https://doi.org/10.26219/heal.aueb.6544

ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

ΣΧΟΛΗ ΕΠΙΣΤΗΜΩΝ & ΤΕΧΝΟΛΟΓΙΑΣ ΤΗΣ ΠΛΗΡΟΦΟΡΙΑΣ
SCHOOL OF INFORMATION SCIENCES & TECHNOLOGY

ΜΕΤΑΠΤΥΧΙΑΚΟ ΣΤΑΤΙΣΤΙΚΗ
MSc IN STATISTICS

Bayesian ANOVA and Multiplicity Adjustment

By

Alkisti Mamougka

A THESIS
Submitted to the Department of Statistics
of the Athens University of Economics and Business
in partial fulfilment of the requirements for
the degree of Master of Science in Statistics

Athens, Greece
February 2025

https://doi.org/10.26219/heal.aueb.6544

https://doi.org/10.26219/heal.aueb.6544

Μπεϋζιανή Ανάλυση Διακύμανσης και Διόρθωση Πολυπλοκοτητας

Άλκηστη Μαμούγκα

ΔΙΑΤΡΙΒΗ
Που υποβλήθηκε στο Τμήμα Στατιστικής
του Οικονομικού Πανεπιστημίου Αθηνών
ως μέρος των απαιτήσεων για την απόκτηση
Διπλώματος Μεταπτυχιακών Σπουδών στη Στατιστική

Αθήνα
Φεβρουάριος 2025

https://doi.org/10.26219/heal.aueb.6544

DEDICATION

To Socrates and Athena, my proudest achievement.

https://doi.org/10.26219/heal.aueb.6544

https://doi.org/10.26219/heal.aueb.6544

ACKNOWLEDGEMENTS

This thesis represents the end of a journey that has fostered growth and support from many great people. I am deeply grateful to everyone who has played a role in making this work possible.

I would like to give my special and sincere thanks to my advisor, Prof. Ioannis Ntzoufras, for the guidance, feedback and encouragement throughout this thesis. I wish to also express my appreciation to all members of Department of Statistics, Athens University of Economics and Business, for their inspiration and support.

I would also like to give an extra special thank you to my parents, who have been the foundation of my love for learning. They are the ones who taught me the importance of education, the habit of questioning and persistence since I was young, gifts that I will forever cherish. To my sister, Anna, my role model and life's inspiration, thank you for showing me what commitment and determination truly mean.

To my children, Socrates and Athena. Every step along this path, every late hour and every challenge was made worthwhile by the promise of you. Your love, questions and laughter remind me of what matters most. To my husband, Ilias, your patience, faith in me and love have made this journey possible, and for that, I will be forever grateful.

And lastly, to my friends and the rest of my family and to all those who supported me in ways both big and small, thank you. This achievement is not mine, but a reflection of the individuals who have walked this path with me.

https://doi.org/10.26219/heal.aueb.6544

https://doi.org/10.26219/heal.aueb.6544

ABSTRACT

Alkisti Mamougka

Bayesian ANOVA and Multiplicity adjustment

February 2025

This thesis explores Bayesian ANOVA as an alternative to frequentist methods. It focuses on its advantages in addressing multiple comparisons and incorporating prior knowledge. Traditional ANOVA methods rely on p-values and require arbitrary multiplicity adjustments, such as Bonferroni corrections, to control the risk of false discoveries. In contrast, Bayesian ANOVA naturally accounts for multiplicity through prior-model probabilities, offering a more coherent framework for inference.

A key component of this study is a simulation-based analysis where we implement Bayesian ANOVA under different prior specifications. The prior choices was examined and how they affect model selection, inference accuracy, and how they handle multiple comparisons. Additionally, we apply Bayesian ANOVA to real-world data, investigating factors influencing marathon performance.

The results demonstrate that Bayesian ANOVA provides a more flexible and informative approach to variance analysis while inherently adjusting for multiplicity. By improving model selection and reducing the risk of false discoveries without arbitrary corrections, Bayesian methods present a strong alternative to frequentist approaches. This study contributes to the broader adoption of Bayesian statistical techniques in experimental research and applied sciences.

https://doi.org/10.26219/heal.aueb.6544

https://doi.org/10.26219/heal.aueb.6544

ΠΕΡΙΛΗΨΗ

Άλκηστη Μαμουγκα

ΜΠΕΫΖΙΑΝΗ ΑΝΑΛΥΣΗ ΔΙΑΚΥΜΑΝΣΗΣ ΚΑΙ ΔΙΟΡΘΩΣΗ ΠΟΛΥΠΛΟΚΟΤΗΤΑΣ

Φεβρουάριος 2025

Η παρούσα διατριβή εξετάζει τη Bayesian ANOVA ως εναλλακτική προσέγγιση έναντι των κλασικών μεθόδων, εστιάζοντας στα πλεονεκτήματά της στη διαχείριση πολλαπλών συγκρίσεων και την αξιοποίηση της α πριόρι πληροφορίας. Οι παραδοσιακές μέθοδοι ANOVA βασίζονται στα p-value και απαιτούν αυθαίρετες διορθώσεις πολυπλοκότητας, όπως η Bonferroni, προκειμένου να ελεγχθεί ο κίνδυνος ψευδών ανακαλύψεων. Αντίθετα, η Μπεϋζιανή ANOVA ενσωματώνει φυσικά τη διόρθωση πολυπλοκότητας μέσω των εκ των προτέρων κατανομών.

Το κυριότερο κομμάτι της μελέτης βασίζεται σε μια προσομοίωση, στην οποία εφαρμόζουμε Bayesian ANOVA με διαφορετικές επιλογές α πριόρι κατανομών. Εξετάζουμε την επίδραση των κατανομών αυτών στην επιλογή μοντέλου, την ακρίβεια των συμπερασμάτων και τη διαχείριση των πολλαπλών συγκρίσεων. Επιπλέον, διερευνούμε την εφαρμογή της Bayesian ANOVA σε πραγματικά δεδομένα, εστιάζοντας σε παράγοντες που επηρεάζουν την απόδοση των αθλητών στον μαραθώνιο.

Τα αποτελέσματα αναδεικνύουν τη Bayesian ANOVA ως μια πιο ευέλικτη και πληροφοριακά πλουσιότερη προσέγγιση στην ανάλυση διακύμανσης, η οποία αντιμετωπίζει τη διόρθωση πολυπλοκότητας εγγενώς, χωρίς την ανάγκη αυθαίρετων προσαρμογών. Με τη βελτίωση της επιλογής μοντέλου και τον περιορισμό του κινδύνου ψευδών ανακαλύψεων, η Bayesian προσέγγιση καθίσταται μια ισχυρή εναλλακτική των κλασικών μεθόδων, ενισχύοντας την υιοθέτηση Μπεϋζιανών τεχνικών στη στατιστική ανάλυση και την εφαρμοσμένη έρευνα.

https://doi.org/10.26219/heal.aueb.6544

https://doi.org/10.26219/heal.aueb.6544

TABLE OF CONTENTS

Πίνακας περιεχομένων

https://doi.org/10.26219/heal.aueb.6544

https://doi.org/10.26219/heal.aueb.6544

https://doi.org/10.26219/heal.aueb.6544

https://doi.org/10.26219/heal.aueb.6544

CHAPTER 1

INTRODUCTION TO ANOVA AND BAYESIAN STATISTICS

1.1 Introduction

Frequentist and Bayesian methods represent two different paradigms in statistical inference, rooted in opposing philosophies. In the context of clinical trials, researchers intend to determine the probability of success for a new treatment (treatment effect). Frequentist methods deal with each affected person's response to the treatment as an unbiased experiment. By enrolling a large number of patients and conducting the experiment repeatedly, the average success rate from these multiple trials is used to estimate the true probability of success for the new treatment.

In contrast, Bayesian methods consider the probability of success a random variable and assign it a distribution when all patients are still unassigned to a treatment. These prior distributions are what we would expect the effect of a new treatment to look like before any data was collected and typically express information derived from previous knowledge about the problem. Bayesian updating occurs as clinical trial data is received and probability distributions for the treatment effect is updated based on actual evidence. This evidence, also called the data likelihood, is a measure of what we can learn from our data about how effective or ineffective the treatment really was. That updated probability distribution is the posterior distribution.

Bayesian ANOVA expands these concepts of Bayesian inference to the analysis of variance (ANOVA), where, for instance, a clinical trial will be testing multiple treatment groups. The aim of traditional ANOVA is to test if there are statistically significant group differences in the mean. However, Bayesian ANOVA allows for a more flexible and informative analysis by incorporating prior knowledge and providing a probabilistic interpretation of the results.

 https://doi.org/10.26219/heal.aueb.6544

1.2 Review of Classical ANOVA

Several studies are focused on the framework of analysis of variance (ANOVA), which is used for comparison between more than two groups. While the means of each group alone may not be important, a comparison of the means across the groups can provide important insights into the corresponding data. Because of the importance of these comparisons, tests on means are widely used in fields such as medicine, engineering, and educational research. Another reason for its extensive use is the simplicity of the application of the ANOVA test. However, like all parametric tests, the ANOVA test must satisfy certain assumptions to ensure the validity of the results. These are that the data are normally distributed within each group, that the within-group variances are equal and that observations are independent.

1.2.1 Challenges in Frequentist Inference

Statistical significance testing is widely used to determine whether the null hypothesis of no effect is to be rejected based on observed data. It is most often determined by comparing the p-value with a pre-chosen level of significance. The method has been severely criticized in the past decades. According to Cohen (1994) and Royal (1997), the null hypothesis is exacting an assumption that barely exists in practice. On the other hand, Wainer (1999) shows examples of when the population may be adequately represented by a point null hypothesis, and Jones and Tukey (2000) suggest even a more liberal stance towards significance testing. In general, the consensus is to use the null hypothesis most appropriately with restraint to represent fairly the population in question, not wholesale.

Several researchers such as Berger and Delampady (1987), Raftery (1995), Harlow, Mulaik, and Steiger (1997/2016), Wagenmakers (2007), and Masson (2011) have pointed out various deficiencies in the statistical significance testing. The above-mentioned scientists are further concerned by issues related to publication bias and questionable research practices. Publication bias describes when studies that return significant findings ($p < 0.05$) are more likely to be published than studies that fail to reject the null hypothesis ($p > 0.05$). Questionable research practices manipulate the analyses to produce significant p-values by selective removal of outliers, reporting

https://doi.org/10.26219/heal.aueb.6544

only significant results, post-hoc selection of covariates, or acquiring more data after having examined results close to the threshold.

The increasing criticism of statistical significance testing has generated interest in alternatives. Wasserstein and Lazar (2016) review proper uses of p-values, while Cumming (2012) suggests the use of Confidence Intervals as summaries of data concerning the parameter of interest. For another approach, Bayesian equivalent credible intervals are available, the comparison of which to CIs is given by Morey et al. (2016). Other propositions are: Trafimov and Marks 2015, that descriptive statistics are more appropriate than the usage of p-values and confidence intervals, and Benjamin et al. 2017 propose changing the significance criteria from 0.05 to 0.005 to reduce publication bias and keep questionable practices in check. Another important approach is the reevaluation of the Fisherian view on p-values expressed by Hurlbert and Lombardi (2009), that considers the p-value as evidence against the null hypothesis without depending on any previously set significance level.

1.2.2 Bayesian Hypothesis Testing and Bayes Factors

This thesis introduces a new alternative to routine significance testing: hypothesis testing via the Bayesian method and the Bayes factor. Initially developed by Jeffreys (1961), it was subsequently developed by Kass and Raftery (1995) and Klugkist, Laudy, and Hoijtink (2005), and again implemented in software by Rouder et al. (2009). As in this chapter, the Bayes factor has a variety of advantages when testing hypotheses.

Bayes factor allows testing of either an alternative or a null hypothesis. The null and the alternative hypothesis typically take the form:

$$H_0 : \textit{The effect is zero}$$

$$H_1 : \textit{The effect is not zero}.$$

This thesis will focus on hypothesis evaluation using the ANOVA model. For instance, for three groups these hypotheses might be written:

$$H_0 : \mu_1 = \mu_2 = \mu_3$$

$$H_1 : \textit{Not } H_0$$

In the frequentist approach F-test and t-test p-values are applied as signs of whether a hypothesis is verified or rejected. These methods were popular among scientists at

https://doi.org/10.26219/heal.aueb.6544

an early stage because they have predetermined formulas and algorithms. As the sample sizes increase, frequentist significance tests will increasingly reject the null hypothesis when it is in fact false. Yet when the null hypothesis is true, p-values are still uniformly distributed between 0 and 1 for all sample sizes. Thus evidence supporting the null cannot be aggregated, no matter the size of the data set, demonstrating that significance tests are set up by nature to reject the null hypothesis but not affirm it.

Therefore scientists use Bayes factor as a more scientifically appealing method. The Bayes factor, which was first proposed by Harold Jeffreys, is a Bayesian substitute for frequentist hypothesis testing. It is most frequently employed when comparing various models through hypothesis testing in order to ascertain which model most closely matches the data. The Bayes Factor BF_{01} measures how much more likely the observed data are under H_0 compared to H_1. This value represents the relative support provided by the data for H_0 versus H_1. A BF_{01} of 1 indicates no preference between H_0 and H_1. When BF_{01} is greater than 1, the data favor H_0. Conversely, if BF_{01} falls between 0 and 1, the data provide greater support for H_1. The Bayes factor has a natural and straightforward interpretation based on reasonable assumptions and has better properties than other methods of inference.

One of the key benefits is that it transcends the hard to accept-or-reject framework of traditional hypothesis testing. Instead, it quantifies the evidence for every hypothesis, including the null hypothesis when relevant. Furthermore, the Bayes factor naturally accommodates multiple hypothesis tests, reducing issues with multiple comparisons and providing a more nuanced method of statistical inference. Third, it allows for Bayesian updating: the evidence for hypotheses is updated as data accumulate. Under the Bayesian framework, one may gather new data and reassess a hypothesis if the first evidence is weak.

Finally, the Bayes factor does not depend upon Type I and Type II error rates controlled by assumed repeated sampling from null and alternative populations. Instead, it emphasizes Bayesian error probabilities, which represent the likelihood of drawing incorrect conclusions based on the data. Unlike traditional error rates, Bayesian error probabilities are directly related to the data at hand and do not rely on repeated sampling. These features make the Bayes factor a flexible and robust tool for the evaluation of hypotheses in experimental research.

https://doi.org/10.26219/heal.aueb.6544

Although Bayes' factor has a number of criticisms, the most prominent one is that it explicitly does not control Type I and Type II error rates as it focuses on Bayesian error probabilities. However, it has also been shown that a Bayesian t-test can be designed to minimize the combined average of Type I and Type II error probabilities. Gu, Hoijtink, and Mulder (2016) have provided such a design. Furthermore, Schönbrodt et al. (2017) have demonstrated that in the case of Bayesian t-tests with updating, the resultant Type I and Type II error rates are either equal to or less than those occurring under traditional significance testing but with reduced sample sizes. Thus, although controlling these errors is not the primary goal of Bayes factors, this does not mean that they are disregarded or unmanageable.

Another objection, answered by Sellke, Bayarri, and Berger (2001) and Mulder (2014), is that, when the simple null hypotheses are being tested (for instance, a mean of zero), the Bayes factor tends to coincide with or change the p-value as an index of evidence against the null. But this does not mean that the benefits of the Bayes factor automatically apply to p-values, nor that this correspondence is general across all hypotheses.

A third comment is that a prior distribution needs to be specified to compute a Bayes factor, and the variance of this prior is necessarily a subjective choice. This is potentially a concern for a researcher who wishes to take more of an objective approach. Of course, for hypotheses containing any equality constraints (a null hypothesis for instance), sensitivity analysis can be carried out with respect to prior variance. Moreover, in the case of informative hypotheses defined only by inequality constraints, the prior variance does not affect the resulting Bayes factors (Hoijtink, 2012). These considerations show that while prior specification is a necessary step, its influence can be put under effective control or minimized depending on the hypothesis being tested.

1.2.3 Bayesian Testing: A Comparison of Software Packages

There are many software packages, all of which quantify the evidence in data for null and alternative hypotheses. This thesis discusses only some of those that are implemented in publicly available software and are therefore accessible to the widest possible range of scientific applications. BayesFactor and bayesanova in R, JAGS,

https://doi.org/10.26219/heal.aueb.6544

and JASP, each of these tools offers different functionalities that enable researchers to efficiently perform Bayesian ANOVA and associated statistical analyzes.

The BayesFactor function from the R package, first described in Rouder et al., 2009, follows the framework established by Jeffreys (1961). This approach uses Jeffreys-Zellner-Siow (JZS) priors, or g-priors, as described by Liang et al. (2008). These priors contain default values for the prior distribution's variance that researchers can change to conduct sensitivity analyses. The BayesFactor package allows users to test the null and alternative hypotheses for ANOVA models, regression models, and contingency tables. This package, along with further examples and other resources, may be found at its CRAN website: (https://richarddmorey.github.io/BayesFactor/)

The bayesanova package offers another way to Bayesian analysis for ANOVA models, offering more flexible prior distribution specification and a clear way to allow researchers to undertake Bayesian model comparisons. In contrast to BayesFactor, bayesanova also supports a number of other types of priors than the JZS priors, thereby enabling more flexible options in Bayesian analyses. It has a wide scope, particularly related to hierarchical modelling and complicated experimental designs.

JAGS is a general-purpose tool for Bayesian analysis, offering flexible syntax to specify complex hierarchical models. This package differs from BayesFactor and bayesanova in that it does not offer default functions for Bayesian ANOVA. Users can specify their own models with MCMC. The flexibility of this makes it a very powerful tool for researchers when they have to work with atypical data structures and models.

Lastly, JASP is a graphical statistical software that encompasses Bayesian hypothesis testing into an intuitive interface. Contrary to the R packages that require knowledge of programming, this is a point-and-click interface, and that is why it will make this tool more approachable for researchers without much experience in coding. Support for Bayesian ANOVA, t-test, regression, contingency table analysis, among many others. JASP uses BayesFactor under the hood for some of its Bayesian analyses, so it is a useful alternative to R scripting for those who prefer to work with a GUI.

The advantages of these tools are related to different research questions and user experiences. BayesFactor allows for easy Bayesian ANOVA and regression with default priors, whereas bayesanova allows flexible specification of priors. JAGS

https://doi.org/10.26219/heal.aueb.6544

allows for the specification of complex Bayesian models, whereas JASP offers a point-and-click interface for Bayesian hypothesis testing, without needing any code. With such diverse tools, these software give a wide avenue to carry out robust Bayesian analysis for any particular research.

1.2.4 Literature Review of Bayesian ANOVA

Bayesian inference has subdivisions based on various probability interpretations. The two main methods that are usually applied in this regard include the objective Bayesian inference introduced by Berger (2006) and the subjective Bayesian inference introduced by Anscombe and Aumann in 1963. The former is associated with the works of Jeffreys (1961), whereas the latter is associated with Brooks (2003), Gilks (1994), and De Finetti (1937). Notably, De Finetti's work in 1937 marked the significant turn of events that led to the revival of Bayesian inference.

In many contexts of social sciences, any data analyzed is multilevel in nature. Consequently, statistical models must have the flexibility to capture between-group variation in order to make good representations of such data. This is typically obtained by allowing some or all model parameters to vary across groups. Such situations arise, for example, in survey data obtained in various locations (such as countries), experimental studies with data obtained in multiple regions, or educational studies with hierarchically grouped data of several levels students within classes, classes within schools and so on.

In such analyses, attention is placed on parameters that change across different group levels. Depending on the discipline, context, and estimation approach, such parameters are also called contextual effects, fixed effects, random effects, or stochastic/varying coefficients in various contexts. It is essential to notice these differences within groups because they determine a lot about the answer to simple questions in empirical social science.

Moreover, group-by-group analysis is also a very useful exploratory step to understand the heterogeneity of parameters. Despite its usefulness, it is often neglected in favor of more complicated techniques such as panel data estimators or Bayesian hierarchical models (Berger, 2006). One of the leading proponents of Bayesian modelling in the social sciences, Andrew Gelman, calls group-by-group analysis a "secret weapon". He calls this a "weapon" due to the powerful insights that

https://doi.org/10.26219/heal.aueb.6544

it can give, while it is "secret" because analysts commonly overlook this step in their haste to apply some advanced modelling technique.

As emphasized by Gelman (2007), the main idea is that the splitting of a large data set into group specific subsets often gives a better fit for the data compared to a pooled or combined analysis. By acknowledging and leveraging group specific variation, researchers can uncover meaningful patterns that would otherwise be obscured in aggregated models.

Bayesian methods have gained much popularity across almost all scientific fields, Poirier (2006). Their appeal mostly depends on the capability to deal with complex problems, like hierarchical modelling or non-nested model comparisons, with relative ease. The advantages of the Bayesian method have made this approach an essential part of data analysis in biology, engineering, economics, sociology, and psychology.

