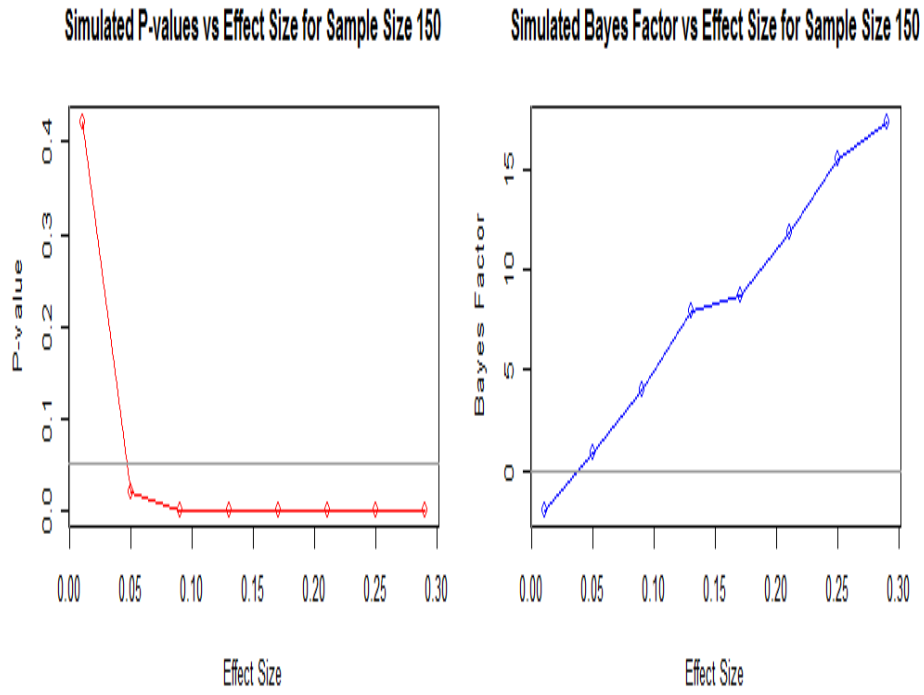

Figure 9: Plots of p-values and Bayes factors against effect size with sample size 150

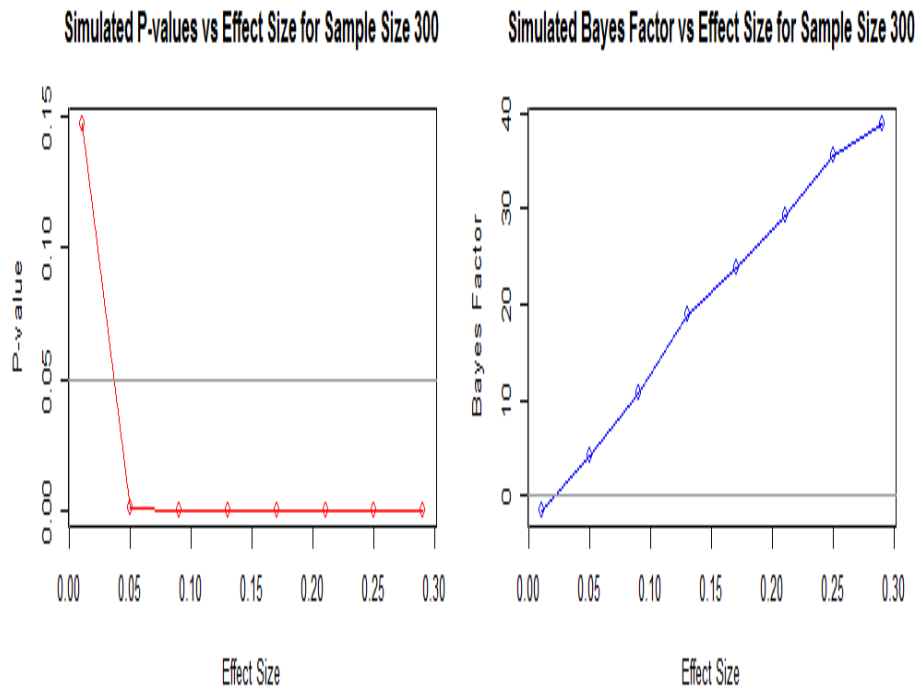

Figure 10: Plots of p-values and Bayes factors against effect size with sample size 300

the first 14 days of life) recorded. The genetic lines labeled RS and SS were selectively bred for resistance and for susceptibility to the pesticide DDT respectively, and the NS line is a non-selected control strain. In this experiment, the effect of the factor *genetic strain* on fecundity is of interest.

Looking at the boxplot and the means plot for the data in Figure 11, we see that the effect of each strain appear to be different. We want to examine the treatment means, or equivalently, the treatment effects to see which pair(s) do not differ significantly using Bayes factor.

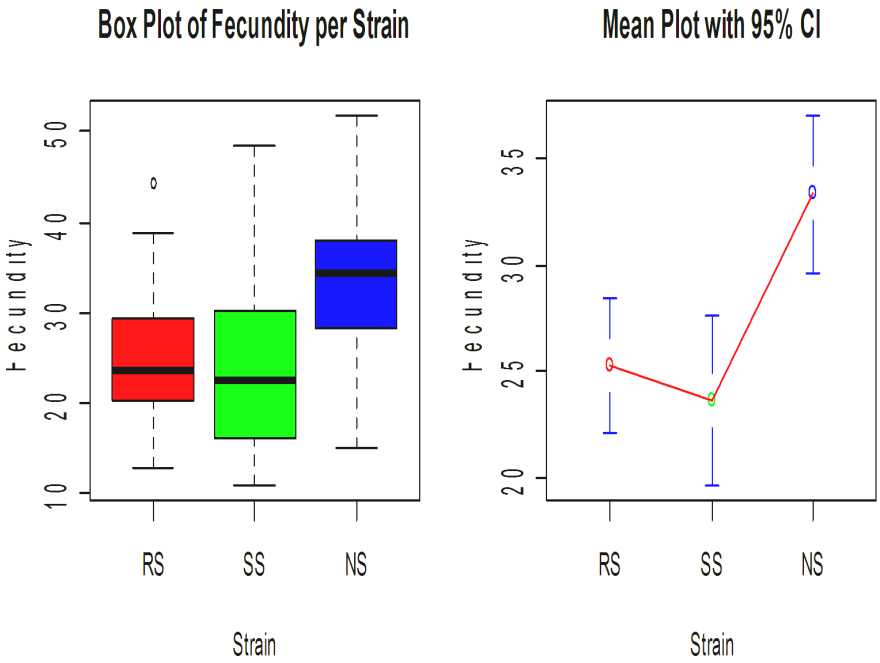

Figure 11: Density plot, box plot and means plot of one-way ANOVA data

The classical ANOVA on these data is given below in Table 3. The F value from the data is given as approximately 8.67 which is greater than $F_{table}(2, 72) \approx 3.14$ from the F distribution table at 5 % significant level and we can reject the null (p-value = 0.000424) that these data came from genetic strain populations where the average fecundity of the flies was the same.

Table 3: ANOVA table for fecundity data of fruit flies

Source	*DF*	*SS*	*MS*	*F*
Strain	2	1362.21	681.11	8.67
Error	72	5659.02	78.6	
Total	74	7021.23		

The Bayesian analysis described in Section 2.2 compares the adequacy of the effects or unequal means (alternative) model against the null model of equal means by evaluating the Bayes factor for this comparison.

The output of the Bayes factor obtained through the BayesFactor package by Morey and Rouder (2014) implemented in R software, gives a Bayes factor of 70.15-to-1 as the probabilistic ratio of evidence of support of the data for the effects model to the null model. From the Bayes factor interpretation in Table I, this is strong evidence for the alternative or effect model by the observed data and hence that genetic strain does affect fecundity (as the classical ANOVA showed).