Rouder et al. (2012) present a broad framework for default Bayes factors in ANOVA designs and describe a method that reduces some of the most central weaknesses of traditional p-values as used in hypothesis testing. The authors discuss asymmetry in p-values whereby it is possible to contradict the null hypothesis but never to confirm it is a frequent consequence of this being so-called paradoxical results, such as Lindley's paradox. By contrast, Bayes factors provide symmetric and graded evidence, which enables researchers to determine the support for or against the null hypothesis in a more straightforward manner. A central innovation of this work is that default Cauchy priors on standardized effects make the approach invariant to unit transformations and more robust for a wide range of experimental designs. These are subsequently generalized to multivariate counterparts, enabling efficient computation of Bayes factors. This flexibility extends to a broad class of fixed, random, and mixed-effect models within and between-subject designs, allowing for wide applicability to different experimental situations. Perhaps most importantly, the utility of the approach in applied research has been made much easier with the development of the BayesFactor R package that simplifies Bayesian computation for researchers making the transition from frequentist approaches to Bayesian inference.

By extending the basics, Rouder et al. (2017) develop further Bayesian ANOVA methods to factorial designs common in experimental psychology. The extension frames ANOVA in terms of a hierarchical model, which may flexibly account for variations of the effect sizes and correlations between different factor levels. It allows a subtle approach, favored by Bayes factors over main and interaction effects in an

https://doi.org/10.26219/heal.aueb.6544

invariance test from rejecting or accepting a hypothesis, for which the traditional setup does not allow. This hierarchical modelling approach allows for adaptive shrinkage, reducing overfitting and enhancing robustness in parameter estimation in cases when studies have complex factor structures and heterogeneous effect sizes. The authors practically introduce how such methods can be done by the BayesFactor package and software JASP. These examples demonstrate how Bayesian methods can easily deal with complex experimental designs and make them, therefore, particularly useful for psychological research where individual differences and experimental manipulations frequently create heterogeneous effect sizes.

1.3 Introduction to Bayesian Statistics and Bayesian ANOVA

1.3.1 Bayesian principles

The real application of Bayesian statistics began in the early 21st century. By the late 1980s, Bayesian statistics was seen as an alternative to “classical” theory. Bayesian theory is based on probability theory, because all the unknown quantities in the model are treated as random variables.

Many statisticians have criticized Bayesian theory because of its apparent subjectivity that requires the specification of a prior distribution in order to make inference. However, the main obstacle to the widespread use of Bayesian methods for data analysis was the computational difficulty in estimating the posterior distribution except in the simple cases. Asymptotic methods solved specific problems, but these could not be generalized.

This began to change in the early 1990s with the rediscovery of Markov Chain Monte Carlo (MCMC) methods by two groups of Bayesians (Gelfand et al., 1992). Physicists have used MCMC methods since the 1950s. Nick Metropolis and his colleagues, who developed one of the first supercomputers of the era, tested their theories in physics using Monte Carlo techniques. The combination of MCMC methods and the speed of personal computers rapidly popularized this computational tool. Consequently, Bayesian statistics became prominent, opening new avenues for statistical analysis. MCMC allowed the establishment and estimation of complex models that could not be handled by conventional methods. Since the founding of the

 https://doi.org/10.26219/heal.aueb.6544

MCMC in Mathematical Sciences in 1990, many important papers have been published on this topic. Between 1990 and 1995, research focused on applying these new methods to a variety of popular models (e.g., Gelman & Rubin, 2004, Dellaportas & Smith, 1993, Gelfand et al., 1992) MCMC development also facilitated random effects and ordered samples.

1.3.2 Types of ANOVA

In this thesis, the focus is on two main types of ANOVA: one-way ANOVA and two-way ANOVA. Each method serves a specific purpose and is applied depending on the nature of the data and the research design.

One-way ANOVA is used to compare the means of three or more independent groups based on a single dependent variable (factor). This approach determines whether there are statistically significant differences between group means.

Two-way ANOVA tests the effect of two independent variables (factors) on a dependent variable and also examines interaction effects between two factors. This method is useful when researchers want to understand how two factors simultaneously influence the dependent variable.

1.3.3 Advantages of the Bayesian approach

Some of the principal advantages of the Bayesian approach and its computational techniques are as follows:

- Numerical Tractability: Bayesian computational techniques, such as Markov Chain Monte Carlo (MCMC), provide a solid infrastructure for parameter estimation of sophisticated models. According to Link et al. (2002), the one reassuring aspect of Bayesian inference is that in most cases, it gives the same results as classical statistical methods, but with the assurance of both reliability and consistency in statistical inference.
- Absence of asymptotics: While classical maximum likelihood (ML) inference is unbiased for large sample sizes (asymptotically), it can be biased for finite sample sizes (Le Cam, 1990). Furthermore, the validity of standard errors and

https://doi.org/10.26219/heal.aueb.6544

confidence intervals depends on large samples. In contrast, Bayesian inference is exact for any sample size.

- Ease of Error Propagation: Classical statistics render it challenging to estimate uncertainty for functions of random variables, i.e., parameters, and one must employ approximations like the delta method (Williams et al., 2002). Bayesian methods simplify the estimation of uncertainties based on posterior samples obtained by Markov Chain Monte Carlo (MCMC) method
- Formal Framework for Combining Information: Bayesian statistics formalizes the combination of prior knowledge (the prior) and new data (the observed data), using the laws of probability. This formal process can significantly improve the precision of estimates (McCarthy and Masters, 2007), with some parameters becoming estimable only through this combination. Classical statistics, on the other hand, assumes complete ignorance of the system being analyzed. However, Bayesian analysis can also allow for an assumption of ignorance if needed, in which case inferences based solely on observed data will often closely resemble classical methods.
- Intuitive Interpretation: Bayesian interpretation of probability is often more intuitive than classical methods since it allows us directly to compute the probability of a parameter taking some value, rather than a probability of observing the particular data set given a null hypothesis. That is why statements like "I am 99% sure that…" are only meaningful in Bayesian inference but not in classical approaches. This is because, in Bayesian analysis, the probability pertains to a parameter, while in classical statistics, it refers to a data set. Furthermore, Bayesian statistics, in melding prior knowledge with current data, provide a mathematical model of the learning process, which parallels how we process information both in science and in daily life.

1.3.4 Comparison of Bayesian approach with frequentist approach

Bayesian and frequentist approaches differ fundamentally in how they treat data and parameters. Bayesian inference views the data as fixed and the parameters as random variables, which capture the uncertainty about their true values. Frequentist inference,

https://doi.org/10.26219/heal.aueb.6544

on the other hand, views the parameters as fixed and the data as random, employing repeated sampling to make inferences. As Link et al. (2010) put it, the Bayesian method directly responds to the question at hand by making probabilistic statements on parameters given the observed data, rather than using estimators' distributions over imaginary future samples.

Bayesian inference estimates a complete probability model. In frequentist inference, there is no associated probability distribution over parameters or hypotheses. Bayesian estimates are therefore of the form $P(hypothesis \mid data)$, whereas frequentist estimates are of the form $P(data \mid hypothesis)$. This means that the term 'hypothesis testing' in frequentist statistics is actually a misnomer. What is being tested is the data, given the hypothesis.

Bayesian inference has a solid axiomatic basis (Cox, 1946), which even frequentists do not question. From the frequentist perspective, Bayesian methods are coherent, but frequentist inference is considered incoherent from the Bayesian perspective. Furthermore, Bayesian inference has a strong decision-theoretic basis (Roberts, 2007; Bernardo & Smith, 2000), wherein the purpose of statistical inference is to facilitate decision-making processes. As such, the Bayesian approach results in optimal decisions within this framework.

A major strength of Bayesian inference is that the uncertainty is included in the probability model, which makes the predictions more realistic. In contrast, frequentist methods do not include uncertainty in the parameter estimates, which makes the predictions less realistic. From a philosophical viewpoint, Bayesian inference corresponds to the idea that knowledge cannot be based only on experimentation but also must include prior information (Roberts, 2007).

In general, Bayesian inference for model comparison makes use of metrics such as the Deviance Information Criterion (DIC), even for hierarchical and complex models. Frequentist fit statistics cannot easily allow the comparison of models that differ in structure or have hierarchical components. Moreover, Bayesian methods avoid overfitting because model parameters are integrated out. Overfitting is an issue only in frequentist approaches. While Bayesian inference is not immune to overfitting, it is much less vulnerable than the frequentist techniques.

Another difference: Bayes's inference relies exclusively on the data at hand, whereas frequentist inference based methods include not only real but also hypothetical future data. In Bayesian methods, the incorporation of prior distributions

https://doi.org/10.26219/heal.aueb.6544

further adds to making more information available for analysis. Hence, the 95% probability intervals obtained from a Bayesian posterior distribution are also usually narrower compared to 95% confidence intervals for the frequentist point estimate.

Finally, Bayesian inference enabled by MCMC algorithms, allows for the estimation of complex models that are beyond the reach of frequentist methods. This computational advantage makes Bayesian inference much more versatile and better suited for modern, intricate statistical modelling.

1.3.5 Reasons Why the Bayesian Approach to ANOVA Has Not Been Widely Adopted

Even though there are many benefits of the Bayesian approach to statistics, it is still surprising that most statisticians still choose classical ANOVA mostly. Many factors account for the restricted implementation of Bayesian ANOVA, many of them derived from both philosophical and practical challenges.

One of the main reasons why people resist Bayesian ANOVA is the perceived subjectivity that arises from the selection of prior distributions. Statisticians argue that when selecting a prior, things may be manipulated entering the study by a mere absence of objective criteria for selecting the right ones. This problem is particularly evident in the case of ANOVA, where the use of prior items makes it more complicated to compare group means which is usually the basic approach.

In a few scenarios, Bayesian methods come most handy, when handling complex models that might not be feasible to analyse using frequentist methods. However, ANOVA is commonly used in cases where the models are relatively simple, such as comparing means across groups. These advantages might not seem that relevant in ANOVA cases, especially when the classical ANOVA which offers a simple and reliable solution is the widely used.

Moreover, computational techniques and Bayesian philosophy are yet to be popular among university students, particularly at undergraduate level. This lack of experience implies that many researchers are not currently acquainted with the computational tools and methods required for implementing Bayesian ANOVA, which hampers its adoption further.

https://doi.org/10.26219/heal.aueb.6544

The need for specialized computational tools was probably the most important barrier to the adoption of Bayesian ANOVA. In the past, a Bayesian ANOVA was executed using handmade codes written in general-purpose programming languages like FORTRAN or specialized statistical software. This requirement, therefore, poses a big challenge to the researchers who are not very competent in both computer programming and statistics and makes the Bayesian ANOVA less accessible to the wider scientific community.

1.4 Multiple Testing Problem

In many different scientific fields, testing hypotheses is an essential phase in the experimental research process. Highlighted in a humorous way by an fMRI study on an Atlantic salmon (Bennett et al., 2009). In this study, the dead salmon was put in an fMRI scanner and asked to identify emotions in human pictures. Remarkably, statistically significant clusters of active voxels were discovered in the salmon's brain although, of course, the salmon was already dead. Although the study was not intended as an attack on fMRI technology, the result was a graphic lesson in how multiple comparisons may lead to spurious conclusions. This problem goes beyond neuroscience. In genetics, for instance, researchers commonly run hundreds of thousands of tests in search of associations between genotypes and phenotypes (Storey & Tibshirani, 2003). In economics, analysts test many trading strategies in hopes of selecting the best performing strategy (Romano & Wolf, 2005). Clinical trials often compare many treatments to find the best (Fleming, 1982). These examples illustrate that the multiple comparisons problem is a common problem across many scientific fields.

Multiple testing arises in high-dimensional contexts where more than one hypothesis is evaluated simultaneously. Using a predefined significance level, such as $\alpha=0.05$ or $\alpha=0.1$, is appropriate for single hypothesis tests, since it ensures the probability of a false discovery (incorrectly rejecting the null hypothesis) is below α. This approach is problematic in the case of many hypotheses being tested. For example, in proteomic experiments, it may be thousands of proteins that are scanned to be differentially present in some particular biological activity. Clinical trials will also often compare treatment and control groups across a large number of outcome

https://doi.org/10.26219/heal.aueb.6544

measures. In genomics, researchers seek to identify differentially expressed genes among hundreds of thousands of possibilities.

The issue lies in the application of traditional single testing strategies. This frequently lead to false rejections of the null hypothesis when applied to multiple hypotheses. Testing each hypothesis independently at a significance level α may seem reasonable, but the probability of obtaining at least one false significant result increases dramatically as the number of tests grows. For example, testing 30 independent hypotheses at α=0.05 results in a probability of observing at least one false significant result of approximately 78%

$$P(\textit{at least one significant result}) = 1 - (1 - 0.05)^{30} \approx 0.785.$$

As the number of hypotheses (m) increases, the likelihood of rejecting a valid null hypothesis due to chance alone $1-(1-\alpha)^m$.

For instance, a microarray experiment testing 10,000 hypotheses would have about 500 genes, by chance alone, meeting the conventional p-value threshold of 0.05 when, in fact, no genes were significant. These types of errors can simply be waste of resources, time, and effort, or much worse, especially in areas involving critical fields like clinical trials and pharmaceutical research.

In order to reduce this problem, multiplicity adjustments are applied in an effort to bound FDR and other metrics of error. This is especially necessary in such techniques as ANOVA, which compares many groups means simultaneously. ANOVA, for example, suffers from the problems of multiplicity when analyzing data containing a large number of factors or interactions.

Adjustment techniques work to restrain error rates within multiple comparisons. Frequentist approaches, including the Bonferroni correction, typically involve a correction of p-values, considering the number of set hypotheses. While effective, these approaches are often conservative and result in a loss of statistical power. Bayesian methods provide a more flexible alternative. Bayesian ANOVA, for example embeds prior knowledge in dynamically updating the probability estimate as new data are added. This probabilistic framework naturally allows multiple hypotheses and greatly reduces false discoveries while keeping the robustness of the analysis.

 https://doi.org/10.26219/heal.aueb.6544

1.4.1 Bayesian approach

In Bayesian approaches to multiplicity test problem, multiplicity is accounted for by specifying appropriate prior probabilities on the hypotheses or models under consideration. This can be done by choosing appropriate prior distributions for the parameters of interest. According to Scott (2009), in many simultaneous testing situations, the Bayesian solution to the problem of multiplicity is straightforward. The Bayesian approach itself does not need any penalty terms, as is the case with frequentist methods. This comes with an implicit penalty which is sometimes referred to as the Occam's razor effect (Jeffreys & Berger,1992) that favors more simplistic models.

For instance, in a microarray study, the task could be to classify observations as either signal or noise. Signals represent meaningful patterns or true effects in the data, while noise corresponds to random changes or irrelevant variations. Assume the signals and noises follow distributions g_1 and g_0, respectively, and the observations x_i are modelled as a mixture:

$$x_i \sim p g_0 + (1-p) g_1$$

Here, p is the prior probability that the null hypothesis is true. Specification of an appropriate prior on p results in the automatic elimination of the multiplicity problem. As strong signals are discovered, p gets pulled toward zero. See Scott and Berger (2006) for a detailed discussion of this approach for the mixture model given here, and also Scott and Berger (2010) for extension to variable selection in linear models.

One of the primary strengths of the Bayesian adjustment is its simplicity and adaptability. If the data have a dependent structure, the Bayesian procedure is unaffected by the dependence, since the prior probability assignment inherently accounts for multiple testing. Moreover, Bayes theorem provides the procedure with a natural adaptation to the dependence structure in the data. This is in contrast to frequentist methods, which might be overly conservative to cope with the presence of dependence, and quite often do not account for such dependence while trying to control the multiplicity. These aspects make the Bayesian approach very appealing for handling multiple testing situations in an effective manner.

https://doi.org/10.26219/heal.aueb.6544

1.5 Closing Remarks

This chapter has provided an overview of classical and Bayesian statistical inference methods, specifically for ANOVA. We covered the fundamental contrasts between frequentist and Bayesian methods, focusing on their philosophical foundations and pragmatic implications. While classical ANOVA remains a widely used hypothesis testing method, its reliance on p-values has been criticized, particularly because of its vulnerability to issues such as multiple comparisons and publication bias.

Bayesian ANOVA, however, offers a more general and informative alternative by incorporating prior information and revising probabilities based on observed data. Through the application of Bayes factors, Bayesian inference offers a coherent solution to hypothesis testing that avoids most of the issues associated with frequentist approaches. Furthermore, recent advances in computation, such as Markov Chain Monte Carlo (MCMC) techniques, have facilitated the growing use of Bayesian approaches to statistical modelling.

Despite these advantages, Bayesian ANOVA remains to find universal application because of practical impediments in terms of prior specification issues, computing complexity, and the overall unawareness among researchers. However as packages like BayesFactor, JAGS, and JASP gain favor, Bayesian procedures are bound to gain wider recognition among scientists.

Besides explaining the merits of Bayesian inference, this chapter also touched upon the problem of multiple hypothesis testing, a significant challenge to statistical analysis. The Bayesian framework naturally accommodates this issue through the use of prior distributions that serve to reduce the risk of false discoveries without the need for extremely conservative corrections.

The next chapter will look at the mathematical foundations of Bayesian inference in greater detail, presenting Bayesian probability rigorously, priors, and key computational techniques. These will form the foundations upon which Bayesian ANOVA can be applied in practice, with additional contrasts between its application and standard frequentist methods.

https://doi.org/10.26219/heal.aueb.6544

https://doi.org/10.26219/heal.aueb.6544

CHAPTER 2

Bayesian ANOVA

2.1 Theoretical Framework – One-Way ANOVA

In this chapter, we focus on the one-way ANOVA model, assuming that the data are generated according to the full model, denoted as M_F. In this context, the data follow the model:

$$Y_{ij} = \mu_i + \varepsilon_{ij},$$
$$j = 1, \ldots, n_i \quad i = 1, \ldots, k$$

where $\mu = (\mu_1, \ldots, \mu_k)$ represents the group means, and σ^2 is the common variance, both of which are unknown parameters. The residual errors ε_{ij} are independent and identically distributed (i.i.d.) from a normal distribution with mean zero and variance σ^2, i.e., $\varepsilon_{ij} \sim N(0, \sigma^2)$.

In classical ANOVA, the focus is on testing the null hypothesis H_0, which assumes that all group means are equal, i.e., $\mu_1 = \mu_2 = \ldots = \mu_k$, against the alternative hypothesis H_1, which posits that at least one pair of group means are unequal.

The test is typically performed using the F-statistic, which is computed as:

$$F = \frac{\sum_{i=1}^{k} n_i \left(Y_i - \bar{Y}\right)^2 / (k-1)}{\sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(\bar{Y}_{ij} - \bar{Y}_i\right)^2 / (n-k)}$$

where $n = \sum_{i=1}^{k} n_i \bar{Y}_i = n_i^{-1} \sum_{j=1}^{n_i} \bar{Y}_{ij}$ and $\bar{Y} = n^{-1} \sum_{i=1}^{k} \sum_{j=1}^{n_i} Y_{ij}$

It is well known that under the null hypothesiH_0, the F statistic follows a $F_{k-1,n-k}$ distribution. Under the alternative hypothesis H_1, the F statistic is distributed according to a non-central $F_{k-1,n-k,\lambda}$ distribution, whose noncentrality parameter λ is given by

https://doi.org/10.26219/heal.aueb.6544

$$\lambda = \frac{1}{\sigma^2} \sum_{i=1}^{k} n_i \left(\mu_i - \bar{\mu} \right)^2$$

where $\bar{\mu} = \frac{1}{n} \sum_{i=1}^{k} n_i \mu_i$

2.2 Bayes Theorem

In the Bayesian framework, the unknown parameters in a statistical model are treated as random variables. Before any data is observed, a prior distribution is used to express our initial beliefs or knowledge about the parameters. Once data becomes available, this prior information can be updated by using the conditional distribution of the parameters given the data. This process of updating from prior to posterior distribution is achieved through Bayes' theorem.