3.3.1 Multiple comparisons with Bayes factor

The result implies we have support for the effects model$\mu(RS) \neq \mu(SS) \neq \mu(NS)$, but at this point we cannot state with any authority which specific pairs are the same, all we can say is that at least one is different. Given the evidence for the effects model that there is difference in means for each strain, with multiple comparison using Bayes factor, we wish to see which pair of means differ by considering the set of hypotheses models:

$$H^{(1,2)} : \mu(RS) = \mu(SS) \neq \mu(NS)$$

$$H^{(1,3)} : \mu(RS) = \mu(NS) \neq \mu(SS)$$

$$H^{(2,3)} : \mu(NS) = \mu(RS) \neq \mu(SS)$$

Since we already have the Bayes factor result in favour of the full model against the null model, we shall be comparing the other three equality constraints against the null model, the full model against each of the three models and lastly the three models paired against each other.

Table 4 shows the Bayes factor of the full model against each of the equality constraint model. Clearly from Table 4, the Full model is preferred with decisive

Table 4: Comparisons of the full model against each three equality constraints

Model	*Full model/RS=SS*	*Full model/RS=NS*	*Full model/SS=SS*
Bayes factor	0.307	22.213	126.533
1/Bayes factor	3.257	0.045	0.008

evidence over the $H^{(2,3)} : \mu(RS) \neq \mu(SS) = \mu(NS)$ unlike the other two constraints with less values for evidence. The implication is that the full model which says that the three group means significantly differs is preferred over $H^{(2,3)} : \mu(RS) \neq \mu(SS) = \mu(NS)$, and implies that SS and NS do significantly differ. While RS and NS also differ, their difference is not as large as that of SS and NS.

Table 5 shows the Bayes Factor Values for Comparison of RS=SS Model against other Two Equality Constraints.

Table 5: Comparison of RS=SS model against other two equality constraints

Model	*RS=SS/RS=NS*	*RS=SS/SS=NS*
Bayes factor	72.288	411.782
1/Bayes factor	0.014	0.002

In Table 5, we get the Bayes factors comparing the model $H^{(1,2)} : \mu(RS) = \mu(SS) \neq \mu(NS)$ against $H^{(1,3)} : \mu(RS) = \mu(NS) \neq \mu(SS)$ and $H^{(2,3)}$:

$\mu(NS) = \mu(RS) \neq \mu(SS)$, that is we are comparing each of the possible pairs of means. Clearly the model $H^{(1,2)} : \mu(RS) = \mu(SS) \neq \mu(NS)$ with Bayes factor of 411.782 has decisive evidence from the data against the other pairs of comparisons and this implies that means of RS and SS do not significantly differ from each other. According to Bayes factor value, SS and NS means differ followed by the means of RS and NS in order of magnitude of mean differences. This can be seen in the plot of means and this shows how Bayes factor incorporates the effect size in its computation.

3.3.2 Comparison of the Bayesian and the Classical Multiple comparisons Results

In summary, we combine the Bayes factors of each test in the set of hypotheses against null to see how each equality model of effects compares against the null model and hence which pair of means differs as supported by the data obtained, we thus obtain Table 6 below.

Table 6: Bayes factor values for comparisons of the full model against each three equality constraints

S/N	*Equality Constraint Model*	*Bayes Factor*	*1/Bayes Factor*
1	Strain(Full model)	70.149	0.014
2	RS=SS	228.288	0.004
3	RS=NS	3.158	0.317
4	SS=NS	0.554	1.805

Clearly from Table 6, the Bayes factor values for equality of means of RS and NS as well as those of SS and NS, is clearly by interpretation *Not worth more than a bare mention*. SS=NS with the least Bayes factor implies that the difference between their means is higher than that of RS=NS.

The outputs of the classical Tukey multiple comparison of means and Duncan's new multiple range test are shown in Table 7 and Table 8. The Tables show that the mean fecundity of strains RS and SS are not significantly different with p-values of 0.7934 and 0.5183 respectively. This result is in consonant with the Bayes factor results in Table 6 supporting the equality of means of RS and SS.