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

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

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

$$f\left(A | B\right) = \frac{f\left(B | A\right) p\left(A\right)}{f\left(B\right)}$$

The equation above is known as Bayes' rule, which forms the foundation of Bayesian inference. It provides a probabilistic framework for learning from data (Bernardo and Smith, 1994). However, the denominator in Bayes' rule, f(B), involves complex high-dimensional integrals that are often analytically impossible to solve. In the past, numerical approximations were used to tackle these integrals, but they were often inadequate or not solvable at all. As a result, despite having theoretically better solutions than classical statisticians, Bayesians were limited in practice to solving only very simple problems. This limitation was eventually overcome with the advent of Markov Chain Monte Carlo (MCMC) methods and Gibbs sampling, which made it feasible to apply Bayesian methods to more complex problems.

 https://doi.org/10.26219/heal.aueb.6544

Hence, after observing data $\left(y_1,y_2,\ldots,y_n\right)$ we calculate the posterior distribution $f\left(\theta|y_1,\ldots,y_n\right)$ 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.

$$f\left(\theta|y_1,y_2,\ldots,y_n\right)=\frac{\prod_{i=1}^{n} f\left(y_i|\theta\right)g\left(\theta\right)}{\int_{-\infty}^{\infty}\prod_{i=1}^{n} f\left(y_i|\theta\right)g\left(\theta\right)d\theta}\propto\prod_{i=1}^{n} f\left(y_i|\theta\right)g\left(\theta\right)$$

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

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

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

$$\prod_{i=1}^{n} f\left(y_i|\theta\right)=f\left(y_1|\theta\right)f\left(y_2|\theta\right)\ldots f\left(y_n|\theta\right)$$

And for k independent unknown parameters,

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

The posterior distribution can often be highly complex, making it challenging to compute integrals or to derive summary statistics like the posterior mean, variance, or probabilities. The development of simulation-based methods, such as Markov Chain Monte Carlo (MCMC), has immensely enabled the practical use of Bayesian inference in the real world. With these approaches, researchers can sample from the posterior distribution, thus allowing for parameters to be estimated even for complex models. As highlighted by Link and Barker (2010), Swain et al. (2009), McCarthy and Masters (2007), Mazzetta et al. (2007), and Gelman (2006), these methods steer clear of having to directly compute the normalizing constant in Bayes' rule, streamlining the estimation process and making the Bayesian approach computationally feasible. As a result, it becomes essential to use Monte Carlo methods to compute the required quantities. However, sampling directly from a complex, high-dimensional distribution is often difficult or even unfeasible. MCMC addresses this by simulating a Markov Chain where the posterior distribution acts as the stationary or limiting distribution.

https://doi.org/10.26219/heal.aueb.6544

The combination of MCMC methods and the rapid advancement of computational power, which these iterative techniques heavily rely on, has driven the Bayesian revolution in statistics (Brooks, 2003). The ability of MCMC algorithms to solve complex computational problems with ease is one of the primary reasons for the recent resurgence of Bayesian statistics, rather than just the flexibility of conducting inference without assuming complete ignorance (i.e., having no prior knowledge of the system under analysis). As a result, the posterior distribution plays a crucial role in Bayesian inference.

2.2.1 Bayesian estimation

In Bayesian estimation (O'Hagan and Forster, 2004, Lindley, 2000, Bernardo and Smith, 1994), uncertainty regarding parameters is expressed through probability distributions. Given a model M, if we aim to estimate the parameters θ, we first establish a prior distribution for these parameters, denoted as f(θ|M). Once we observe data Y, this prior distribution f(θ|M) is updated to form the posterior distribution f(θ|Y,M). This update follows Bayes' rule.

$$f(\theta|Y,M)=\frac{f(Y|\theta,M)f(\theta|M)}{f(Y|M)}=\frac{f(Y,M|\theta)f(\theta|M)}{\int f(Y,M|\theta)f(\theta|M)\,d\theta}f(Y|\theta,M)f(\theta|M)=\text{Likelihood}\times\text{ prior}$$

Thus, the posterior distribution of the parameter is directly proportional to the product of the likelihood and the prior. In Bayesian parameter estimation, the goal is to determine the posterior distribution of the model parameters f(θ|Y,M). On the other hand, in Bayesian model selection, the emphasis shifts to f(Y|M), which represents the marginal likelihood of the data given model M.

2.2.3: Bayesian model selection

In Bayesian model selection, different statistical models or hypotheses are given prior probabilities. For instance, if we have two competing models, M_1 and M_2, they are assigned prior probabilities $f(M_1)$ and $f(M_2)$. Once the data has been observed, the comparative likelihood of M_1 and M_2 is determined by the ratio of their posterior probabilities, known as the posterior ratio.

https://doi.org/10.26219/heal.aueb.6544

$$\frac{f(M_1|Y)}{f(M_2|Y)}=\frac{f(M_1)f(Y|M_1)}{f(M_2)f(Y|M_2)}$$

The posterior odds are determined by multiplying the prior odds with the ratio of marginal likelihoods, a term referred to as the Bayes factor (Kass and Raftery, 1995, Berger and Sellke, 1987, Dickey, 1971, Jeffreys, 1961). The Bayes factor measures how much the evidence from the data shifts the prior odds to posterior odds, and importantly, it is independent of the prior probabilities of the models $f(M_1)$ and $f(M_2)$. It serves to quantify the strength of the data in favor of one model over another.

In linear regression and ANOVA, two notable models are the null model M_0, which includes only the intercept but no predictors, and the full model M_F, which incorporates all the predictors of interest. A key issue with the Bayes factor in such cases is its sensitivity to the choice of the prior distribution for the parameters being tested (Gelman, 2008, Berger 2006, Press et al., 2003, Gelman et al., 1996).

When there is limited information about the phenomenon being studied, the prior distribution for the parameters should be relatively uninformative. However, to prevent paradoxical outcomes, it's crucial that the prior distribution isn't excessively uninformative. The Jeffreys-Lindley-Bartlett paradox (Robert, 1993, Berger & Delampady, 1987, Lindley, 1980, Jeffreys, 1961) is a major area of interest that occurs if one employs Bayesian hypothesis testing using vague, uninformative priors. Specifically, when these priors are employed, the Bayes factor favors the null hypothesis strongly regardless of the available data. This is because the marginal likelihood, f(Y|M), is determined as an average of the likelihood over the prior distribution. If the prior is too diffuse, its probability mass lies in vast regions of the likelihood space where it is very unlikely, thus reducing the overall marginal likelihood and unfairly increasing the null model support.

https://doi.org/10.26219/heal.aueb.6544

2.3 Prior distributions

2.3.1 Conjugate Prior distributions

A prior distribution that belongs to a family of distributions D with parameters α is referred to as conjugate to the likelihood function $f(y|\theta)$ if the posterior distribution $f(\theta|y)$ remains within the same family of distributions. Thus, if $\theta \sim D(\alpha)$, then the posterior will also follow a distribution from the same family, denoted as $D(\bar{\alpha})$, where $\bar{\alpha}$ represents the posterior parameters.

In many cases, the target posterior distribution cannot be solved analytically. This issue was mitigated by using conjugate priors, which have the useful characteristic of leading to posteriors that belong to the same family of distributions as the prior. Bernardo and Smith (1994) provide detailed examples of conjugate priors. In this context, α and $\bar{\alpha}$ correspond to the parameters of the prior and posterior distributions, respectively. In simpler models, these posterior parameters are often expressed as a weighted combination of the prior parameters and the maximum likelihood estimates.

2.3.2 Non-informative and weakly-informative prior distributions

A prior distribution is considered weakly informative if it is a proper distribution but it is constructed in such a way that it conveys less information than any actual prior knowledge. On the other hand, non-informative prior distributions are designed to enable Bayesian inference for parameters where little is known beyond the data being analyzed.

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 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 $f(\mu) \sim Normal(\mu_0,\tau^2)$ and $f(\sigma^2) \sim InverseGamma(\alpha,\beta)$

However, in this situation, the posterior distribution takes an unknown form, making it difficult to compute posterior summaries or marginal densities directly.

 https://doi.org/10.26219/heal.aueb.6544

When conjugate priors are either unavailable or deemed unrealistic, alternative methods such as asymptotic approximations (e.g., Laplace approximation as proposed by Kass and Raftery, 1995) or numerical integration can be employed. Another popular approach is to use simulation-based techniques, which generate samples from the posterior distribution.

2.3.3 Conditionally conjugate prior distributions

A family of prior distributions $f(\theta)$ is considered conditionally conjugate for θ if the conditional posterior distribution $f(\theta|y)$ belongs to the same family as the prior. From a computational standpoint, conditional conjugacy implies that if it is possible to sample θ from the prior distribution, it is also possible to use a Gibbs sampler to draw samples from the posterior distribution of θ. More importantly, conditional conjugacy allows for an intuitive interpretation of the prior distribution in terms of equivalent data.

2.3.4 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

$$a_j \sigma^2 \sim Normal(\mu, gV\sigma^2)$$

where α_j represents the components of the prior mean vector.

where $\sigma^2 \sim InverseGamma\left(\frac{v_0}{2}, \sigma_0^2 \frac{k_0}{2}\right)$

Where g is the parameter controlling the overall magnitude of the prior variance and

$$V = \left(X^T X\right)^{-1}$$

A common default choice is $g=n$, where n is the number of observations. This is because $g=n$ corresponds to adding prior information equivalent to one additional data point (Kass and Wasserman, 1995). Another option for g is to set it to the square of the number of predictors in the model, $g = k^2$, which is known as the Risk Inflation Criterion (Foster and George, 1994). Fernandez et al. (2001) suggested using $g=max\{n,k^2\}$ as a "benchmark prior," a choice that simplifies posterior computations

https://doi.org/10.26219/heal.aueb.6544

by reducing the number of prior variance parameters to a single value. This makes it easier to compare different values of g, a topic explored by Liang et al. (2008).

When no information is available, the above prior set up is simplified by letting the matrix $V=XI_j$ 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

$$a_j\sigma^2 \sim Normal\left(\mu, gV\sigma^2\right) \text{ for } j=1,\ldots,J$$

Alternatively, one can assume that all parameters are a priori independent. In this case, the prior is not conjugate, and Markov Chain Monte Carlo (MCMC) methods must be used to estimate the posterior distribution. However, this setup is still conditionally conjugate, meaning the conditional posterior distributions of the α_j remain in the same family, enabling the use of an efficient Gibbs sampler for inference. The Bayes factor for the full model to the null model is:

$$BF=\left(1+g\right)^{(n-k-1)/2}\cdot\left[1+g\left(1-R^2\right)\right]^{-(n-1)/2}$$

2.3.5 Jeffrey-Zellner-Siow Prior

The JZS prior is a well-known non-informative prior for Bayesian hypothesis testing and regression modelling, specifically in Bayesian model comparison. It was introduced as an improvement over Zellner's g-prior, improving on some of its limitations through the utilization of a Cauchy prior structure to achieve increased robustness. The JZS prior puts a prior distribution on g=n, typically of scaled inverse-gamma or Cauchy form. This alleviates issues with sensitivity to the specification of g and provides automatic adaptation to the data structure. The JZS prior is particularly useful when objective Bayes methods are sought, as it maintains scale invariance and gives rise to marginal likelihoods that are well-defined even in small samples. It is widely applied in Bayesian hypothesis testing, including Bayes Factors for linear regression models, and has been extensively studied in the context of Bayesian ANOVA and variable selection. The Bayes factor for the full model to the null model is:

$$BF=\left(1+g\right)^{(n-k-1)/2}\cdot\left[1+g\left(1-R^2\right)\right]^{-(n-1)/2}$$

https://doi.org/10.26219/heal.aueb.6544

2.3.6 Hyper g prior

The Hyper-g prior is a family of prior distributions for he parameter g for which the probability density function is of the form

$$p(g)=\frac{(a-2)}{2}(1+g)^{-a/2}, g>0$$

The value of a typically lies in the range 2 to 4 and controls the behavior of the prior. In practice, common choices are $a=3$ and $a=4$. An arbitrary choice of proportionality constant can produce undefined Bayes factors.

The Bayes factor (BF) for the comparison of the full model to the null model is given by:

$$BF=\frac{(a-2)}{2}\int_0^\infty (1+g)^{(n-k-1-a)/2}\left[1+g\left(1-R^2\right)\right]^{-(n-1)/2} dg$$

2.4 Foundation and Interpretation of the Bayes Factor

The Bayes Factor (BF_{01}) measures how much more likely the data are to be observed under the null hypothesis H_0 compared to the alternative hypothesis H_1. In other words, BF_{01} represents the relative level of support the data provide for H_0 versus H1. A BF_{01} of 1 indicates no preference for either hypothesis. When BF_{01} is greater than 1, the data favor H_0, whereas a value between 0 and 1 suggests stronger support for H_1.

Mathematically, the Bayes Factor is formulated as the marginal likelihood ratio of the data for each hypothesis:

$$BF_{01}=\frac{p(y|H_0)}{p(y|H_1)}$$

where $p(y|H_0)$ and $p(y|H_1)$ are the probability of observed data y under the null hypothesis and alternative hypothesis, respectively. This is a measure of how much more probable the data is under one of the two hypotheses than the other.

Based on Bayes' theorem, the posterior probability of a hypothesis H under observed data y is given by:

$$p(H|y)=\frac{p(y|H)\,p(H)}{p(y)}$$

https://doi.org/10.26219/heal.aueb.6544

Since p(y) is the same for both hypotheses, the posterior odds of versus are:

$$\frac{p(H_0|y)}{p(H_1|y)} = BF_{01} \times \frac{p(H_0)}{p(H_1)}$$

where $p(H_0)$ and $p(H_1)$ are the prior probabilities of H_0 and H_1 respectively. The Bayes Factor thus updates the prior odds into the posterior odds, determining how the data modify our beliefs about the hypotheses.

The Bayes factor determines the support for a hypothesis by balancing its relative fit against its relative complexity by comparing H_0 with H_1. A strong hypothesis not only fits the data well but also avoids unnecessary complexity, being both specific and parsimonious. This principle aligns with Occam's razor that suggests that when two hypotheses explain the data equally the simpler one is favored. Therefore, if the observed data aligns well with H_0, this simpler hypothesis will be preferred over the more complex H_1. As demonstrated by Hoijtink (2012, pp. 59, Section 3.7.1), under certain conditions, the Bayes factor can be expressed as the ratio $BF_{01} = \frac{f_0}{c_0}$, where f_0 represents the relative fit of H_0, and c_0 represents its relative complexity compared to H_1.

BF	Interpretation
> 100	Extreme evidence for H1
30 – 100	Very strong evidence for H1
10 – 30	Strong evidence for H1
3 – 10	Moderate evidence for H1
1 – 3	Anecdotal evidence for H1
1	No evidence
1/3 – 1	Anecdotal evidence for H0
1/10 – 1/3	Moderate evidence for H0
1/30 – 1/10	Strong evidence for H0
1/100 – 1/30	Very strong evidence for H0
< 1/100	Extreme evidence for H0

Table 2.1: Bayes Factor thresholds for interpreting the significance of evidence

Researchers often pose the question of the appropriate size of the Bayes factor BF_{01} necessary to draw definitive conclusions about hypotheses. Specifically, they seek to

https://doi.org/10.26219/heal.aueb.6544

determine how large BF_{01} should be to favor the null hypothesis and how small it should be to support the alternative hypothesis. This reflects a common desire for a clear threshold, similar to the α-level of 0.05 used in null hypothesis significance testing (NHST), which aids in deciding which hypothesis to accept. However, the Bayes factor differs from NHST as it does not lead to a binary decision of rejecting or not rejecting H_0. Instead, it quantifies the degree of support in the data for the competing hypotheses. When BF_{01} is approximately 1, there is no clear preference for either the null or the alternative hypothesis, indicating that the Bayes factor may be inconclusive. In such cases, additional data may be required to provide stronger evidence for one hypothesis over the other. It is evident that a Bayes factor of 100 (or 0.01) indicates strong support for the null hypothesis (or the alternative hypothesis) with a Bayesian error probability so low (0.01) that a decisive conclusion about which hypothesis is best can be confidently made. When BF_{01} is 10 (or 0.1), there remains a preference for H_0 (or H_1) but with a Bayesian error probability of 0.09, the alternative hypothesis cannot yet be disregarded. In contrast, a BF_{01} of 2 (or 0.5) suggests uncertainty about whether it is advisable to favor H_0 over H_1 or vice versa, as the Bayesian error probability rises to 0.33. Therefore, formal threshold values for interpreting the Bayes factor are unnecessary. The evidence for the competing hypotheses is adequately conveyed by the Bayes factor itself.

Null Hypothesis Significance Testing focuses on the null hypothesis, and the result is either its rejection or non-rejection. In contrast, when using the Bayes factor to compare H_0 and H_1, both hypotheses are treated equally, without assigning one as the null or alternative. The Bayes factor is derived by calculating the probability of the observed data under each hypothesis, and this ratio determines the preference. This means the Bayes factor can favor either H_0 or H1, depending on which hypothesis better explains the data.

2.5 Modelling data and parameters that vary by groups

Consider y_{ij} as the response variable, where $i=1,\ldots,n_j$ indexes the observations within each group j=1,…,J and let n denote the total number of observations. Within each group j the mean of y is α_j . For simplicity, we assume homoskedasticity across groups, such that the variance of y_j is $V(y_j)=\sigma^2$ for all j. The means $\bar{y}_1,\ldots,\bar{y}_j$ and

https://doi.org/10.26219/heal.aueb.6544

an estimate of the common variance σ^2 are sufficient statistics for the data if we assume a normal model for the data. Our objective is to explore the potential variation of these means across groups, which can be achieved through a hierarchical model. The model is specified as follows:

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

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

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

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

$$y_{ij}|\alpha_j,\sigma^2\sim Normal(\alpha_j,\sigma^2) \quad (2.1)$$

$$\alpha_j|\mu_0,\omega^2\sim Normal(\mu_0,\omega_0^2) \quad (2.2)$$

Equation 2.1 is a normal model for the data, with the parameters a_j (the group means) and σ^2 (the within-group variance). Equation 2.2 describes how the group means α_j vary across groups: μ_0 is the mean of the distribution of group means, while ω_0 represents the variance of this distribution (also known as the between-group variance). Here σ^2 is the within-group variance for each group j.

To complete the model specification, prior distributions must be defined for the hyperparameters μ_0 and ω_0, as well as for σ^2, the within-group variance. These prior distributions reflect our initial beliefs about the parameters before observing the data.

2.6: Derivation of posterior distributions

Assume the model $y_{ij}\sim Normal(\alpha_j,\sigma^2)$ and $\alpha_j\sim Normal(\mu_0,\omega_0^2)$ where $i=1,\ldots,n_j$ indexes observation with group j, $j=1,\ldots,J$. Then $\alpha_j|y_{ij},\sigma^2,\mu,\omega^2\sim Normal(\bar{\mu}_j,v_j)$ or

$$\mu_j=\left(\frac{\frac{\mu_0}{\omega_0^2}+\frac{n\overline{y}_j}{\sigma^2}}{\frac{1}{\omega_0^2}+\frac{n_j}{\sigma^2}}\right) \text{ and } v_j=\left(\frac{1}{\omega_0^2}+\frac{n}{\sigma^2}\right)^{-1}$$

 https://doi.org/10.26219/heal.aueb.6544

And where $\overline{y_j}=n_j^{-1}\sum_{i=1}^{n_j} y_{ij}$ is the maximum likelihood estimate of a_j (i.e the group mean)

2.6.1 Posterior distribution from Normal prior and normal likelihood case

This approach is used when we have continuous data and aim to infer how the group means vary assuming that the variance σ^2 is known and that the observations y_{ij} follow a normal distribution. Given the observed data $y_1,\ldots,y_n$ groups $j=1,\ldots,J$ we focus on estimating the model parameters $\theta=\left(\alpha_1,\ldots,\alpha_J,\mu_0,\omega_0^2,\sigma^2\right)$

The Bayesian model is structured as follows:

- $y_{ij} \sim Normal\left(a_j,\sigma^2\right)$ model for the data
- $a_j \sim Normal\left(\mu_0,\omega_0^2\right)$ model of how the means vary across the groups, where $i=1,\ldots.,n_j$ are observation in the samples j, $j=1,\ldots,J$

The posterior distribution for the model was estimated by assuming a normal likelihood for the data, with a specified mean and variance σ^2. A conjugate normal prior was chosen for μ, with a mean of μ_0 and a variance ω_0^2.

$f\left(\alpha_j|y,\sigma^2\right)\propto f\left(y|\alpha_j,\sigma^2\right)f\left(\alpha_j\right)\propto likelihood\times prior$

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

$\alpha_j|y_j,\sigma^2,\mu_0,\omega_0^2 \sim Normal(u,V)$

Posterior distribution ∝ likelihood × prior

$f\left(\alpha_j|y,\sigma^2\right)\propto f\left(y|\alpha_j,\sigma^2\right)f\left(\alpha_j\right)\propto Normal\left(\alpha_j,\sigma^2\right)\times Normalprior\left(\mu_0,\omega_0^2\right)$

Let θ represent the parameter of interest, then

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

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

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

 https://doi.org/10.26219/heal.aueb.6544

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

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

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

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

The terms are then separated into several components. Since certain terms do not include α_j, they can be represented by a constant T, which will be absorbed into the normalizing constant.

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

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

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

$$f\left(\alpha_j|y\right)\propto\exp\left[-\frac{1}{2}\left(\frac{1}{\omega_0^2}+\frac{n}{\sigma^2}\right)\left(\alpha_j^2-2\alpha_j\left(\frac{v_0}{\omega_0^2}+n\frac{\bar{y}}{\sigma^2}\right)\right)+T\right]$$

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

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

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

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

https://doi.org/10.26219/heal.aueb.6544

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

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

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

The posterior precision is the sum of the prior precision and the data precision. Also, we examine the relationship between the prior mean and the posterior mean in greater detail. The posterior mean is:

$$\mu_j = \left(\frac{\dfrac{\mu_0}{\omega_0^2} + \dfrac{n\bar{Y}}{\sigma^2}}{\dfrac{1}{\omega_0^2} + \dfrac{n}{\sigma^2}}\right) = \frac{\left(\dfrac{\omega_0^2\sigma^2 + n\{\bar{y}\omega_0^2}{\omega_0^2\sigma^2}\text{¿}\right)}{\left(\dfrac{\sigma^2 + n\omega_0^2}{\omega_0^2\sigma^2}\right)} = \frac{\omega_0^2\sigma^2 + n\{\bar{y}\omega_0^2}{\sigma^2 + n\omega_0^2} = \frac{\omega_0^2\sigma^2}{\sigma^2 + n\omega_0^2} + \frac{n\bar{y}\omega_0^2}{\sigma^2 + n\omega_0^2}\text{¿}$$

As the sample size n increases, the data mean has a stronger influence compared to the prior mean. Also, when ω_0^2 decreases (indicating lower prior variance and higher prior precision), the prior mean plays a more significant role in estimation.

2.6.2: 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 and scale parameter β_0

$f(\sigma^2|y,\mu)\ \ f(y|\mu,\sigma^2)f(\sigma^2)$

Posterior distribution is therefore,

$Normal(\mu,\sigma^2) \times$ Inverse Gamma(α_0,β_0), where $\alpha_0 = \dfrac{v_0}{2}$ and $\beta_0 = v_0\dfrac{\sigma_0^2}{2}$

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

https://doi.org/10.26219/heal.aueb.6544

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

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

$$\theta^{-n/2}\exp\left(-\frac{\sum_{i=1}^{n}\left(y_i-\mu\right)^2}{2\theta}\right)\times\theta^{\left(-\theta_0-1\right)}\exp\left(-\frac{\beta_0}{\theta}\right)=\theta^{\left(-\left(n/2+\alpha_0+1\right)\right)}\exp\left(-\left[\frac{\beta_0}{\theta}+\frac{\sum_{i=1}^{n}\left(y_i-\mu\right)^2}{2\theta}\right]\right)$$

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

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

This is a density function of an Inverse Gamma distribution with parameters α_1 and β_1.

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

$$a_1=a_0+\frac{n}{2}$$

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

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

2.7: Hierarchical One-Way ANOVA

Analysis of Variance (ANOVA) is an extension of the t-test for a situation involving more than two groups. ANOVA exists in a number of different forms, including one-way ANOVA which examines one factor and two-way or multi-way ANOVA which examines two or more factors and may examine main effects and interactions.

https://doi.org/10.26219/heal.aueb.6544

In this chapter, we present one-way ANOVA and introduce the random effects concept. Group effects such as means are assumed to be drawn from a distribution in random-effects models, often a normal distribution. We start by constructing and analyzing fixed-effects ANOVA before proceeding to random-effects ANOVA.

The full formulation of a normal, one-way ANOVA model as a Bayesian hierarchical model is as follows:

$$y_{ij}|a_j,\sigma^2 \sim Normal\left(a_j,\sigma^2\right)$$

$$a_j|\mu_0,\omega_0 \sim Normal\left(\mu_0,\omega_0^2\right)$$

$$\mu_0 \sim Normal\left(b_0,B_0\right)$$

$$\sigma_0^2 \sim inverse-Gamma\left(\frac{v_0}{2},\sigma^2\frac{v_0}{2}\right)$$

$$\omega_0 \sim inverse-Gamma\left(\frac{\kappa_0}{2},\kappa_0\frac{\omega_0^2}{2}\right)$$

A model that incorporates unit-specific heteroskedasticity is achieved when the within-unit variance parameter, σ^2, is allowed to vary across units. In this case, instead of a single variance σ^2, we introduce separate variance parameters for each unit, denoted as σ_1^2, σ_2^2,..., σ_j^2. The hyperparameters for the normal prior, specifically the mean b_0 and variance B_0 along with the hyperparameters for the model's parameters, are contained in the vector $\theta=(\alpha_1,...,\alpha_J,\mu_0,\sigma^2)$.

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

$$\begin{aligned} f(\theta) &= f\left(\alpha_1,\ldots,\alpha_J,\mu_0,\sigma^2,\omega_0^2\right) \\ &= f\left(\alpha_1,\ldots,\alpha_J|\mu_0,\omega_0^2\right) f\left(\mu_0\right) f\left(\sigma^2\right) f\left(\omega_0^2\right) \\ &= \prod_{i=1}^{n} f\left(\alpha_j|\mu_0,\omega_0^2\right) f\left(\mu_0\right) f\left(\sigma^2\right) f\left(\omega_0^2\right) \end{aligned}$$

2.8 Posterior Analysis and Decision Making

In Bayesian statistics, one does not aim to estimate a point, as classical statistics does with the maximum likelihood estimate, but instead to find the whole posterior distribution of an unknown (e.g., a parameter, function of parameters, prediction, or residual). This distribution gives a better sense of uncertainty.

https://doi.org/10.26219/heal.aueb.6544

The posterior distribution can be plotted graphically using tools such as histograms or kernel density plots. For summarizing point estimates, measures of central tendency such as the mean, median, or mode of the posterior can be used. Additionally, the standard deviation of the posterior is a measure of uncertainty, just as in frequentist statistics the standard error is.

Unlike a traditional statistics 95% confidence interval, Bayesian analysis makes use of the credible interval (CRI), i.e. the interval including 95% of the posterior probability. Credible intervals are specified differently, and among the most used is the highest posterior density interval (HPDI). In this study, we employed 95% credible intervals as the range between the 2.5th and 97.5th percentiles of the posterior distribution for all parameters.

2.9 From Estimation to Prediction

Predictions represent the expected values of the response variable for future samples, hypothetical values of the explanatory variables in a model, or, more broadly, any unobserved quantity. Predictions play a key role in:

- Presenting the results of an analysis, and
- Gaining insight into what a model implies. For instance, understanding the biological significance of an interaction or a polynomial term may not be straightforward from parameter estimates alone.

Since predictions depend on both the parameters and data (i.e., values of covariates), their posterior distributions can be used for inference. Typically, the mean of the posterior distribution is used as the predicted value, along with a 95% credible interval (CRI) to indicate a range of uncertainty, effectively serving as a 95% prediction interval.

The one-way ANOVA can be structured in multiple ways. For our approach, we used a mean parameterization for the linear model in a fixed-effects one-way ANOVA:

$$y_i = \mu_j + \epsilon_i$$

where $\epsilon_i \sim Normal\left(0, \sigma^2\right)$

https://doi.org/10.26219/heal.aueb.6544

In this setup, y_i is the i-th observation within group j, μ_j represents the mean for group j, and ϵ_i is the random deviation of i from its group mean μ_j. This deviation is assumed to follow a normal distribution centered at zero with constant variance σ^2.

In a fixed effect ANOVA, the group means μ_j are treated as unknown constants that are estimated directly. However, if we assume a distribution for these group means, we get a random effects ANOVA:

$$y_i = \mu_j + \epsilon_i$$

where $\epsilon_i \sim Normal\left(0, \sigma^2\right)$

and $\mu_j \sim Normal\left(\mu_0, \omega^2\right)$

Here, μ_j still represents the group mean, and ϵ_i is still the residual, but μ_j values are no longer independent. Instead, they are assumed to be drawn from a higher level normal distribution with mean μ_0 and variance ω^2. These parameters μ_0 and ω^2 are called hyperparameters because they describe the distribution from which the μ_j values are drawn.

In practice, fixed effect factors might include categories like gender or crop variety in agricultural studies. Random effects factors, on the other hand, could represent time (e.g., year, month, or day) or location, such as blocks in an experiment or other spatial units where repeated measures are taken. This approach reflects the idea that these levels are influenced by a shared random process, creating a correlation among the effects within a random-effects factor.

When factor levels are treated as fixed, they are considered unrelated or independent. Moving from a fixed effects ANOVA to a random effects ANOVA is beneficial for several reasons:

- It allows inference to be extended to a larger population,
- It improves the handling of system uncertainty, and
- It enhances the efficiency of estimation.

Firstly, interpreting the effects being studied as a random sample from a population enables generalization to that population. This generalization is achieved by modelling the process that generates the random effects' values, often by assuming a normal distribution for the α_1 values (mentioned above). Secondly, specifying factor effects as random acknowledges that if the study were repeated, a different set of

https://doi.org/10.26219/heal.aueb.6544

effects would likely be observed, resulting in different parameter estimates. A random-effects model accounts for this additional uncertainty in system inference. Thirdly, under the random-effects assumption, estimates are not made independently but are instead influenced by one another, becoming dependent. Specifically, individual estimates are "drawn toward" the overall mean μ, making them closer to μ than they would be under a fixed-effects approach. This property is why random-effects estimators are known as "shrinkage estimators." Estimates with less precision and smaller sample sizes are more strongly pulled toward the mean. When effects are indeed interchangeable, shrinkage provides better estimates (e.g., with lower prediction error) than fixed-effects estimates (Gelman, 2007).

2.10 Two way ANOVA

The one-way ANOVA model can be easily extended to accommodate additional categorical variables.

In a two-way ANOVA design, Factor A has a levels and Factor B has b levels, arranged in a factorial setup where each combination of Factor A and Factor B levels is tested in every replicate. The experimental model for this design can be expressed as:

$$y_{ijk} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij} + \epsilon_{ijk}$$

where:

- μ is the overall mean effect,
- τ_i represents the effect of the i-th level of Factor A,
- β_j is the effect of the j-th level of Factor B,
- $(\tau\beta)_{ij}$ captures the interaction effect between Factor A and Factor B, and
- $\epsilon_{ijk\backslash}$ denotes a random error term.

Here, i=1,2,…,a, j=1,2,…,b and k=1,2,…,n where n represents the number of replicates for each treatment combination.

In a Bayesian framework, the two-way ANOVA model is reformulated as a regression model to simplify applying Bayesian techniques with different prior distributions (as suggested by Zellner et al.). This approach allows for testing the

 https://doi.org/10.26219/heal.aueb.6544

effects of the main factors, A and B, as well as their interaction, by comparing four potential ANOVA model structures.

The regression model for the two-way ANOVA experiment is structured as:

$$Y=X\beta$$

where X is a design matrix created based on the factors. This matrix may not have full column rank, as we apply sum-to-zero constraints (meaning the coefficient for the last level of each factor is set to the negative sum of the other levels' coefficients).

For the full two-factor ANOVA model, we include:

- the grand mean μ
- two levels each for the main effects of A and B,
- and interaction terms (excluding the third levels of A and B due to the sum-to-zero constraint).

Thus, the model can be expressed as:

$$Y=1_n\mu+X_{p1}\beta_{p1}+X_{p2}\beta_{p2}+X_{t1}\beta_{t1}+X_{t2}\beta_{t2}+X_{p1\times t1}\beta_{p1\times t1}+X_{p2\times t1}\beta_{p2\times t1}$$

where each term represents the corresponding levels of Factors A, B and their interaction, with constraints to ensure sum-to-zero properties for identifiability.

The two-way ANOVA model containing only the main effects, without interaction effects, is

$$Y=1_n\mu+X_{p1}\beta_{p1}+X_{p2}\beta_{p2}+X_{t1}\beta_{t1}+X_{t2}\beta_{t2}$$

The two-way ANOVA model containing only the main effect A is

$$Y=1_n\mu+X_{p1}\beta_{p1}+X_{p2}\beta_{p2}$$

The two-way ANOVA model containing only the main effect B is

$$Y=1_n\mu+X_{t1}\beta_{t1}+X_{t2}\beta_{t2}$$

The two-way ANOVA model without main and interaction effects (the null model) is

$$Y=1_n\mu$$

Let BF_i (where i=1,2,3,4) represent the Bayes factor for the i-th model compared to the null model, as described by Rouder et al. In this context, the full model is considered a larger model compared to the other models, meaning the reduced models

https://doi.org/10.26219/heal.aueb.6544

are nested within the full model. To examine the significance of different factors, we compare the reduced models to the full model.

To test the interaction effect, we compute the Bayes factor BF_{12} which is the Bayes factor for comparing the larger model with the reduced model. To find BF_{12} we first calculate the Bayes factors BF_{10} and BF_{20}, and the ratio of these Bayes factors is given by:

$$\frac{BF_{10}}{BF_{20}} = BF_{12}$$

This follows from the inverse and transitive properties of Bayes factors. Similarly, the ratios BF_{23} and BF_{24} can be derived by using BF_{20} and BF_{30} and BF_{20} and BF_{40} respectively to assess the effects of the main factors. However, relying on a single prior does not provide definitive conclusions about the impact of main and interaction effects.

2.11 Conclusion

This chapter provided an extensive overview of Bayesian ANOVA. It gives details to theoretical aspects and comparing it to traditional one-way and two-way ANOVA models. We initially outlined the one-way ANOVA model, detailing its Bayesian formulation and prior distribution usage via different priors. Bayesian estimation, model comparison, and hypothesis testing based on Bayes factors were also covered in the chapter.

One of the key benefits of Bayesian ANOVA is that it is hierarchical in nature, thus allowing modelling group-level variation. Bayesian inference provides a richer understanding of group differences, uncertainty estimation, and parameter estimation via posterior distributions. The use of prior information also renders decision-making more informed, while computational advances such as MCMC methods allow one to conduct inference even in complex models.

Extending beyond one-way ANOVA, we introduced the Bayesian two-way ANOVA, demonstrating how factorial designs may be expressed within a Bayesian framework. The fact that the two-way ANOVA could be expressed as a regression model allowed for more flexible hypothesis testing since different model

https://doi.org/10.26219/heal.aueb.6544

specifications could be tested using Bayes factors. The approach enables researchers to avoid the limitations of p-value-based inference.

The next chapter will focus on multiplicity adjustments in Bayesian ANOVA. We will explore how Bayesian approaches naturally incorporate multiplicity correction through hierarchical modelling, shrinkage priors, and prior probability adjustments, providing a more coherent alternative to traditional frequentist correction methods.

https://doi.org/10.26219/heal.aueb.6544

https://doi.org/10.26219/heal.aueb.6544

CHAPTER 3

Multiplicity adjustment

The problem of multiplicity in statistics is: how should we adjust our criteria of significance when we give data multiple chances to surprise us? For instance, if we flipped a coin ten times and got ten heads in a row, that would be remarkable, but if we flipped it 100,000 times, a run of ten heads becomes almost sure. Between these two extremes lies a gray area in which we must consider how to quantify our degree of skepticism.

The fundamental challenge of multiplicity adjustment lies in moderating the threshold as the number of potential findings increases. This is a concern that extends to many modern experiments, such as microarrays, fMRI, environmental sensor networks, and proteomics. These various fields all share a common essential goal: to isolate the important useful information (signals) from unwanted variation (noise). The tendency toward Type I errors means many discoveries will require followup validation, and therefore could prove to be costly and fruitless pursuits. There lies the appeal of using testing methods with good frequentist properties. However, it is also possible to make the case for a Bayesian approach.

In some simultaneous-testing scenarios, the Bayesian approach to addressing multiplicity is almost as straightforward as the issue itself. You begin by modelling the "signal" observations with one distribution, F_1 and the "noise" observations with another, F_0. Define w as the unknown probability that an observation belongs to the signal distribution, and assume each observation y_i follows a mixture model:

$$y_i \sim w \cdot F_1 + (1-w) \cdot F_0 \tag{3.1}$$

Using Bayes' rule, you compute the posterior probabilities that each yi is from F_1. This Bayesian framework adjusts for multiplicity naturally. The posterior probabilities self-correct according to the data at hand, through the common mixing weight w.

https://doi.org/10.26219/heal.aueb.6544

3.1 Bayesian Approaches to Addressing Multiplicity

Multiplicity is a built-in issue in statistical inference that arises when multiple hypotheses are simultaneously tested. Frequentist techniques attempt to control the total error rate through adjustments such as the Bonferroni correction, but Bayesian techniques provide an alternative framework that automatically deals with multiplicity based on prior distributions and model probabilities. The Bayesian perspective of multiplicity includes two general topics: multiple hypothesis testing and multiple comparisons.

3.1.1 The Role of Prior Probabilities in Hypothesis Testing

One of the key aspects of Bayesian statistics is the application of prior probabilities, which are at the heart of hypothesis testing. Unlike frequentist methods that apply p-values and significance levels, Bayesian methods use prior beliefs and update these based on observed data to provide the probability that a hypothesis is true.

One of the earliest influential works on Bayesian hypothesis testing was by Harold Jeffreys (1939), who tested the hypothesis that a mean parameter θ in a normal model was zero or not zero.

Jeffreys' hypothesis testing for the mean θ under a normal model assumes:

- $H_0 : \theta = 0$
- $H_1 : \theta \neq 0$

In other words, he claimed that if θ was assumed to be nonzero, the corresponding prior distribution π(θ) needs to be symmetric about zero and scaled relative to the variance of the sampling model. His framework emphasized the importance of correctly setting prior probabilities so that several tests could be run at once.

3.1.2 Jeffreys' Approach to Correcting for Selection

Jeffreys also identified the problem of selection bias in multiple testing, whereby the chance of finding significant results rises merely because tests are being repeated. He termed this the problem of "correcting for selection." In a practical sense, when testing multiple alternative hypotheses, each has to be corrected for the fact that it was

https://doi.org/10.26219/heal.aueb.6544

selected from a wider set of possibilities. Jeffreys gave a mathematical correction for this, showing that if there are m competing alternative hypotheses, their respective probabilities must be scaled correspondingly.

Let $H_1^{(1)}, H_1^{(2)}, \ldots, H_1^{(m)}$ be competing alternatives. Then instead of assigning full prior weight $\pi(H_1)$ to each, Jeffreys' idea was:

$\pi(H_1^{(i)}) = \frac{\pi(H_1)}{m}$, for i = 1, ..., m

This ensures:

$$\sum_{i=1}^{m} \pi(H_1^{(i)}) = \pi(H_1)$$

So you spread the total prior for "some alternative" across the specific alternatives.

Later researchers developed on Jeffreys' work to establish more formal Bayesian solutions to multiple testing. Berry (1988) formulated the discrete mixture model, a procedure that broadens empirical Bayes methods by estimating prior probabilities from information instead of having them fixed at specific levels. This extension offers a more accommodating strategy for coping with multiple hypothesis testing.

Scott and Berger (2006) and Bogdan et al. (2008) extended the statistical properties of Bayesian decision rules in multiplicity. Concurrently, other nonparametric competitors such as those with Dirichlet process priors (Do et al., 2005) provided increased flexibility in parameter models under uncertainty.

In addition, Bayesian multiple testing procedures have been linked to false discovery rate (FDR) control, a common frequentist method introduced by Benjamini and Hochberg (1995). Studies by Muller et al. (2006) and Bogdan et al. (2008) highlighted how the Bayesian and frequentist solutions can be reconciled, particularly when controlling false discoveries in high-scale testing scenarios.

3.1.3 Bayesian Approaches to Multiple Comparisons

Multiple testing is interested in testing individual hypotheses, multiple comparisons is interested in making different groups of an experiment. This type of multiplicity is typical in ANOVA-type problems, in which interest is in determining if there are significant differences among several groups of samples.

https://doi.org/10.26219/heal.aueb.6544

One of the first Bayesian procedures for multiple comparisons was introduced by Duncan (1961, 1965), who gave a decision-theoretic procedure for comparing group means in a one-way ANOVA setting. His procedure used t-statistics that are used to compare differences between group means and expressed the decision procedure in terms of Bayes rules. Possibly the most potent weapon Bayesian statistics has to offer in the multiple comparisons problem is hierarchical models. These offer a formal approach to estimating group differences by pooling data across comparisons in a way that leads to better estimates.

Initial hierarchical Bayesian models, such as Tiao and Tan (1965), provided a formal framework for multiple comparison control with the bonus of protection from false positives. Later, Gopalan and Berry (1998) generalized this approach by incorporating Dirichlet process priors, which allowed more general grouping of groups with similar structure.

Beyond hierarchical structures, Bayesian approaches have also looked at clustering-based models to manage multiplicity. These models do not possess fixed group membership but instead derive clusters from observed data. Berry (1988) initially proposed a clustering-based Bayesian model, which was subsequently extended by researchers such as Hartigan (1990) and Crowley (1997), who constructed product-partition models that regulate the number of comparisons in terms of a chosen cohesion function.

An alternative is Bayesian model selection, in which models for potential group differences are compared probabilistically and models differ from each other. Scott and Berger (2010) explained how Bayesian model selection would naturally correct for multiplicity without the need for explicit p-value adjustments since prior model probabilities automatically correct for overfitting.

3.1.4 A Comparative Overview

It is useful to contrast Bayesian multiplicity adjustment with frequentist methods in an attempt to better understand Bayesian multiplicity control. Standard techniques like the Bonferroni correction reduce multiple comparison threats by adjusting significance levels, while Bayesian procedures treat the problem differently through altered prior distributions and model probabilities.There have been two differing perspectives on Bayesian multiplicity control:

 https://doi.org/10.26219/heal.aueb.6544

- Jeffreys' Method (1961): Concentrates on prior probability corrections based on the number of rival hypotheses.
- Bayesian Hierarchical Models: Utilize shrinkage estimators to pull estimates toward the group mean, reducing variance and stabilizing results.

A few studies have compared the relationship between Bayesian and frequentist approaches:

- Westfall et al. (1997): Showed that in some cases, Bonferroni-adjusted p-values are approximately equal to Bayesian inference.
- Meng and Dempster (1987): Asserted that implicit post-hoc corrections may be circumvented in Bayesian models as prior probabilities contain the multiplicity adjustment.
- Storey (2003) and Efron et al. (2001): Provided Bayesian methods for FDR control.