Table 7: Tukey multiple comparison of means

Strain	*Diff*	*Low. CI*	*Uppr. CI*	*p-value*
SS - RS	-1.628	-7.629	4.373	0.793
NS - RS	8.116	2.115	14.117	0.005
NS - SS	9.744	3.743	15.745	0.001

We can see that the model of equal means of strains RS and SS given by the equality constraint $H^{(1,2)} : \mu(RS) = \mu(SS) \neq \mu(NS)$ has the highest Bayes factor of 228.288 and this shows decisive evidence from the data that while the

Table 8: Duncan's new multiple range test

Strain	*Diff*	*Low. CI*	*Uppr. CI*	*p-value*
SS - RS	-1.628	-6.627	3.371	0.518
NS - RS	8.116	3.117	13.115	0.002
NS - SS	9.744	4.485	15.003	0.0003

different genetic strains affect fecundity of the fruit flies, the group means of RS and SS are not significantly different.

3.4 *MCMC posterior sampled simulation results*

We performed Bayesian analysis using Gibbs Sampler on the data of fecundity of the fruit fly *Drosophila melanogaster* by Hand et al., (1994), with our prior specifications on our parameters as stated in Section 2.1 and Section 2.2. The purpose is to check the veracity of the result by Baayes factor. The analysis was carried out with the software called Windows Bayesian analysis using Gibbs Sampling (WinBUGS) and the following output was obtained. Table 9 below gives the summary statistics of the posterior samples. The nodes beta[1], beta[2] and beta[3] are the mean parameters for the three group of strains, m is the overall mean parameter and s is the model standard deviation.

Table 9: MCMC posterior summaries after 102000 iterations and 1000 burn-in iterations

node	*mean*	*sd*	*MC error*	*2.5 %*	*median*	*97.5 %*	*start*	*sample*
beta[1]	-1.348	1.292	0.00727	-4.037	-1.273	0.936	1001	102000
beta[2]	-1.333	1.284	0.00966	-4.083	-1.207	0.759	1001	102000
beta[3]	2.680	1.639	0.0157	-0.068	2.681	5.906	1001	102000
m	27.420	1.077	0.00327	25.280	27.420	29.520	1001	102000
s	9.276	0.833	0.00427	7.812	9.218	11.080	1001	102000

The posterior distribution shows that the three means are different as we saw from the value of the Bayes factor reported. We also see that for the large number of iterations the posterior overall mean and variance are very close to the prior overall mean and the variance (prior mean and standard deviation 27.42, 9.74; posterior mean and standard deviation 27.42, 9.28), with their 95 % credible intervals.

The posterior box plot of Figure 12 shows that from the posterior distributions of the model the mean effects of the first two levels of the factor are near the same with the third level significantly different. This is consistent with the Bayes factor result and we see that the posterior mean effects for the last level of the factor strain is very different from the other two strains as we see in Table 6.

Lastly, we see from the density plots in Figure 13 that the simulated posterior distributions of the parameters are in consonant with our prior specifications with the noninformative priors placed on the mean and variance of the models.

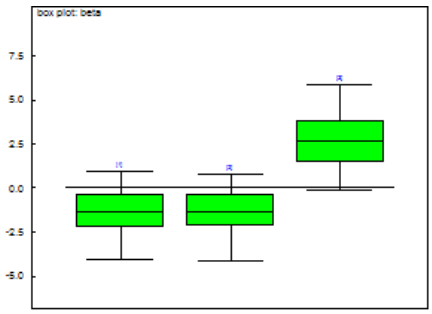

Figure 12: Posterior box plot for the three groups

The density of the third level parameter we see is the different one from the first two. The posterior density of the group level variance g also shows that mass falls off for very small and very large values of g; that is, g is constrained to be somewhat near 1.