3.2 Strategies for Managing Multiplicity in Statistical Inference

Multiplicity is a basic problem in statistical inference, where multiple hypotheses are to be tested in the same data. Different statistical schools of thought treat the issue differently: frequentist approaches employ direct rate of error adjustment, while Bayesian approaches employ multiplicity in the form of prior distributions. This section addresses prominent techniques used to treat multiplicity from both perspectives.

3.2.1 Frequentist Approaches to Multiplicity Control

One of the significant concerns in various hypothesis testing is retaining control on at least one Type I error (false negative) probability. Family-Wise Error Rate (FWER), probability of at least one false rejection of a null hypothesis, is extensively used as a measure of false-positive risk (Hochberg & Tamhane, 1987).

One of the most commonly used FWER control methods is the Bonferroni correction (Bonferroni, 1936), which reduces the significance level by dividing it by the number of tests conducted. Although this method works well to eliminate false positives, it

https://doi.org/10.26219/heal.aueb.6544

tends to be too conservative and may result in an increased number of Type II errors (false negatives) because of its stringent requirements.

To reduce the conservativeness of FWER control, another approach known as False Discovery Rate (FDR) was proposed by Benjamini and Hochberg (1995). Rather than emphasizing on avoiding entirely false positives, FDR controls the ratio of rejected null hypotheses that are false positives.

$$FDR = E\left[\frac{R}{V}\right] \text{ where V=Number of false positives, R=Total rejections}$$

This approach is particularly helpful in high-dimensional data analysis (e.g., genomics, neuroimaging), where it is acceptable to make some errors as long as they are proportionally bounded. FWER and FDR corrections both have the consequence of constraining incorrect statistical conclusions, but FDR is generally preferred in large-scale hypothesis testing.

3.2.2 Bayesian Perspective on Multiplicity Adjustment

Unlike frequentist approaches, which make direct corrections after the tests have been conducted, Bayesian ones incorporate multiplicity adjustment into the model directly. This is done by employing prior beliefs about hypothesis probabilities and updating them with observed data.

A natural way in which Bayesian methods are responsive to multiplicity is through hierarchical models, which have an organized structure for collapsing hypotheses into groups. Such models use shrinkage estimators, which pull distinct estimates toward a shared mean (Gelman, Hill, & Yajima, 2012). This helps to stabilize statistical inference over a collection of tests.

Nevertheless, hierarchical models do not behave like true tests of hypotheses. Their behavior depends on sample size and data variability (Gelman & Loken, 2013) such that they perform optimally for effect-size estimation purposes but are less adapted to binary accept/reject style decisions.

Another Bayesian technique involves dynamically adjusting prior probabilities based on the number of hypotheses to test. Instead of assigning fixed prior probabilities for individual hypotheses, it considers the larger chance of false positives if many tests are being conducted.

 https://doi.org/10.26219/heal.aueb.6544

For example, in model choice problems, there is a prior distribution that can be put on the probability that the parameter is in a model. This is a technique typically modeled as a Bernoulli process that helps in control of variables input while making sure that statistical requirements are upheld.

$$\gamma_j \sim Bernoulli(\pi)$$

Where:

- γ_j=1 means variable j is included
- π is the prior inclusion probability

A technique referred to as null control directly ensures the total probability assigned to the null hypothesis across all comparisons will remain the same (for example, 0.5). It is a correction, as explained by Williams, Heathcote, Nesbitt, and Eidels (2016), that works in a manner akin to Bonferroni-type frequentist adjustments, where significance thresholds are dynamically tweaked based on the number of comparisons undertaken.

Bayesian null control is also analogous to frequentist post-hoc adjustments. Westfall, Johnson, and Utts (1997) demonstrated that, under certain conditions, Bonferroni-adjusted p-values and Bayesian posterior probabilities scale proportionally.

3.2.3 Bayesian Techniques for Estimating the Proportion of True Effects

In contrast to examining single hypotheses separately, Bayesian approaches can measure the overall ratio of true effects in a data set. It assumes that signal-to-noise ratios come from a structured distribution rather than treating hypotheses as if they are independent. For instance, Stephens and Balding (2009) applied this method in genetic research, defining prior probabilities from previously observed effects. This is comparable to False Discovery Rate (FDR) control since it estimates how many null hypotheses rather than correcting single comparisons.

A higher-order generalization of the single proportion model is to place a distribution over the proportion of true hypotheses. This renders hypotheses as dependent, rather than independent, and this has the effect of reducing overconfidence in false discoveries.

This procedure has been utilized in:

- Feature selection for regression models (Scott & Berger, 2006, 2010)

https://doi.org/10.26219/heal.aueb.6544

- Statistical inference from graphs (Bogdan, Ghosh, & Tokdar, 2008)
- Genetic association studies (Carvalho & Scott, 2009, Li & Shang, 2015)

One of the advantages of this method is that posterior probabilities respond dynamically as more hypotheses are added, and it is a robust option for large-scale hypothesis testing.

	Declared.non.true.signal	Declared.true.signal	Total
True signal	U	V	m_0
Non-true signal	T	S	$m - m_0$
Total	$m - R$	R	m

Table 3.1: Summary of possible outcomes in multiple hypothesis testing. Of the m total hypotheses, let m_0 represent the number of true null hypotheses, and $m-m_0$ the number of actual alternative hypotheses. Type I errors (false positives) are denoted by S, and Type II errors (false negatives) by U.

3.3 Null Control in Bayesian Multiplicity Adjustment

The concept of null control is central in Bayesian multiplicity adjustments to ensure that the prior probability of no effect observation as a whole remains constant under numerous hypothesis tests. This is particularly important when dealing with cases involving numerous comparisons since the accumulation of noise variables biases statistical inference.

To observe this, consider the following example in which we perform k = 6 pairwise comparisons. If we conduct k = 6 independent tests and assign a fixed prior probability π_0 to the null hypothesis H_0 for each test, the overall probability that all k null hypotheses are true is:

$$\prod_{i=1}^{k} P(H_{0,i}) = \pi_0^k$$

For example, if $\pi_0 = 0.5$ and k = 6, this results in:

$$P(Global\,Null) = 0.56 = 0.015625$$

demonstrating that even moderate values of k rapidly diminish the global null probability under uniform priors.

https://doi.org/10.26219/heal.aueb.6544

In conclusion, if we place a uniform prior probability on each null hypothesis, the prior belief that there are no effects is lower than the overall probability of observing no true differences. The more noise variables we add to the scenario, the lower the probability falls, and thus it is apparent that a more systematic approach to dealing with prior model probabilities is required.

Several approaches have been proposed for null control, and two broad approaches are:

- Jeffreys' Method (1938), which employs a uniform probability framework but with independence assumption between tests.

$$P(H_{0,i})=\pi_0 \text{ for all i, and } P(H_{0,i})=\pi_0^k \text{ for } i=1,\dots,k$$

- Westfall's Method (1997), which involves dependence among comparisons in order to calibrate previous probability estimation. Compared to methods that assume independence between tests, Westfall's method reduces prior probabilities for each hypothesis proportionally with the number of tests. The aim is to keep the combined prior probability of no true effects constant across scenarios. As additional comparisons are included, prior belief assigned to each single null must be reduced to not have the total probability of no effect falling too sharply.

3.3.1 Jeffreys' Method for Null Control: A Conservative Approach

One of the earliest null control procedures was described by Harold Jeffreys (1938), who introduced a simple method grounded in independent comparisons. The basic idea is that the probability of a null hypothesis being true needs to be proportional to the number of alternative hypotheses under consideration

Jeffreys' method assigns an equal prior probability to each alternative hypothesis and its complement (the null hypothesis). This can be expressed as:

$$p(H_0)+p(H_1)=1$$

where H_0 represents the null hypothesis, and H_1 represents the alternative. Given k alternative hypotheses, the probability that all alternative hypotheses are false (all comparisons are truly null) is calculated as:

$$P(all\, H_1\, are\, false)=(1-p)^k$$

https://doi.org/10.26219/heal.aueb.6544

where p is the probability of an alternative hypothesis being true. Making this equal to pre-specified overall prior probability of no effect allows the equation to be rearranged to obtain proper prior probability allocation:

$$p = 1 - (P(H_0))^{1/k}$$

This ensures that the probability assigned to each alternative hypothesis remains consistent with the expected number of true nulls.

While Jeffreys' method provides an easy answer to multiplicity control, it has one significant flaw: it assumes that all comparisons are independent. In reality, there are typically dependencies between hypotheses, particularly in experimental research where multiple comparisons are influenced by shared factors. This makes the method excessively conservative, with too much loss in power and too much risk of failing to detect real effects.

3.3.2 Westfall's Dependency-Based Null Control Method

This chapter originates from the development of the dependency-based null control method presented by Tim de Jong (2020), extending Jeffreys approach by modeling group means that are interdependent in the sense of a common hierarchical structure.

Observing the limitations of Jeffreys' approach, Westfall et al. (1997) proposed an extension which makes formal provision for interdependence of comparisons. Their method increases the accuracy of prior probability assignment by shifting focus from individual pairwise comparisons to the group pattern beneath them.

Westfall's model assumes that within an experimental design, the mean values of different groups are drawn from a grand mean underlying μ. The model holds two possibilities for each group mean μ_i

- It is equal to the grand mean μ with probability τ.
- It is drawn from a continuous distribution G with probability $1-\tau$

This can be generalized to m groups, so the probability that group means all are equal can be written as:

$$P(all\ \mu_i = \mu) = \tau^m$$

 https://doi.org/10.26219/heal.aueb.6544

By solving for τ, we obtain an expression that incorporates dependencies between group comparisons.

To show the diffirences, we consider m=4 groups, resulting in k=6 pairwise comparisons. Using $p(H_0)=0.5$ Jeffrey's method yields $p(H_{0i})=0.5^{\frac{1}{6}}=0.891$ while Westfall's method gives $p(H_{0i})=0.5^{\frac{2}{4}}=0.707$. When m=2 and k=1, both methods default to $p(H_{0i})=0.5$.

Feature	Jeffreys.Method	Westfalls.Method
Focus	Individual comparisons (k)	Group means (m)
Dependency Handling	Assumes independence of tests	Incorporates dependencies via τ
Adjustment	More conservative	Less conservative, higher power
Example (m = 4, k = 6)	Null prob: ≈ 0.89	Null prob: ≈ 0.71
Applicability	Simpler, but less realistic	Better suited for correlated data

Table 3.2: Comparison of Jeffreys' and Westfall's Methods for Null Control in Bayesian Hypothesis Testing

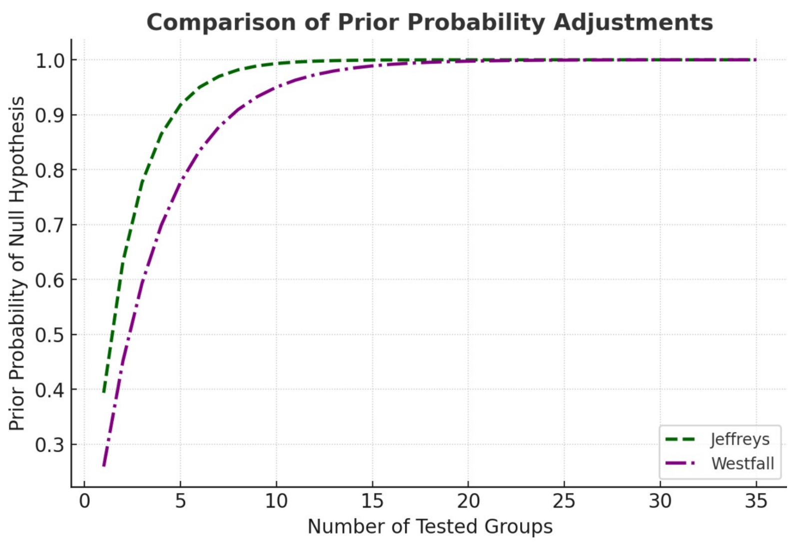

Figure 3.1: Single Null Hypothesis Prior Under Null Control Assuming 0.5 Total Prior for No Effects

https://doi.org/10.26219/heal.aueb.6544

While Jeffreys' method is simple and conservative in null control, it will tend to underestimate the probability of true effects due to its independence assumption. Westfall's method, however, is more realistic and flexible since it considers that multiple comparisons often share common sources of variation.

3.4 Priors Adjusting for Multiplicity

3.4.1 Dirichlet Process Prior

The Dirichlet Process (DP) was introduced by Ferguson (1973). It provides flexible, infinite-dimensional modelling, in the sense that the number of groups (or clusters) is not required to be specified beforehand. Rather, the model discovers the number of clusters from the data. A distribution G follows a Dirichlet Process with concentration parameter α and base distribution G_0, denoted as:

$$\text{DP} \ (G \sim DP(a, K))$$

Here:

- G_0 is the base distribution, the prior on the parameter distribution.
- α is the concentration parameter, which is accountable for the tendency of new clusters to form.

The Dirichlet Process is often applied in Bayesian nonparametric models, and more specifically in clustering problems, where it can serve as a versatile alternative to models that have a fixed number of clusters.

When introducing a new observation θ_{j+1}, we have two possibilities:

1. It forms a new cluster with probability:

$$P\left(\theta_{j+1} \text{ is new}\right) = \frac{\alpha}{\alpha + j}$$

2. It joins a pre-existing cluster with probability proportional to the number of times the cluster has been observed:

$$P\left(\theta_{j+1} = \theta_i\right) = \frac{n_i}{\alpha + j}$$

where n_i is the number of times the value θ_i^* has been observed.

The Dirichlet Process controls overfitting and multiple testing by varying the number of clusters automatically based on the data. Compared to the Uniform Prior

https://doi.org/10.26219/heal.aueb.6544

over Partitions, which is indifferent to all possible partitions, the Dirichlet Process favors fewer clusters but with some degree of flexibility. The prior over partitions is:

$$\pi(\rho|\alpha)=\frac{\alpha^{|\rho|}\cdot\Gamma(\alpha)}{\Gamma(n+\alpha)}\cdot\prod_{c\in\rho}\Gamma(|c|)$$

where:

- $|\rho|$ is the number of clusters in the partition.
- $|c|$ is the size of each cluster.

Although the base distribution of the Dirichlet process is infinite-dimensional, the prior over partitions can be integrated out. This implies nonparametric modelling - the number of parameters is not fixed - a parametric formulation for the partitions, i.e., the number of partitions is fixed. For this reason, this kind of approach is sometimes referred to as a product partition model in the literature, e.g., Quintana, 2006, Quintana & Iglesias, 2003. This makes the model appropriate for our application, since we are working with a fixed number of parameters.

3.4.2 Beta-binomial Prior

The beta-binomial model prior is widely used in stochastic search variable selection for linear regression (George & McCulloch, 1993) and Bayesian model averaging (e.g., Hinne et al., 2020; Hoeting et al., 1999). Given K total elements (e.g., predictors, clusters, or partitions), the probability of selecting j of them follows a Beta-Binomial distribution:

$$BB(j|K,\alpha,\beta)=\binom{K}{j}\frac{B(j+\alpha,K-j+\beta)}{B(\alpha,\beta)}$$

where:

- α and β are hyperparameters controlling the prior shape.
- $B(a,b)$ is the Beta function, ensuring normalization.

This prior naturally penalizes complexity. This means that it discourages selecting too many elements when applied to variable selection or forming too many partitions in clustering problems.

In a partitioning scheme, we can interpret the Beta-Binomial prior as a probabilistic regularization of the count of groups. Instead of every partition being uniformly

https://doi.org/10.26219/heal.aueb.6544

probable (the uniform prior), the Beta-Binomial prior has a structured way of favoring sensible quantities of partitions.

- The prior shrinks estimates toward a moderate number of partitions.
- Larger values of α and β enforce stronger regularization, reducing unnecessary partitions.
- Smaller values of α and β allow more flexibility, but still control excessive partitioning.

For example, if we are partitioning K elements, the probability of having i partitions follows:

$$BB(i \mid K-1, \alpha, \beta)$$

where the total number of partitions is linked to the number of imposed inequalities (constraints on the groups).

The Beta-Binomial prior naturally adjusts for multiplicity in the following ways: Penalizing excessive complexity: Adding more partitions (or predictors in regression) is discouraged unless strongly supported by data.

- Flexibly modelling uncertainty: By choosing appropriate values of α and β we can control the likelihood of adding more partitions.
- Avoiding overfitting: The Beta-Binomial prior favors simpler models unless the data suggests otherwise.

For instance:

- Setting $\alpha = \beta = 1$ leads to a U-shaped distribution, giving more probability to extreme cases (very few or very many partitions).
- Setting $\alpha = 1$ and $\beta = \lambda K$ ensures a monotonic decrease, discouraging excessive partitions.

In summary, this prior on the beta-binomial distribution regulates the number of inequalities and hence the control over model complexity and interpretability in partition selection.

3.4.3 Uniform Prior

Lastly, we also consider a uniform prior over partitions. This prior assumes that all partitions are likely with equal probability. Unlike other priors, i.e., the Dirichlet Process Prior that favors a structured clustering process, or the Beta-Binomial Prior

https://doi.org/10.26219/heal.aueb.6544

that penalizes over-complexity, the Uniform Prior does not favor any specific grouping of elements. If we have a total of K elements, each valid partition has an equal probability:

$$P(\rho)=\frac{1}{B_K}$$

where B_K is the total number of possible partitions of a set of size K (related to Bell numbers). This means that every way of grouping elements is treated equally, regardless of the number or size of the groups. When a new element θ_{j+1} is added to an existing partition, it can either:

1. Form a new group with probability:

$$P_{\pi U}=\frac{B_{K-j+1,r+1}}{B_{K-j+1,r+1}+rB_{K-j+1,r}}$$

2. Join an existing group with probability proportional to the number of ways that can happen:

$$P\left(\theta_{j+1}=\theta_{\iota}\right)=\frac{rB_{K-j+1,r}}{B_{K-j+1,r+1}+rB_{K-j+1,r}}$$

where:

- $B_{K-j+1,r+1}$ counts the number of ways to place θ_{j+1} into a new subset.
- $B_{K-j+1,r}$ counts the number of ways to place θ_{j+1} into one of the existing r subsets, multiplied by r since there are r available choices.

The Uniform Prior treats all partitions as equally probable, without preference for smaller or larger clusters and does not adjust for multiplicity in the same way as other priors. Since it assigns equal weight to all partitions, it does not penalize unnecessary complexity or favor simpler models. This can lead to overfitting in cases where too many partitions are created.

3.5 Hierarchical Modeling of Prior Inclusion Probabilities via Beta-Bernoulli Distributions

Scott and Berger (2006, 2010) proposed a hierarchical Bayesian approach to accommodate model uncertainty and correct for multiplicity. Although their method

https://doi.org/10.26219/heal.aueb.6544

was originally developed for variable selection and regression, it can be adapted to pairwise contrasts.

For this model, every one of the m parameters (or variables) is added or excluded from a model, which results in a model space of size 2m. Equal prior probability given to all models implies that every parameter has a 50% opportunity of being included, which does not correct for multiple testing and can overestimate false discoveries.

To address this, Scott and Berger introduced a hierarchical prior. In this the inclusion of each parameter is treated as a Bernoulli trial with inclusion probability q. Instead of fixing q, they assume it follows a Beta distribution, typically Beta(1,b). This formulation allows flexibility. Higher values of b shows a stronger preference for smaller, simpler models (fewer included variables), while b=1 corresponds to a uniform prior. The overall prior probability of a model is then derived by integrating over q, producing probabilities that depend on the number of included parameters in each model.

This method has the benefit of having model probabilities uniformly distributed over model sizes, rather than over specific models. That is, all models with the same number of variables included are equally likely, but more complex models are penalized unless the data strongly favor them. This has multiplicity control built in by favoring simpler models unless complexity evidence is strong.

3.5.1 Extending the SB Framework

Scott and Berger's approach to variablinclusion into regression can be modified for dealing with pairwise comparisons. Instead of modeling individual regression coefficients' inclusion, we focus on differences or no differences among group means. This requires a change in the number of models: rather than 2^m models for m variables, we look at the $k = \binom{2}{m}$ potential pairwise comparisons among m groups. The prior model probabilities are then distributed over configurations having varying numbers of non-zero differences. For instance, with m=3, we have k=3 comparisons Granting independence among comparisons, the prior probability is spread evenly over all possible configurations (e.g., 0, 1, 2, or 3 differences).

https://doi.org/10.26219/heal.aueb.6544

This independent-treatment assumption is false because comparisons are not actually independent. Whether the groups A and B differ or not can influence what is plausible between groups B and C. With increasing noise, the actual differences may increase overall, but some are structurally related. In order to provide logical coherence, we model group differences as a network where groups are nodes and comparisons are edges. With m=3 for example three groups can form a triangle of differences but not all permutations when comparisons are interdependent. This leads us to distribute the previous probabilities only over model classes that are possible, and not over configurations that do not satisfy logical constraints.

Shaffer (1986) proposed a systematic method to identify valid hypothesis structures under dependence by dividing all possible equivalence relations between groups. This allows us to specify sets of null hypotheses that are coherent under rejection of others. Using this approach, we recursively identify feasible hypothesis subsets. For example, for m=3, that five distinct null configurations are achieved, rather than the eight configurations we'd have with complete independence. That figure is the Bell number for m=3. Having narrowed down the valid hypothesis space, we assign prior probabilities across permissible subsets through a weighted formula, in proportion to the number of valid configurations in each class. The probability given to a particular model is in proportion to its frequency in the set of possible models. With, for instance, m=4, there are 15 possible hypothesis models (the Bell number for m=4). A four comparisons active model between six possible would be given a probability of $\frac{6}{90}=\frac{1}{15}$, since there are six models in the class. This restricted distribution provides a closer approximation to prior belief with consistency to logical group relationships.

3.6 Bayesian ANOVA and Marginal Likelihood for Group Comparisons

So far, we have been guided by model probabilities in the past in what we do with multiplicity, but here we shift to marginal likelihoods. While the t-test is okay for comparing two groups, it is not extensible to multiple group comparisons. For instance, when making inferences about models that have three or more groups, we need an alternative framework.

 https://doi.org/10.26219/heal.aueb.6544

To address this, we adopt the Bayesian one-way ANOVA approach of Rouder et al. (2012), generalizing the argument of the t-test to multiple groups. In the important special case of comparing only two groups, the approach provides the same result as the standard Bayesian t-test.

One issue with the Bayesian ANOVA model is that there is an identifiability problem, since the number of parameters at the group level exceeds the number of groups. This was resolved by Rouder et al. (2012) by placing a sum-to-zero constraint on the group-level effects, which renders the model identifiable.

$$\sum \omega_i = 0 .$$

Similar to the t-test, the model terms ω_i in the Bayesian ANOVA are reparameterized as effect sizes $d_i = \frac{\omega_i}{\sigma}$. Common parameters across all models are assigned the same uninformative prior:

$$\pi\left(\mu, \sigma^2\right) \propto \frac{1}{\sigma^2}$$

For models that include the parameter d, the specification mirrors that of the t-test:

$$d_i \sim Normal\left(0, \sigma_d^2\right)$$

which simplifies to a Cauchy distribution after marginalizing out the variance:

$$d_i \sim Cauchy$$

Although Bayesian ANOVA is useful, it does not provide direct inference regarding individual pairwise differences (δ). It only checks if group means are equal in general. Moreover, placing a sum-to-zero constraint on δ's raises doubts about their interpretability.

A more practical alternative is the idea of relabeling. For example, if we are comparing a difference between three groups and we assume that groups 1 and 2 are equal, then the two groups can be relabeled into one. The remaining comparison (e.g., between the combined group and group 3) can be compared by a t-test or simple ANOVA.

This method allows testing of intermediate models, where just some groups are considered equal. Rouder et al. (2012) agree with this argument, and they further say that researchers typically have a specific contrast in mind, not the whole model.

By comparing evidence from all the intermediate models and the full model, we can better understand group-level relationship structure. This relabeling method has

https://doi.org/10.26219/heal.aueb.6544

already been applied in other Bayesian analyses of interest for pairwise group comparisons (e.g., Gopalan & Berry, 1998; Neath & Cavanaugh, 2006).

3.9 Conclusions

This chapter discussed the problem of multiplicity adjustment in statistical inference, with a specific emphasis on Bayesian methods. The problem of multiple testing and multiple comparisons occurs in most scientific applications and may heighten the risk of false discoveries unless properly addressed. Traditional frequentist methods, such as Bonferroni adjustment and False Discovery Rate (FDR) adjustment, offer solutions for this issue but are too conservative or rigid in more complex scenarios. Bayesian methods provide a natural resolution of multiplicity using prior knowledge and dynamically updated model probabilities. We described how hierarchical models provide a natural resolution of multiple comparisons using shrinkage estimation, pulling estimates toward group means. In addition, objective adjustment of prior model probabilities offers a systematic approach to managing multiplicity without loss of statistical power.

Lastly, we reviewed Bayesian techniques, including: Null control methods, such as Jeffreys' and Westfall's, which re-tune earlier model probabilities to have a reasonable balance between true discovery and false positives. The Scott & Berger (SB) procedure, which determines the proportion of true effect by a hierarchical prior framework, making it particularly useful in regression and high-dimensional hypothesis testing. Pairwise group comparisons in Bayesian inference, with emphasis on the use of network-based dependency models and Dirichlet process priors to manage multiplicity in group comparison pairs. We also investigated applications of these methods to real-world problems, including Bayesian ANOVA and hierarchical partitioning of hypotheses, demonstrating how Bayesian methods enhance model selection and inference under multiple testing. Simulation experiments confirmed that Bayesian methods offer a flexible and computationally efficient alternative to standard correction methods without compromising interpretability and robustness.

In the next chapter, we will illustrate the application of Bayesian multiplicity adjustment in real experimental designs. We will discuss case studies and simulation-

https://doi.org/10.26219/heal.aueb.6544

based analyses to show how Bayesian ANOVA and model selection procedures can be used in practice.

https://doi.org/10.26219/heal.aueb.6544

CHAPTER 4

Bayesian Analysis of Model Parameters Using Bayes Factors

4.1 Data Description

The data for this analysis was from the research paper "Cardiac Output and Performance during a Marathon Race in Middle-Aged Recreational Runners" by Billat et al. (2012). The paper focused on examining cardiac performance measures as determinants of marathon performance in recreational athletes, especially cardiac endurance and cardiac cost.

The study included 14 middle-aged male recreational runners, all experienced in endurance sports. The mean characteristics of the participants are as follows: age: 37±6 years, weight: 71±8 kg, and height:178±6 cm. Each participant was subjected to an incremental treadmill test before the marathon to determine their maximal physiological values, including maximal oxygen uptake (VO_2max), heart rate (HRmax), stroke volume (SVmax), and cardiac output (COmax). Data collection during the marathon race were made with continuous monitoring of heart rate, stroke volume, and cardiac output via impedance cardiography.

Core question for this thesis is: Analyse if there is any correlation between performance at the marathon and an individual's ability to maintain a high fractional utilization of maximal stroke volume and cardiac output. performance. The data set included both laboratory and real race measurements, so that cardiac dynamics can be compared in controlled and competitive settings. The main performance measure was the average running speed during the marathon (v_{Mar}), expressed as a fraction of maximal stroke volume speed (vSV_{max}) and maximal oxygen uptake speed ($vVO2_{max}$). Additionally, changes in cardiac parameters such as heart rate per meter run (HRS), stroke volume per meter run (SVS), and cardiac output per meter run (COS) were analyzed over successive segments of the marathon.

https://doi.org/10.26219/heal.aueb.6544

	Characteristics		Values during the incremental test					
Runner	**Age**	**BMI**	**SV_max**	**HR_max**	**CO_max**	**vLT**	**vSV_max**	**vPeak**
1	27	18.3	147	189	25.6	12	11	15
2	30	24	180	191	26.1	10	11	14
3	39	24.2	182	200	26.7	11	12	16
4	36	23.3	174	192	28	14	15	17
5	39	21.6	197	189	33.7	15	14	18
6	36	22.5	179	199	29.9	12	11	16
7	34	20.8	163	187	25.5	15	13	17
8	37	23.2	198	190	35.4	13	13	16
9	45	22.1	185	179	23.3	12	13	16
10	47	21	166	190	27.4	15	13	19
11	33	23.4	175	181	29.5	12	15	16
12	36	23.6	180	187	22.6	12	12	16
13	35	24.6	180	186	27.4	11	12	16
14	46	20.8	163	180	24.3	12	12	14
Mean	**37.1**	**22.4**	**176.4**	**188.6**	**27.5**	**12.6**	**12.6**	**16.1**
SD	*5.8*	*1.7*	*13.5*	*6.2*	*3.6*	*1.6*	*1.3*	*1.4*

Table 4.1: Characteristics and Incremental Test Performance Metrics of Marathon Runners

	Performance Metrics			
Runner	**T_mar (hrs)**	**vMar (km/h)**	**vMar% vSV_max**	**vMar% vVO2_max**
1	3.51	11	100	77.3
2	4.58	8.5	77.3	60.7
3	4.08	10.2	85	72.9
4	3.04	13.8	92	81.2
5	2.52	14.7	105	81.7
6	3.51	11	100	73.3
7	2.58	14.2	109.2	83.5
8	3.05	13.7	105.4	85.6
9	3.39	11.6	89.2	72.5
10	2.57	14.3	110	79.4
11	3.45	11.2	74.4	70
12	5.1	8	66.7	53.3
13	4.02	10.5	87.5	70
14	3.54	10.8	90	77.1
Mean	**3.5**	**11.68**	**92.26**	**74.18**
SD	**0.76**	**2.15**	**13.42**	**8.89**

Table 4.2: Performance Metrics of Runners during the Marathon Race

https://doi.org/10.26219/heal.aueb.6544

After data collection, athletes were divided into groups based on their physiological responses. More precisely, they were divided into those with high SV and low SV, and further into high VO2max and low VO2max. Such categorization enables the comparison of how these different physiological profiles affect marathon performance and endurance capacity.

4.1.1 Explanation of Key Metrics

Endurance performance in marathon running is influenced by running efficiency and physiological capacity. To understand how a runner sustains a particular speed over 42.195km, one has to observe not only his/her speed but also the response of his/her cardiovascular system to the prolonged effort. By monitoring some of the key physiological markers, we can assess how effectively runners manage aerobic energy production, cardiac efficiency, and physiological strain during the race.

This chapter introduces the major parameters of running velocity and cardiovascular function, providing insight into the relationship between cardiac function and endurance performance during marathon conditions.

The average running speed is the mean speed kept by each one of the participants in the 42.195 km marathon. It is one of the major measures of performance and corresponds to the ability to maintain a constant pace in view of long-lasting physical work.

To contextualize v_{Mar} within the physiological capabilities of the runners, it is expressed as a fraction of two key speeds derived from the pre-race laboratory tests:

1. Maximal Stroke Volume Speed (vSVmax): This is the speed at which the stroke volume (SV) reaches its maximum during an incremental treadmill test. VSVmax provides insight into the runner's capacity to optimize cardiac output at a specific workload.
2. Maximal Oxygen Uptake Speed (vVO2max): This is the speed corresponding to the maximal oxygen consumption (VO2) during the treadmill test. It reflects the upper limit of aerobic energy production and is a fundamental indicator of endurance performance.

By expressing vMar as a percentage of vSVmax and vVO2maxv the analysis highlights how well each participant utilized their maximal physiological potential during the marathon.

https://doi.org/10.26219/heal.aueb.6544

Three cardiac parameters were evaluated as indicators of performance and physiological strain during the marathon:

1. Heart Rate Per Meter Run (HRS): This metric represents the number of heartbeats required to cover one meter of distance. It is calculated by normalizing the heart rate (HR) to the running speed. A lower HRS value suggests greater cardiac efficiency.

2. Stroke Volume Per Meter Run (SVS): This measures the volume of blood pumped by the heart in one beat, normalized to the distance run. It reflects the heart's capacity to deliver oxygen and nutrients per unit of distance.

3. Cardiac Output Per Meter Run (COS): This parameter combines stroke volume and heart rate to quantify the total blood flow supplied by the heart per unit of distance. COS serves as an indicator of the "cardiac cost" of maintaining a specific pace and can reveal the efficiency of the cardiovascular system under sustained exercise.

These parameters were monitored and analyzed over successive segments of the marathon, allowing for a detailed assessment of changes in cardiac dynamics and their relationship to performance.

4.2 Analysis of Marathon Performance Using ANOVA

A one-way ANOVA, has been done to calculate the F statistic and the accompanying p-value of the dataset. Marathon time (Tmar) measurements have been analyzed. Specifically, the dataset was analyzed using two different models:

- Marathon Time as a function of maximal stroke volume speed (vSVmax).
- Marathon Time as a function of maximal oxygen uptake speed (vVO2max).

One-way ANOVA assumes the following statistical model:

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

In each model:

- y_{ij} represents the marathon time for the i-th participant in the j-th group (e.g., groups based on stroke volume or oxygen uptake speed).
- μ_j is the mean marathon time for batch j.

https://doi.org/10.26219/heal.aueb.6544

- τ_j are fixed contrasts (deviations from the overall mean), constrained such that their sum equals zero.
- E_{ij} are random, independent, and identically distributed deviations from the gtoup mean ($N\left(0,\sigma^2_{within}\right)$).

4.2.1 Frequentist approach

One-way ANOVA is a statistical method that can be used to investigate the impact of (k>2) levels of a single factor on measured observations. The hypotheses that can be tested for, in order to check whether the levels of the factor significantly affect the values observed, are as follows:

$$\text{Ho: } \mu_1=\mu_2=\ldots=\mu_j$$

Versus

$$\text{H1: } \mu_1\neq\mu_2\neq\ldots\neq\mu_j$$

To compare the means of J populations, we consider J groups, each with n_j observations. The model is expressed as:

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

where j=1,…,J (groups or samples) and i=1,2,…,n_j.

For one-way ANOVA to yield valid results, the following assumptions must be satisfied:

- The population at each factor level is (approximately) normally distributed.
- Observations are obtained independently from the populations defined by the factor levels.
- All populations have the same variance.
- Random error terms (ε_{ij}) are independent and identically distributed (iid).
- $\varepsilon_{ij}\sim N(0,\sigma^2)$

Under these assumptions, for a factor level i, the population distribution is $N(\mu_i,\sigma^2)$.

The hypothesis H_0 is tested by computing the F-statistic:

$$F=\frac{MS_{between}}{MS_{within}}F\left(J-1,N-J\right)$$

https://doi.org/10.26219/heal.aueb.6544

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.

4.2.2 Performance in Marathon Completion from One-Way ANOVA Analysis

To investigate the physiological determinants of marathon performance, we analyzed the effects of stroke volume and maximal oxygen uptake speed on marathon completion time using one-way ANOVA. The results revealed significant relationships between these cardiovascular factors and endurance performance.

The effect of Stroke Volume was significant ($F(2,10)=5.998$, $p=0.019$). This finding demonstrates that the ability to sustain a high proportion of maximal stroke volume speed during the marathon is a critical factor influencing Marathon Time. Runners in the higher maximum Stroke volume achieved faster completion times, suggesting a direct relationship between cardiovascular efficiency and marathon performance. Stroke volume, as a major determinant of cardiac output, represents the ability of the heart to pump oxygenated blood in order to satisfy the metabolic requirements during prolonged exercise. The highly significant effect of stroke volume emphasizes the need to optimize stroke volume utilization during endurance events and reiterates the role of cardiovascular changes for improved performance.

Lastly the analysis of maximal oxygen uptake speed revealed a highly significant effect on Marathon Time ($F(2,10)=16.09$, $p<0.001$). This result shows the importance of aerobic capacity in endurance performance. Runners who sustained higher percentages of their maximal oxygen uptake speed during the marathon completed the race significantly faster. Maximal oxygen uptake is a fundamental determinant of endurance performance. It reflecting the body's ability to utilize oxygen during aerobic metabolism. The strong significance of Oxygen uptake demonstrates that maintaining a high fraction of aerobic capacity is essential for optimizing marathon performance. These findings align with previous research emphasizing the critical role of aerobic power and metabolic efficiency in determining endurance running success.

https://doi.org/10.26219/heal.aueb.6544

Description of Model	F-Statistic	P-Value
Effect of Maximal Stroke Volume on Marathon Time	5.998	1.94e-02
Effect of Maximal Oxygen Uptake on Marathon Time	16.090	7.48e-04

Table 4.3: Summary of One-Way ANOVA Results for Performance Metrics and Marathon Completion Time

4.2.3 A Frequentist Two-Way ANOVA

Next, two factors were examined for independent and interactive influences on marathon completion time: Stroke Volume and Oxygen Uptake. This two-way ANOVA approach allows for the evaluation of both main effects and their interaction and offers information into the physiological contributions to endurance performance.

The hypotheses for two-way ANOVA can be tested as follows:

1. Main Effect of Factor A:
    - $H_0 : \mu_{A1} = \mu_{A2} = \ldots = \mu_{Ak_1}$
    - H_1 : At least one μ_{Ai} is different
2. Main Effect of Factor B:
    - $H0 : \mu_{B1} = \mu_{B2} = \cdots = \mu_{Bk2}$
    - H_1 : At least one μ_{Bi} is different
3. Interaction Effect Between Factors A and B:
    - H_0: The effect of Factor A is the same at all levels of Factor B.
    - H_1: The effect of Factor A depends on the level of Factor B.

The two-way ANOVA allows significant insight into how Stroke Volume and Oxygen Uptake together affect completion time in marathon running. The main effects were both significant: stroke volume significantly impacted marathon time, indicating that the differences across the categories of stroke volume contribute much to the variation in performance. Secondly, Oxygen Uptake has been found to

https://doi.org/10.26219/heal.aueb.6544

significantly affect marathon time, underlining once again the contribution of aerobic capacity to endurance performance.

Effect	Df	Sum_Sq	Mean_Sq	F_value	Pr_F
Stroke Volume	2	3.94	1.97	11.15	0.01
Oxygen Uptake	2	2.03	1.01	5.75	0.03
Interaction	1	0.06	0.01	0.08	0.78
Residuals	7	1.24	0.18	NA	NA

Table 4.4 : Two way ANOVA results

4.3 Bayesian approach in data analysis

Analysis of marathon performance via Bayesian ANOVA reveals information about the role of physical performance measures. It focuses on two main predictors: Maximum stroke volume speed and Maximum oxygen uptake speed. We sought to understand their impact on marathon success. Different Bayesian analyses reflect different assumptions about the data and parameters. Zellner $g=k$ scales the prior variance based on the number of predictors in the prior model by providing a moderate level of normalization. This prevents overfitting in simpler models. Zellner $g=n$ would adjust variance based on sample size. This will be useful by data when the data set is large. The JZS prior is often a more regularized version of the previously described priors. It acts conservatively and usually functions well across a wide variety of situations and requires less tweaking. Finaly, the Hyper-g prior introduces more flexibility by treating the variance parameter g as random, enabling the model to adapt dynamically to the data's complexity. This makes it suitable for detecting micro-variation in practice between groups and accounting at different levels. It provides a robust framework for statistical inference.

4.3.1 Bayesian One-Way ANOVA

In table 4.4, are the results from Bayesian ANOVA analyses of these three key variables, under alternative prior assumptions for the analysis of the effects on

https://doi.org/10.26219/heal.aueb.6544

marathon performance, for Marathon Speed, Stroke Volume, and Oxygen Uptake. As shown, Bayes Factors for models with oxygen uptake are all above 30 for every specification of the prior. For stroke volume, Bayes Factors range from approximately 4.01 to 4.29, indicating a low degree of evidence that stroke volume is important for predicting marathon performance.

Variable	Prior	Bayes.Factor..BF.
Stroke Volume	Zellner g = n	4.29
	Zellner g = k	4.09
	JZS	4.19
	Hyper-g	4.01
Oxygen Uptake	Zellner g = n	30.89
	Zellner g = k	30.48
	JZS	30.66
	Hyper-g	30.35

Table 4.4: Bayes ANOVA Results for Different Priors

The higher Bayes Factor values for oxygen uptake across different priors indicate its association with endurance performance and support the hypothesis that this variable is an important determinant in sustained aerobic activity. In contrast, while stroke volume is a critical determinant of cardiac output, the relatively lower Bayes Factors of this variable may indicate it acts more subtly and perhaps in concert with other physiological parameters rather than as a direct predictor. The observed differences in Bayes Factors across prior choices further support the need for sensitivity analyses with regard to priors in Bayesian modelling to ensure that results are not sensitive to specific distributional assumptions.

4.3.2 Bayesian Two-Way ANOVA for Interaction Effects

The Bayesian analysis of the interaction effect of stroke volume (SV) and oxygen uptake (VO2) yielded Bayes Factors ranging from 0.95 to 1.09 over different priors. These numbers indicate weak or no evidence for the interaction of these two variables. In all cases, priors resulted in BF values close to 1, suggesting that the inclusion of an interaction term between SV and VO2 does not provide a substantially better fit than models without interaction. Thus, these results suggest that stroke

https://doi.org/10.26219/heal.aueb.6544

volume and oxygen uptake make independent contributions to marathon performance, with no significant interactive effect. These results suggest that nonlinear dependencies between these physiological variables may exist that required the examination of alternative hierarchical modelling approaches or larger data sets.

Prior	Bayes.Factor..BF.
Zellner g = n	1.09
Zellner g = k	1.05
JZS)	1.06
Hyper-g	1.02

Table 4.5: Bayes Factors for the Interaction Effect (SV × VO_2) Under Different Priors

4.4 Simulation Example Using the SB-Corrected Prior Framework

At first we started out with a simple example to demonstrate the SB method in combination with the Rouder et al. (2012) framework. A dataset with m=4 groups and $k=\binom{4}{2}$ = 6 pairwise comparisons between groups is used. This situation is based on the dataset used in the previous chapters. The 4 groups in the dataset each had n=50 observations and were drawn from a normal distribution. Only the first group had a different mean from the original dataset, which was increased by 0.5.