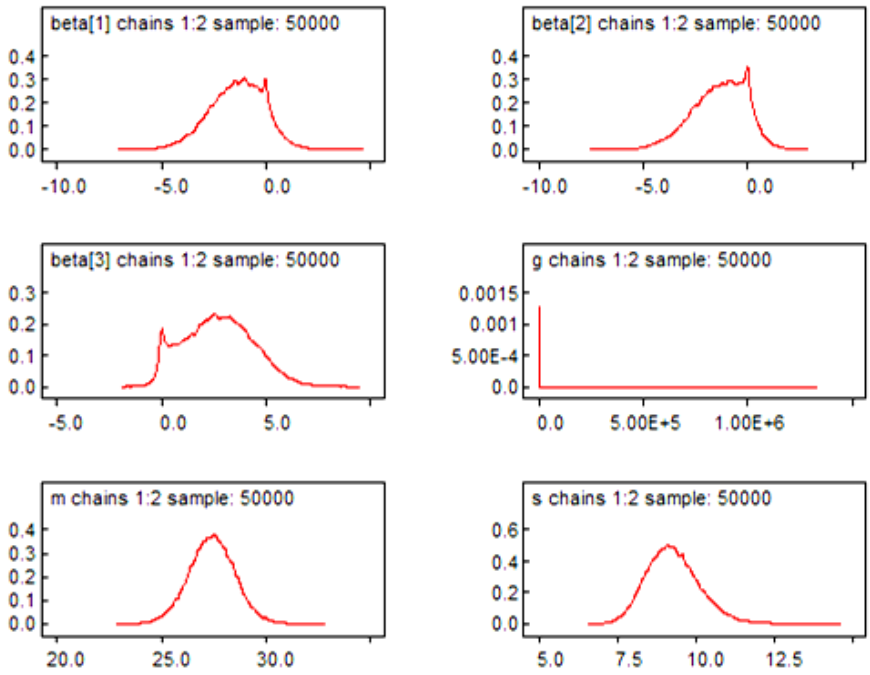

Figure 13: Posterior density plot of the parameters

From the foregoing we see that the posterior distribution has helped us to examine the veracity of the favoured full model, as reported by Bayes factor, since the posterior simulated data look like the actual data obtained.

4. Conclusion

One major goal in the analysis of experimental data is the assessment of models that specify constraints among factors. In meeting this goal, there is a general need to be able to state evidence for effects. The Bayes factor provides an approach to measuring evidence from data for competing theoretical positions. In this regard, it provides a principled approach to accumulating evidence for null

hypotheses; this is an advantage over the null hypothesis significance p-value testing.

We were able to see from our study of one-way balanced design, how p-value used in test of hypotheses is confounded by the influence of sample sizes and effect sizes through our simulated data. We saw that unlike p-value that is both influenced by sample size and effect size, Bayes factor was only affect by sample size, that is, as sample size increases, the Bayes factor favours the alternative hypothesis the more as expected in large samples. Bayes factor was able to show support for null hypothesis for reasonable smaller sample sizes, unlike p-value which always does not necessarily give support for the null, but leaves one with lack of sufficient evidence for the alternative.

We further saw how the posterior distribution gives a way to examine the veracity of one model over another by checking the result of Bayes factor with the samples from the posterior distributions to see if the posterior simulated data look like the actual data. We also saw how Bayes factor can be used to do multiple comparison tests by comparing each pair of means for equality using their Bayes factor. Data set from a one-way balanced study of the fecundity of the fruit fly Drosophila melanogaster by Hand, et al., (1994) was analysed to see how the three group means of the factor "genetic strain" affect fecundity of the flies. Having gotten an effect of strain on fecundity using Bayes factor, we saw how Bayes factor was used to test for pair-wise equality of means, which were represented by models, and each model was tested and evaluated to see which pair of equal means was supported from the data.

We therefore propose that Bayes factor be used in place of p-value in making decisions in one-way ANOVA. With the availability of powerful statistical computing programs like R, the Bayesian approach can be readily applied these days. And in cases of ANOVA where we want to compare more than two competing models like in two-way design and other designs, we now have a graded statistical measure for support of any competing model without bias to any particular null model (intercept-only model).