To calculate the prior probabilities, we use Shaffer's method to define the possible model classes. Shaffer's method is a structured approach used to define model classes in statistical hypothesis testing, particularly in the context of multiple group comparisons. The method classifies candidate models according to the number of significantly differing group means rather than examining each pairwise difference in isolation. Shaffer's procedure is very useful in the context of Bayesian ANOVA because it assists in reducing model selection complexity at the expense of statistical simplicity. For 4 groups this is S(4)={0,3,4,5,6}. These classes represent different configurations of group differences. The models are distributed among the classes as follows:

- M_0: One model assumes all groups have the same mean.

https://doi.org/10.26219/heal.aueb.6544

- M_3: Four models assume one group differs while the other three share the same mean.
- M_4: Three models assume two groups have the same mean, while the other two form a separate pair.
- M_5: Six models assume all four groups have different means.
- M_6: One model assumes two groups are equal, while the other two are distinct.

Based on these distributions, the SB-corrected prior probabilities for each model class are calculated:

$$p(M_0)=\frac{1}{5}\frac{4!}{4!}=\frac{1}{5}$$

$$p(M_3)=\frac{1}{5}\frac{4!}{3!1!}=\frac{1}{20}$$

$$p(M_4)=\frac{1}{5}\frac{4!}{(2!)^3}=\frac{1}{15}$$

$$p(M_5)=\frac{1}{5}\frac{4!}{(2!)(1!)^2}=\frac{1}{30}$$

$$p(M_6)=\frac{1}{5}\frac{4!}{(1!)(4!)}=\frac{1}{5}$$

The uncorrected prior probabilities for each model are assigned equally across all models. The total number of possible models is calculated using the Bell number, Bn(4)=15. As a result, in the uncorrected scenario, each model is given a prior probability of $\frac{1}{15}$, regardless of which class it belongs to.

The Bell number is named after mathematician Eric Temple Bell. It counts the number of ways a set of elements can be divided into non-empty mutually exclusive subsets. These subsets must collectively cover the entire set with no overlap between elements in different subsets. The first few Bell numbers are 1, 1, 2, 5, 15, 52, 203, 877, 4140, and 21147. The Bell number B(n) can be calculated using this formula:

$$B(n+1)=\sum_{k=0}^{n}\binom{n}{k}B(k),$$

where $B(0)=1$

https://doi.org/10.26219/heal.aueb.6544

The Bayes factors for each model were calculated using a Bayesian ANOVA approach with the BayesFactor package (Morey & Rouder, 2015). These models were defined by relabeling groups. For instance, when $\mu_1=\mu_2$ and $\mu_3=\mu_4$ the groups were relabeled as {12} and {34}. The Bayes factor for this model, denoted $BF_{M4,z}^{A0}$ was calculated as:

$$BF_{A0}^{M4,z}=\frac{p(y|\mu_{12},\mu_{34})}{p(y|\mu_1=\mu_2=\mu_3=\mu_4)}$$

This approach generated 15 Bayes factors for all model configurations. To compute the posterior odds, the prior model probabilities were converted into prior odds, which were then multiplied by the corresponding Bayes factors. For example, the posterior odds for model z in class M_4 are given by:

$$\frac{p(M_{4,z}|y)}{p(M_0|y)}=\frac{p(M_4)}{p(M_0)}\times BF_{A0}^{M4,z}$$

The posterior probability for each model was then calculated by normalizing the posterior odds across all models. For model z, the posterior probability is computed as:

$$p(M_{4,z}|y)=\frac{\frac{p(M_{4,z}|y)}{p(M_0|y)}}{\sum\sum\left(\frac{p(M_{i,j}|y)}{p(M_0|y)}\right)}$$

Here, $|S(\cdot)|$ represents the number of model classes, and $|M_i|$ is the number of models in each class. This step converts Bayes factors to posterior probabilities when all Bayes factors have the same denominator model, meaning each Bayes factor provides evidence relative to a fixed model.

To measure the quantification of inclusion of some effects (e.g., pairwise group differences), posterior inclusion probabilities were computed by adding posterior probabilities across all models containing that effect. Prior inclusion probabilities were likewise computed, and inclusion Bayes factors were computed as posterior to prior inclusion probability ratios.

https://doi.org/10.26219/heal.aueb.6544

The simulation was run 1,000 times for stable estimates. The inclusion threshold was 0.5. All comparisons with the manipulated group included had a posterior inclusion probability above 0.5, while comparisons with noise only fell below the threshold.

Comparing the SB-corrected scenario to the uncorrected uniform prior, there are some notable differences:

- Prior inclusion probabilities are more conservative under SB, reducing the baseline chance of including any effect.
- Posterior inclusion probabilities are decreased for all real effects, as anticipated due to multiplicity adjustment.
- Despite the reduced probabilities, the SB method still maintains correct identification of true effects without spuriously including noisy comparisons.

4.4.1 Discussion of the results

Effect	Prior_incl	Post_incl.	Inclusion_BF	Prior_incl_corrected	Post_incl_corrected	Inclusion_BF_corrected
δ12	0.667	0.910	1.366	0.6	0.853	1.421
δ13	0.667	0.910	1.365	0.6	0.854	1.424
δ14	0.667	0.911	1.366	0.6	0.856	1.427
δ23	0.667	0.417	0.626	0.6	0.412	0.686
δ24	0.667	0.424	0.636	0.6	0.409	0.682
δ34	0.667	0.417	0.626	0.6	0.413	0.689

Table 4.6: Prior and Posterior Inclusion Probabilities with Scott & Berger Correction

The prior inclusion probability represents the probability that a given effect is included in the model before observing the data. This probability is determined by the structure of the model space and the way we define the set of possible models. In our analysis, we obtained a prior inclusion probability of 0.667, which stems from a structured approach to model selection where each effect is assigned an equal probability of being included based on combinatorial model counting.

In Bayesian model selection, we consider a set of possible models that can include or exclude specific effects. Each effect (denoted as $\delta_{12},\delta_{13},\delta_{14},\delta_{23},\delta_{24},\delta_{34}$) can either:

- Be included in a given model.

 https://doi.org/10.26219/heal.aueb.6544

- Be excluded from the model.

Since we are working with a Bayesian ANOVA framework, models are defined based on the grouping structure of means rather than individual pairwise comparisons. The total number of possible models that can be constructed depends on how these effects are included or excluded in different model classes. The prior inclusion probability is determined by the proportion of models in which a particular effect appears. In Bayesian model averaging, a common approach is to assume a hierarchical prior where each effect is included independently with a certain probability.

For a hierarchical Bayesian ANOVA, the probability of including an effect follows the structure of binomial inclusion priors, where the probability is computed as:

$$P(\textit{Effect Included}) = \frac{\textit{Number o fModels Containing the Effect}}{\textit{Total Models Considered}}$$

In this setup, we assume that each effect is included uniformly across models, ensuring a balanced distribution of prior probabilities.

Since we have six effects and a structured way of selecting models, the expected proportion of models containing a given effect is determined as follows:

1. The total number of possible models is calculated using the Bell number for four groups, which is B(4) = 15.
2. By examining the distribution of models within this space, each effect appears in 10 out of the 15 models in a Bayesian ANOVA framework.
3. Thus, the prior inclusion probability is:

$$\mathrm{P}(\textit{Effect Included}) = \frac{10}{15} = 0.667$$

This result is consistent with Bayesian binomial prior structures, where a uniform Beta(1,1) prior on model inclusion leads to an expected inclusion probability of 0.667.

The posterior inclusion probabilities gives updated probabilities for each effect after incorporating data evidence. These probabilities allow us to determine which effects are likely to be meaningful in explaining variance across groups.

- Effects associated with the first group (e.g., δ_{12},δ_{13},δ_{14}) exhibit high posterior inclusion probabilities (>0.5 both before and after correction). This indicates strong evidence that these effects contribute meaningfully to differences between groups.

https://doi.org/10.26219/heal.aueb.6544

- Conversely, effects that do not involve the first group (e.g., $\delta_{23}, \delta_{24}, \delta_{34}$) show lower posterior inclusion probabilities, below the 0.5 inclusion threshold. This suggests that these effects are less relevant and may represent noise.

This pattern supports the conclusion that group 1 significantly differs from the others, whereas the differences between the remaining groups are less pronounced.

Also Bayes factors show the strength of evidence in favor of including an effect in the model. The inclusion Bayes factors (BF_incl) measure how much more likely it is that an effect should be included rather than excluded.

- For the first set of effects ($\delta_{12}, \delta_{13}, \delta_{14}$), the Bayes factors are well above 1, indicating substantial support for inclusion.
- For the other effects ($\delta_{23}, \delta_{24}, \delta_{34}$), the Bayes factors are below 1, suggesting that these effects do not contribute significantly to explaining the variance.

The SB correction had the following effects:

- Lower Prior Inclusion Probabilities: The SB correction reduced the prior inclusion probability for each effect, making the prior more conservative.
- Reduced Posterior Inclusion Probabilities: All posterior inclusion probabilities decreased slightly after applying the SB correction. This is expected since adjusting for multiplicity controls for false positives.
- Stronger Evidence for True Effects: Despite the correction, the effects involving the first group still exceeded the 0.5 threshold, while non-significant effects remained below it. This indicates that the correction effectively maintains sensitivity while controlling for false discoveries.

4.5 A simulation Study for Multiple Cases

In the following section, we perform a more comprehensive simulation study in order to investigate the multiplicity adjustment across different priors in more detail. We considered an ANOVA model as before, but now varying the total number of groups to $K \in \{4, 6, 10\}$ and sample sizes per group of $n \in \{50, 100, 250, 500\}$. In addition, we have varied the percentage of true (in)equalities among groups, considering cases with 0%, 25%, 50%, 75%, and 100% equalities.

https://doi.org/10.26219/heal.aueb.6544

For K = 4, there are 3 possible (in)equalities, meaning that models could have 0, 1, 2, or 3 equalities in the true structure. For K = 6, there are 5 possible (in)equalities, with to models with 0, 1, 2, 3, 4, or 5 equalities and for K = 10 the number of inequalities are 9 with models with 0, 2, 4, 6, 8, or 9 equalities. Given a specific number of equalities, we randomly sampled a partition from all possible partitions matching the given number of equalities and used this structure to generate data. We generated each unique combination 100 times, analyzing each generated dataset under the same prior specifications as throughout the above. We evaluated both familywise error control and statistical power. Results for K=4 showed similar trends to other group sizes.

4.5.1 Description of Priors Used in the Simulation Study

In our simulation study, we evaluated different Bayesian prior structures to analyze their effects on familywise error control and statistical power. These priors influence the probability assigned to different hypotheses and thus play a crucial role in determining which models are selected. Below, we describe the key priors used:

1. Beta-Binomial (BB) Prior: The Beta-Binomial prior is a hierarchical Bayesian approach that places a Beta distribution as a prior on the probability of including effects in the model. The Beta parameters (α,β) control the distribution of inclusion probabilities:
    - $BB(\alpha=1,\beta=1)$: Assigns equal probability to each model (uninformative).
    - $BB(\alpha=1,\beta=K)$: Encourages fewer effects to be included as K increases, leading to a more conservative model.
    - BB(α=1,$\beta=\binom{K}{2}$)): Adjusts the prior based on a binomial structure, making it more flexible than the uniform Beta prior.
2. Dirichlet Process (DP) Prior: The Dirichlet Process prior is a nonparametric approach that allows for flexible clustering of group means. The concentration parameter (α) determines the strength of belief in group similarity:
    - DP (α=0.5): Encourages a stronger grouping of means, leading to fewer detected differences.

https://doi.org/10.26219/heal.aueb.6544

- DP(α=1): Balances prior information and data-driven inference.
- An adaptive approach by Gopalan and Berry (1998) selects α such that the prior probabilities of the null and alternative hypotheses are approximately equal, i.e., $p(H_0)=p(H_1)$. providing balanced inference across multiple comparisons.

3. Westfall et al. (1997) Prior: This is a frequentist-inspired Bayesian approach that prioritizes strong control over Type I errors (false positives). It is highly conservative, making it suitable for settings where strict error control is necessary.

4. Pairwise Bayes Factors (Bfs): The pairwise Bayes factor approach evaluates each comparison independently. While flexible, it can inflate false positives when multiple comparisons are made.

5. Uniform Prior: The uniform prior assigns equal probability to all models, making it the least informative prior. While it provides maximum flexibility, it can overestimate model complexity, leading to weaker familywise error control.

4.5.2 Familywise Error Rate

The results in Figure 4.1 show main differences between approaches in risk management of making the wrong inferences for testing multiple hypotheses under inequality constraints. Without any inequalities present in the true model, several methods such as the uniform prior, the pairwise Bayes factor approach, and the symmetric Dirichlet process are discovered to have relatively high error rates, meaning that these approaches are not conservative here. This may be because they tend to shrink all models equally plausible or fail to shrink strongly towards more parsimonious hypotheses. Alternatively, beta-binomial prior based approaches and certain Dirichlet process settings appear to have better control over false inferences, possibly because they place more emphasis on simpler or more structured models by design.

https://doi.org/10.26219/heal.aueb.6544

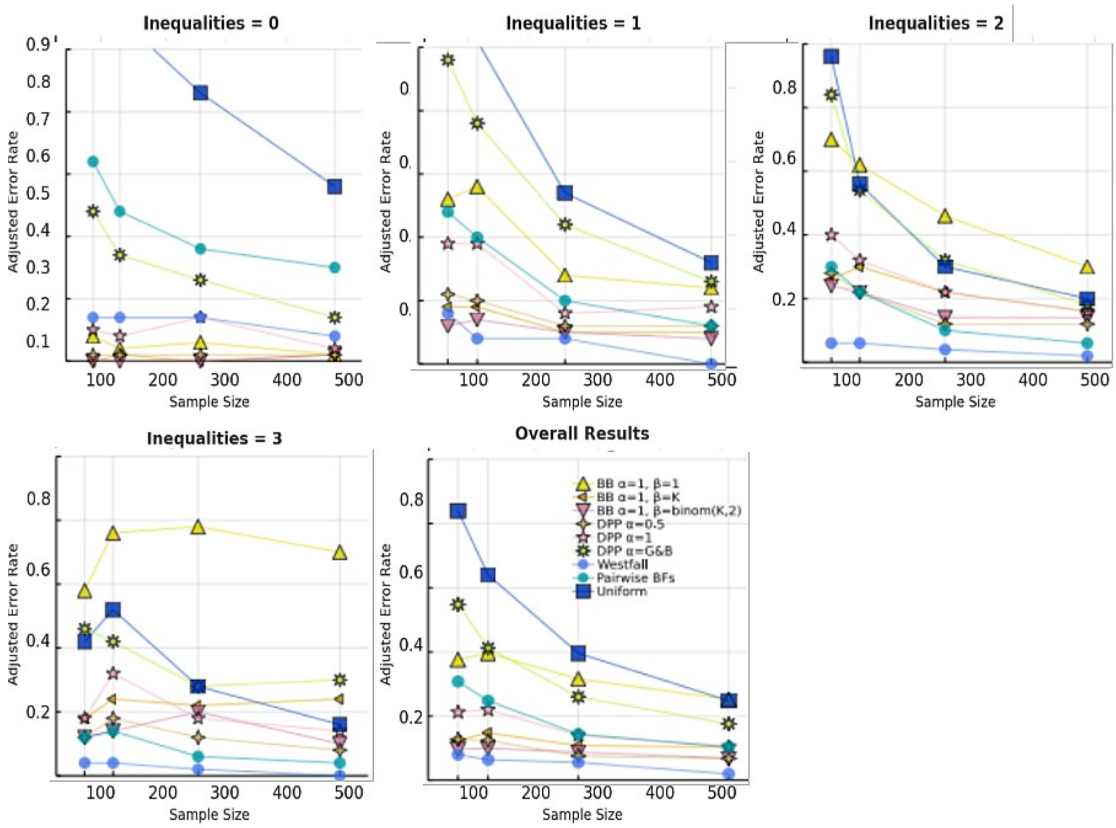

Figure 4.1: Familywise error rate across different priors and sample sizes for a model with 0, 1, 2 and 3 true inequalities when K=4 groups. The last graph presents the average familywise error rate across all inequality scenarios.

Surprisingly, adding small levels of actual inequalities to the model seems to alter this equilibrium. The error rates for the once-risky procedures begin declining, suggesting that some of them can be more compatible with cases of actual group differences. This could be a result of a lower number of options for wrong rejections, because actual constraints cut down on potential blunders. This benefit, however, does not occur uniformly for all techniques. The more conservative priors suffer a decrease in performance slightly as inequalities are introduced, and no trend is evident beyond this, indicative of a more complex interaction between model structure and prior assumptions.

If the performance is averaged over all configurations, however, the trend is more definitive. Methods with built-in multiplicity control or hierarchical shrinkage, such as the Westfall et al. procedure and beta-binomial priors with adaptive hyperparameters, show uniformly good familywise error control. These approaches appear to have a good compromise between flexibility and parsimony even when the

https://doi.org/10.26219/heal.aueb.6544

number of comparisons is large. Other approaches that rely on the equally likely assumption or do not provide multiplicity adjustments, e.g. the uniform prior and some symmetric Dirichlet priors, do badly. But in all of the approaches, the influence of sample size is evident: larger sets provide more solid performance, reducing the impact of the prior and allowing the data to play a more powerful role in determining the conclusions.

4.5.3 Statistical Power

Figure 4.2 shows the false discovery rates (FDR) of the methods, evaluated under varying numbers of true equalities in the model and increasing sample sizes. When all of the group means are equal (i.e., the null model is fully true), the beta-binomial prior with $\beta=1$ performs optimally, producing the lowest FDR. In contrast, Westfall et al.'s (1997) method has considerably higher FDR levels in this setting. This reflects an essential trade-off: while Westfall's method is so successful at controlling the familywise error rate (FWER), it appears too conservative to detect true null hypotheses and therefore rejects proportionally more false discoveries when all the hypotheses are truly null.

As the number of true equalities increases, most methods exhibit declining FDR performance. This is due to the compression of between-group differences. As more parameters are constrained to be equal, the effective signal is diminished, and the average pairwise contrasts are pulled towards zero. As a result, it becomes increasingly difficult to distinguish true from false hypotheses, especially for methods with more extreme multiplicity adjustment.

One exception to this trend is the uniform prior, which exhibits improving FDR with increasing sample size. This is consistent with the uniform prior's lack of structure—it has little influence on the model space so the likelihood dominates as more data are seen.

https://doi.org/10.26219/heal.aueb.6544

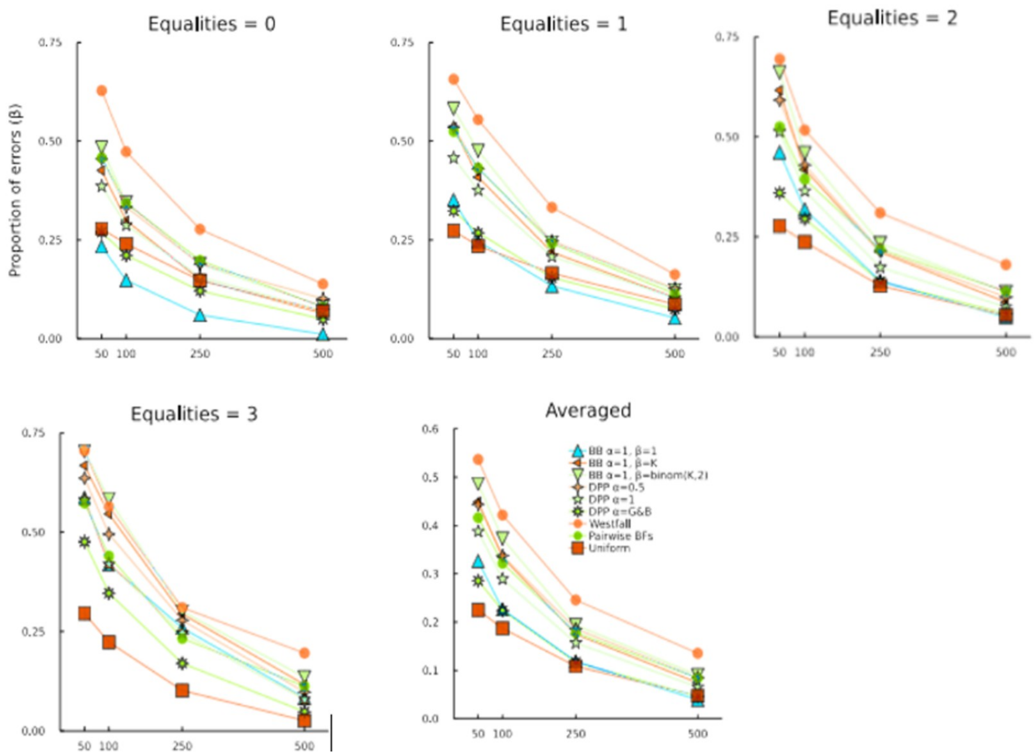

Figure 4.2: False positive rate across priors and sample sizes for K=4 groups, considering 0 to 3 true inequalities. The last panel shows the average error rate.