References

Andrade, P. C, Rocha, L. C., and Mendes da Silva, M. (2017). Bayesian Multiple Comparisons Procedures for CRD in R Code. International Journal of Probability and Statistics 2017, 6(3): 45-50

Andraszewicz, S., Scheibehenne, B., Rieskamp, J., Grasman, R., Verhagen, J. and Wagenmakers, E. J. (2015). An Introduction to Bayesian Hypothesis Testing for Management Research. Journal of Management, Vol. 41, No. 2, February 2015, 521-543. DOI: 10.1177/014 9206314560412. Retrieved from http://www.ejwagenmakers.com/2015/AndraszewiczEtAl2015.pdf

Bayarri, M. J. and Garcia-Donato, G. (2007). Extending Conventional Priors for Testing General Hypotheses in Linear Models. Biometrika, 94 , 135-152.

Burnham, K. P, and Anderson, D. R. (2002). Model Selection and Multimodel Inference: A Practical Information-Theoretical Approach. 2nd ed., Springer-Verlag.

Coe, R. (2002). It's the effect Size, Stupid. What effect Size is and Why is it Important. Paper presented at the Annual Conference of the British educational research Association, University of Exeter, England, 12-14 September 2002.

Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences (2nd. ed.). Lawrence Erlbaum Associates.

Efron, B. (2005). Bayesian, Frequentists and Scientists. Journal of the American Statistical Association, Vol. 100, No. 469, pp. 1-5.

Gelman, A. (2005). Analysis of Variance, Why it is more Important than Ever. Annals of Statistics, No 33, pg 1 - 53.

Guo, R. and Speckman, P. L. (2009). Bayes factor consistency in linear models. In the 2009 International Workshop on Objective Bayes Methodology, Philadelphia, June 5-9, 2009.

Hand, D. J., Daly, F., Lunn, A. D., et al. (1994) Small Data Sets. Boca Raton, FL: Chapman and Hall/CRC.

Hansen, M. H. and Yu, B. (2001). Model Selection and the Principle of Minimum Description Length. Journal of American Statistical Association. Vol. 96, No 454, pp 746-774. Retrieved from http://www.stat.ucla.edu/ cocteau/papers/pdf/old/main.pdf

Hossein, P. Bayesian Hypothesis Testing. Retrieved April 2017 from http://www.probabilitycourse.com/

Howell, D. C. (2012). Statistical Methods for Psychology. 8th Edition, Belmont, CA ISBN-13:978-1-111-83548-4. Retrieved from http://www.uvm.edu/ dhowell/methods8/Supplements/oldChapter12.pdf

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

Kass, R. E. and Raftery A. E. (1995). Bayes Factor. Journal of the American Statistical Association, Vol. 90, No 430, 773-795.

Koech, M. K., Otieno, A. R., Kimeli, V. and Koech, E. K. (2014). Posterior F-value in Bayesian Analysis of Variance using WinBugs. Mathematical Theory and Modeling Vol. 4, No. 5, 2014.

Kotz, S. and Nadarajah, S. (2004). Multivariate t distributions and their Applications. Cambridge: Cambridge University Press.

Lambers, J. (2009). MAT 460/560 Lecture Notes: Gaussian Quadrature. Retrieved from http://math.usm.edu/lambers/math460/fall09/lecture31.pdf

Lindley, D. V. (1957). A Statistical Paradox. Biometrika, Vol. 44, Issue 1-2, pg 187-192. Doi:10.1093/biomet/44.1-2.187.JSTOR 2333251.

Morey, R (2015). Multiple Comparisons with BayesFactor, Part 1. Retrieved from http://bayesfactor.blogspot.com/2015/01/multiple-comparisons-with-bayesfactor-1.html.

Morey, R. D. and Rouder, J. N. (2014). BayesFactor 0.9.6 Comprehensive R Archive Network. Retrieved from http://cran.r-project.org/web/ packages/BayesFactor/index.html

Morey, R. D., Romeijn, J. W. and Rouder, J. N. (2013). The Humble Bayesian Model Checking from a Full Bayesian Perspective. British Journal of Mathematical and Statistical Psychology, 66, pp68-75. Retrieved from http://dx.doi.org/10.1111/j.2044-8317.2012.02067.x

Ntzoufras, I. (2009). Bayesian Modelling Using WinBUGS. An Introduction. Wileyinterscience. John Wiley and Sons Inc. Publication.

Oehlert, G. W. (2010). A First Course in Design and Analysis of Experiments. ISBN, 0-7167.3510-5.

Orloff, J. and Bloom, J. (2014). Comparison of Frequentist and Bayesian Inference. Retrieved from http://ocw.mit.edu/courses/readings on 23/05/2017.

Pratte, M. S. and Rouder, J. N. (2011). Hierarchical Single-and Dual Process Models of Recognition Memory. Journal of Mathematical Psychology, 55, pp 36-46.

Raftery, A. E. (1995). Bayesian Model Selection in Social Research. Sociological

Methods, Vol 25, (1995), pp 111-196, Cambridge, MA: Blackwell.
Raftery, A. E. (1999). Bayes factor and BIC: Comments on a critique of the Bayesian information criterion for model selection. Sociological methods and Research, 27, pp 411-427.
Ronald, L. W. and Nicole A. L. (2016). The ASA's Statement on p-Values: Context, Process and Purpose. The American Statistician, 70:2, 129-133, DOI:10.1080/00031305.2016.1154108. Retrieved from http://amstat.tandfonline.com/doi/full/10.1080/00031305.2016.1154108
Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., and Iverson, G. (2009). Bayesian t tests for Accepting and Rejecting the Null Hypothesis. Psychonomics Bulletin and Review, 2009, 16 (2), pp 225-237.
Rouder, J. N., Morey, R. D., Speckman, P. L., and Province, J. M. (2012). Default Bayes Factors for ANOVA Designs. Journal of Mathematical Psychology, 56, pp 356-374.
Rouder, J. N., Morey, R. D., Verhagen J., Swagman, A. R. and Wagenmakers, E. J. (2016). Bayesian Analysis of Factorial Designs. Retrieved from http://ejwagenmakers.com/inpress/RouderEtAlinpressANOVAPM.pdf
Sellke, T., Bayarri, M. J. and Berger, J. O. (2001). Calibration of P Values for Testing Precise Null Hypotheses. American Statistician, 55 , 62-71.
Shiffrin, R. M., Lee, M. D., Kim, W., and Wagenmakers E. J. (2008). A Survey of Model Evaluation Approaches with a Tutorial on Hierarchical Bayesian Methods. Retrieved from http://www.socsci.uci.edu/ mdlee/shiffrinEtAl2008.pdf
Siegfried, T. (2016). Expert Issue Warning on Problem with P values: Misunderstandings about Common Statistical Test Damage Science and Society. Retrieved from http://www.sciencenews.org/blog/context/experts-issue-warning-problems-p-values.
Wagenmakers, E. J., and Farrell, S. (2004). Notes and Comments: AIC Model Selection Using Akaike Weights. Psychonomics Bulletin and Review, 11 (1), 192-196. Retrieved from http://www.ejwagenmakers.com/2004/aic.pdf
Wagenmakers, E. J. (2007). A Practical Solution to the Pervasive Problems of P–value. Psychonomics Bulletin and Review 2007, 14(5), pp 779-804
Weakliem, D. L. (1999). A Critique of the Bayesian Information Criterion for Model Selection. Sociological methods and Research. 27: pp 59-97. Retrieved from http://www.stat.washington.edu/pdf
Zellner, A. and Siow, A. (1980). Posterior Odds ratios for Selected Regression Hypotheses. In J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith (Eds.), Bayesian Statistics: Proceedings of the First International Meeting held in Valencia (Spain) (pp. 585603). University of Valencia.