The last panel presents results averaged across all equality configurations. The pattern that emerges is consistent with the inverse relationship between FWER and FDR: procedures that previously had poor control of FWER (e.g., the uniform prior, symmetric Dirichlet process prior, and beta-binomial prior with $\beta=1$ tend to give lower FDRs in this case. These methods are more liberal, with the consequence that there is greater power to detect null effects but also greater danger of Type I errors in other cases. Westfall et al.'s method, although conservative in maintaining FWER at or below nominal levels, pays the price in terms of high FDR, most likely due to an inflation of false negatives.

Moderately performing methods, like pairwise Bayes factors, Dirichlet process priors with moderate concentration parameters, and beta-binomial priors with $\beta=K$ lie somewhere between these extremes. Notably, as sample size increases, methodological differences begin to fade. This is as it should be in the Bayesian paradigm: larger samples decrease the influence of prior assumptions, leading to more agreement between methods as inference becomes increasingly data-driven.

https://doi.org/10.26219/heal.aueb.6544

4.5.4 Discussion of the Simulation

Our findings show that no approach is systematically the best across all situations. Although some priors have good familywise error control, this is not necessarily at the expense of statistical power, and vice versa. The Beta-Binomial prior with α=1 performed best overall, minimizing both Type I and Type II errors. Its relative performance declined, however, with models beyond the null (all equal) and full model (all different) models. This suggests that although it is an excellent general-purpose prior, it may not be the optimal choice when capturing more subtle group differences. Beta-Binomial priors and Dirichlet Process (DP) priors produced similar patterns, reflecting that both methods are well-suited to varying degrees of equality across group means. They are generalizable enough to cover all cases and thus are robust choices for Bayesian ANOVA.

Out of the different priors, there were trade-offs. Westfall method was strongest at controlling false positives but only at the cost of higher false negatives. That means it is too conservative and therefore should only be used in situations where not making false discoveries is more vital than detecting real effects. On the other hand, Pairwise Bayes Factors and Uniform Priors generated more false positives at a higher error rate for overestimating differences between groups. The Dirichlet Process Priors with α=0.5 and α=1 had a fair balance between Type I and Type II errors, particularly for moderate levels of inequality. They were less conservative than Beta-Binomial priors but more flexible than standard frequentist adjustments and therefore offered an effective compromise between conservative and flexible approaches.

The impact of sample size and model complexity was another key factor influencing prior performance. As sample size increased, all priors showed reduced differences in performance, suggesting that data-driven inference dominates in large samples, making the choice of prior less influential in high-powered studies. Additionally, when the number of true inequalities increased, the performance of pairwise Bayes factors, the uniform prior, and the Westfall method improved. This suggests that when the true structure contains more differences, the risk of false positives decreases naturally, reducing the need for strict multiple comparison adjustments.

In conclusion, prior choice should be guided by the study goal, whether the priority is to prevent false positives or to achieve maximum sensitivity to true effects.

https://doi.org/10.26219/heal.aueb.6544

 https://doi.org/10.26219/heal.aueb.6544

CHAPTER 5

Discussion and Final Remarks

This thesis considered Bayesian ANOVA as an alternative to frequentist techniques and presented its advance in dealing with multiple comparisons as well as accommodating prior information into statistical inference. Classical ANOVA relies fundamentally on p-values and requires multiplicity adjustments on an ad hoc basis that give the conclusions unnecessarily conservative and at a loss in terms of power. Bayesian ANOVA, on the other hand, addresses multiplicity adjustment inherently through prior-model probabilities giving a better and probabilistic framework for testing hypotheses.

In Chapter 1, we set out the background and motivation for this work and the need for a Bayesian alternative to ANOVA. We outlined the limitations of frequentist ANOVA, particularly in how it handles multiple comparisons, and how Bayesian statistics offer a more sensible statistical paradigm. The objectives of the thesis were also set out, along with an overview of the approach adopted.

Chapter 2 provided a formal introduction to Bayesian ANOVA, explaining its theoretical foundations. We reviewed Bayesian probability theory, prior distributions, posterior inference, and Bayes factors as a model comparison tool. The concept of Bayesian model averaging was also discussed, showing how Bayesian ANOVA incorporates model uncertainty in a principled way. Additionally, this chapter introduced the notation and mathematical formulations used throughout the thesis.

Chapter 2 provided a formal introduction to Bayesian ANOVA and covered its theory. We had gone through Bayesian probability theory, prior distributions, posterior inference, and Bayes factors as a measure of model comparison. The idea of Bayesian model averaging was also touched upon in order to underscore how Bayesian ANOVA expresses model uncertainty in a principled way.

Chapter 3 addressed the problem of multiplicity adjustment and how Bayesian methods handle this issue in a natural way. We provided Bayesian solutions to multiplicity, e.g.,Scott & Berger's multiplicity correction, hierarchical priors, and shrinkage-based methods. A number of priors, including the Dirichlet Process prior,

https://doi.org/10.26219/heal.aueb.6544

Beta-Binomial prior, and Westfall's method, were explored as means of controlling false positives in Bayesian hypothesis testing.

Chapter 4 outlined a detailed simulation study contrasting the performance of Bayesian ANOVA with different prior choices. The study contrasted the impact of priors on model selection, inference accuracy, and false discovery control. We compared different Bayesian priors on their power-error trade-off. The results showed that Bayesian ANOVA provides a more flexible and informative analysis platform with decreased risk of false discovery, eliminating the need for the use of stringent multiple comparison adjustments.

While this work shows the advantages of Bayesian ANOVA, aspects of it need to be looked into further. Future research may extend Bayesian ANOVA to more complex experimental designs, such as mixed-effects models and hierarchical models, for which Bayesian analysis has theoretical strength. Random effects in Bayesian ANOVA hold particular promise as they allow variability at the level of the individuals with a solid probabilistic paradigm. Furthermore, its application in longitudinal analysis can offer more insight into experimental conditions' temporal changes.

Sensitivity analysis is another region that requires increased research. More extension of exploration on the effects of informative and weakly informative priors would enhance the dependability and robustness of Bayesian ANOVA uses. Selection techniques for the prior could further provide increased model flexibility and accuracy of inference.

A possible direction for extension is Bayesian ANOVA applications in high-dimensional data settings such as genomics and machine learning, where several hypothesis testing constitutes a main issue. Its consideration in fields such as neuroscience and bioinformatics, where controlling the false discovery is significant, would further validate its application.

Overall, this study is just a part of the growing body of literature on promoting Bayesian analysis methods in statistics. The study confirms that Bayesian ANOVA is not just a viable but sometimes even a superior alternative to classical methods, particularly in complex experimental designs in which numerous hypothesis tests are conducted.

https://doi.org/10.26219/heal.aueb.6544

References

Abramovich, F., Benjamini, Y., Donoho, D. L., & Johnstone, I. M. (2006). Adapting to unknown sparsity by controlling the false discovery rate. *The Annals of Statistics, 34*(2), 584–653. https://doi.org/10.1214/009053606000000074

Anscombe, F. J., & Aumann, R. J. (1963). A definition of subjective probability. *The Annals of Mathematical Statistics, 34*(1), 199–205 https://doi.org/10.1214/aoms/1177704255

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science, 7*(6), 543–554. https://doi.org/10.1177/1745691612459060

Bennett, C. M., Baird, A. A., Miller, M. B., & Wolford, G. L. (2009). Neural correlates of interspecies perspective taking in the post-mortem Atlantic salmon: An argument for multiple comparisons correction. *Journal of Serendipitous and Unexpected Results, 1*(1), 1–5.

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., ... & Johnson, V. E. (2017). Redefine statistical significance. *Nature Human Behaviour, 2*(1), 6–10. https://doi.org/10.1038/s41562-017-0189-z

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological), 57*(1), 289–300.

Bernardo, J. M., & Smith, A. F. M. (1994). *Bayesian theory.* John Wiley & Sons.

Berry, D. A., & Hochberg, Y. (1999). Bayesian perspectives on multiple comparisons. *Journal of Statistical Planning and Inference, 82*(1-2), 215–227. https://doi.org/10.1016/S0378-3758(99)00044-0

Berger, J. O. (2006). The case for objective Bayesian analysis. *Bayesian Analysis, 1*(3), 385–402. https://doi.org/10.1214/06-BA115

Berger, J. O., & Delampady, M. (1987). Testing precise hypotheses. *Statistical Science, 2*(3), 317–335. https://doi.org/10.1214/ss/1177013238

Berger, J. O., & Sellke, T. (1987). Testing a point null hypothesis: The irreconcilability of P values and evidence. *Journal of the American Statistical Association, 82*(397), 112–122. https://doi.org/10.2307/2289130

Berry, D. A., & Christensen, R. (1979). Empirical Bayes methods: Theory and applications. Springer.

https://doi.org/10.26219/heal.aueb.6544

Billat VL, Petot H, Landrain M, Meilland R, Koralsztein JP, Mille-Hamard L. Cardiac output and performance during a marathon race in middle-aged recreational runners. ScientificWorldJournal. 2012;2012:810859. doi: 10.1100/2012/81085

Blackwell, D., & MacQueen, J. B. (1973). Ferguson distributions via Pólya urn schemes. *The Annals of Statistics, 1*(2), 353–355. https://doi.org/10.1214/aos/1176342372

Bogdan, M., Ghosh, J. K., & Tokdar, S. T. (2008). A comparison of the Benjamini-Hochberg procedure with some Bayesian rules for multiple testing. In *Beyond Parametrics in Interdisciplinary Research: Festschrift in Honor of Professor Pranab K. Sen* (pp. 211–230). Institute of Mathematical Statistics. https://doi.org/10.1214/193940307000000243

Brooks, S. (2003). Bayesian computation: Principles and practice. *Statistical Modelling, 3*(3), 181–198. https://doi.org/10.1191/1471082x03st054oa

Carvalho, C. M., & Scott, J. G. (2009). Objective Bayesian model selection in Gaussian graphical models. *Biometrika, 96*(3), 497–512.

Casella, G., Moreno, E., & Girón, F. J. (2014). A consistent high-dimensional Bayesian variable selection model. *Bernoulli, 20*(3), 1401–1427. https://doi.org/10.3150/13-BEJ531

Clyde, M. (2017). *BAS: Bayesian variable selection and model averaging using Bayesian adaptive sampling* (R package version 1.5.3). Retrieved from https://cran.r-project.org/web/packages/BAS/

Cohen, J. (1994). The earth is round (p < .05). *American Psychologist, 49*(12), 997–1003. https://doi.org/10.1037/0003-066X.49.12.997

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis.* Routledge.

Dahl, D. B., & Newton, M. A. (2007). Multiple testing and false discovery control in Bayesian selection models. *Journal of the American Statistical Association, 102*(479), 517–526.

De Finetti, B. (1937). La prévision: Ses lois logiques, ses sources subjectives. *Annales de l'Institut Henri Poincaré, 7*, 1–68.

Dellaportas, P., & Smith, A. F. M. (1993). Bayesian inference for generalized linear and proportional hazards models via Gibbs sampling. *Applied Statistics, 42*(3), 443–459. https://doi.org/10.2307/2984486

Dickey, J. M. (1971). The weighted likelihood ratio, linear hypotheses on normal location parameters. *The Annals of Mathematical Statistics, 42*(1), 204–223.

Do, K. A., Müller, P., & Tang, F. (2005). A Bayesian mixture model for differential gene expression. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 54*(3), 627–644.

https://doi.org/10.26219/heal.aueb.6544

Duncan, D. B. (1961). Bayesian probability and multiple comparisons. *Biometrics, 17*(2), 252–259.

Duncan, D. B. (1965). A Bayesian approach to multiple comparisons. *Technometrics, 7*(1), 171–222.

Efron, B., Tibshirani, R., Storey, J. D., & Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association, 96*(456), 1151–1160. https://doi.org/10.1198/016214501753382129

Foster, D. P., & George, E. I. (1994). The risk inflation criterion for multiple regression. *The Annals of Statistics, 22*(4), 1947–1975.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics, 1*(2), 209–230.

Fernandez, C., Ley, E., & Steel, M. F. J. (2001). Benchmark priors for Bayesian model averaging. *Journal of Econometrics, 100*(2), 381–427.

Fleming, T. R. (1982). Clinical trials: Discerning hype from substance. *Statistics in Medicine, 1*(4), 357–374. https://doi.org/10.1002/sim.4780010403

Gopalan, R., & Berry, D. A. (1998). Bayesian Multiple Comparisons Using Dirichlet Process Priors.Journal of the American Statistical Association, 93(443), 1130–1139. https://doi.org/10.2307/2669856

Gelfand, A. E., Hills, S. E., Racine-Poon, A., & Smith, A. F. M. (1992). Illustration of Bayesian inference in normal data models using Gibbs sampling. *Journal of the American Statistical Association, 87*(417), 97–104. https://doi.org/10.1080/01621459.1992.10475182

Gelfand, A. E., & Rubin, D. B. (2004). Markov chain Monte Carlo methods in biostatistics. *Statistical Science, 19*(1), 92–114.

Gelman, A. (2007). Struggles with survey weighting and regression modelling. *Statistical Science, 22*(2), 153–164. https://doi.org/10.1214/088342307000000035

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis, 1*(3), 515–533. https://doi.org/10.1214/06-BA117A

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (1996). *Bayesian data analysis* (3rd ed.). Chapman & Hall/CRC.

Gilks, W. R. (1994). Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series D (The Statistician), 43*(1), 169–178. https://doi.org/10.2307/2348941

Gopalan, R., & Berry, D. A. (1998). Bayesian multiple comparisons using Dirichlet process priors. *Journal of the American Statistical Association, 93*(443), 1130–1139.

Hartigan, J. A. (1990). Partition models. *Communications in Statistics - Theory and Methods, 19*(8), 2745–2756.

https://doi.org/10.26219/heal.aueb.6544

Henderson, D. J., & Ghosh, J. (2019). Flexible Bayesian multiple comparison adjustment using Dirichlet process and beta-binomial model priors. *Bayesian Analysis*, *14*(1), 223–256. https://doi.org/10.48550/arXiv.2208.07086

Hinne, M., Gronau, Q. F., Bergh, D. J. van den, & Wagenmakers, E.-J. (2020). A Bayesian approach for estimating and testing small effects. *Psychological Methods, 25*(6), 746–769.

Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian model averaging: A tutorial. *Statistical Science, 14*(4), 382–401.

Hoijtink, H. (2012). *Informative hypotheses: Theory and practice for behavioral and social scientists.* Chapman & Hall/CRC Press.

Hochberg, Y., & Tamhane, A. C. (1987). *Multiple comparison procedures.* Wiley.

Ishwaran, H., & James, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association, 96*(453), 161–173.

Jeffreys, H. (1938). Significance tests when several degrees of freedom arise simultaneously. *Proceedings of the Cambridge Philosophical Society, 34*(1), 217–219.

Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford University Press.

Jeffreys, H., & Berger, J. O. (1992). Ockham's razor and Bayesian analysis. *American Scientist, 80*(1), 64–72.

Johnstone, I. M., & Silverman, B. W. (2004). Needles and straw in haystacks: Empirical Bayes estimates of possibly sparse sequences. *The Annals of Statistics, 32*(4), 1594–1649.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association, 90*(430), 773–795. https://doi.org/10.1080/01621459.1995.10476572

Kass, R. E., & Wasserman, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the American Statistical Association, 90*(431), 928–934.

Klugkist, I., Laudy, O., & Hoijtink, H. (2005). Inequality constrained analysis of variance: A Bayesian approach. *Psychological Methods, 10*(4), 477–493. https://doi.org/10.1037/1082-989X.10.4.477

Kruschke, J. K., & Liddell, T. M. (2017). Bayesian data analysis for newcomers. *Psychonomic Bulletin & Review, 24*(2), 567–580

Li, J., & Shang, Y. (2015). Bayesian multiple testing under dependence. *Biometrika, 102*(2), 481–493

https://doi.org/10.26219/heal.aueb.6544

Lindley, D. V. (1980). The Bayesian approach to hypothesis testing. *Journal of the Royal Statistical Society: Series D (The Statistician), 29*(4), 293–337.

Lindley, D. V. (2000). The philosophy of statistics. *Journal of the Royal Statistical Society: Series D (The Statistician), 49*(3), 293–337.

Liang, F., Paulo, R., Molina, G., Clyde, M. A., & Berger, J. O. (2008). Mixtures of g priors for Bayesian variable selection. *Journal of the American Statistical Association, 103*(481), 410–423. https://doi.org/10.1198/016214507000001337

Link, W. A., Barker, R. J., & Sauer, J. R. (2010). *Bayesian inference: With examples from ecology.* Academic Press.

Mazzetta, C., Solari, A., & Ferrari, P. A. (2007). Bayesian hierarchical models for multiple comparisons. *Statistical Methods & Applications, 16*(1), 35–49.

McCarthy, M. A., & Masters, P. (2007). Profiting from prior information in Bayesian analyses of ecological data. *Journal of Applied Ecology, 42*(6), 1012–1019. https://doi.org/10.1111/j.1365-2664.2005.01101.x

Meng, X. L., & Dempster, A. P. (1987). Bayesian p-values. *Journal of the American Statistical Association, 82*(398), 397–401.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics, 21*(6), 1087–1092. https://doi.org/10.1063/1.1699114

Mitra, R., Mueller, P., & Ji, Y. (2017). Bayesian graphical models for inference on interactions. *Statistical Science, 32*(1), 86–105.

Morey, R. D., & Rouder, J. N. (2015). *BayesFactor: Computation of Bayes factors for common designs* (R package version 0.9.12-4.2). Retrieved from https://richarddmorey.github.io/BayesFactor/

Neath, A. A., & Cavanaugh, J. E. (2006). Bayesian pairwise multiple comparisons with a control. *Journal of Statistical Planning and Inference, 136*(11), 3950–3964.

O'Hagan, A., & Forster, J. J. (2004). *Bayesian inference* (2nd ed.). Oxford University Press.

Poirier, D. J. (2006). *The Bayesian approach to econometrics* (2nd ed.). Wiley.

Press, S. J., Tanur, J. M., & Mosteller, F. (2003). *Bayesian statistics: Principles, models, and applications.* Wiley-Interscience.

Rasmussen, C. E., Ghahramani, Z., & Williams, C. K. I. (1999). Bayesian treatment of the latent variable Gaussian model. *Neural Computation, 11*(8), 1847–1863.

Robert, C. P. (1993). A note on Jeffreys-Lindley paradox. *Statistical Science, 8*(1), 48–54.

 https://doi.org/10.26219/heal.aueb.6544

Romano, J. P., & Wolf, M. (2005). Stepwise multiple testing as formalized data snooping. *Econometrica, 73*(4), 1237–1282. https://doi.org/10.1111/j.1468-0262.2005.00615.x

Rouder, J. N., Engelhardt, C. R., McCabe, S., & Morey, R. D. (2017). Model comparison in ANOVA. *Psychological Methods, 22*(2), 356–374. https://doi.org/10.1037/met0000071

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology, 56*(5), 356–374. https://doi.org/10.1016/j.jmp.2012.08.001

Scott, J. G. (2009). The impact of priors on Bayesian multiple testing. *Journal of Statistical Planning and Inference, 139*(12), 4038–4048. https://doi.org/10.1016/j.jspi.2009.06.017

Scott, J. G., & Berger, J. O. (2006). An exploration of aspects of Bayesian multiple testing. *Journal of Statistical Planning and Inference, 136*(7), 2144–2162.

Scott, J. G., & Berger, J. O. (2010). Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *The Annals of Statistics, 38*(5), 2587–2619. https://doi.org/10.1214/10-AOS792

Sellke, T., Bayarri, M. J., & Berger, J. O. (2001). Calibration of p-values for testing precise null hypotheses. *The American Statistician, 55*(1), 62–71.

Shaffer, J. P. (1986). Modified sequentially rejective multiple test procedures. Journal of the Royal Statistical Society: Series B (Methodological), 48(1), 81-92.

Stephens, M., & Balding, D. J. (2009). Bayesian statistical methods for genetic association studies. *Nature Reviews Genetics, 10*(10), 681–690.

Storey, J. D., & Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences, 100*(16), 9440–9445. https://doi.org/10.1073/pnas.1530509100

Swain, P. S., Elowitz, M. B., & Siggia, E. D. (2009). Bayesian inference in dynamic systems. *Proceedings of the National Academy of Sciences, 106*(51), 21051–21056.

T. de Jong, (2020). *A Bayesian approach to the correction for multiplicity* (Master's thesis, University of Amsterdam). https://doi.org/10.31234/osf.io/s56mk

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. *Bayesian Inference and Decision Techniques, 6*, 233–243.

https://doi.org/10.26219/heal.aueb.6544

Statistical Software and R Packages

Heck, D. W., & Wagenmakers, E. J. (2021). *bayesanova: Bayesian analysis of variance*. R package version 1.0. https://cran.r-project.org/package=bayesanova

Morey, R. D., & Rouder, J. N. (2015). *BayesFactor: Computation of Bayes factors for common designs*. R package version 0.9.12-4.2. https://cran.r-project.org/package=BayesFactor

Plummer, M. (2003). *JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling*. Retrieved from https://mcmc-jags.sourceforge.io

JASP Team. (2024). *JASP (Version 0.17.2) [Computer software]*. https://jasp-stats.org/

Appendix

A. Simulation Results for K = 10

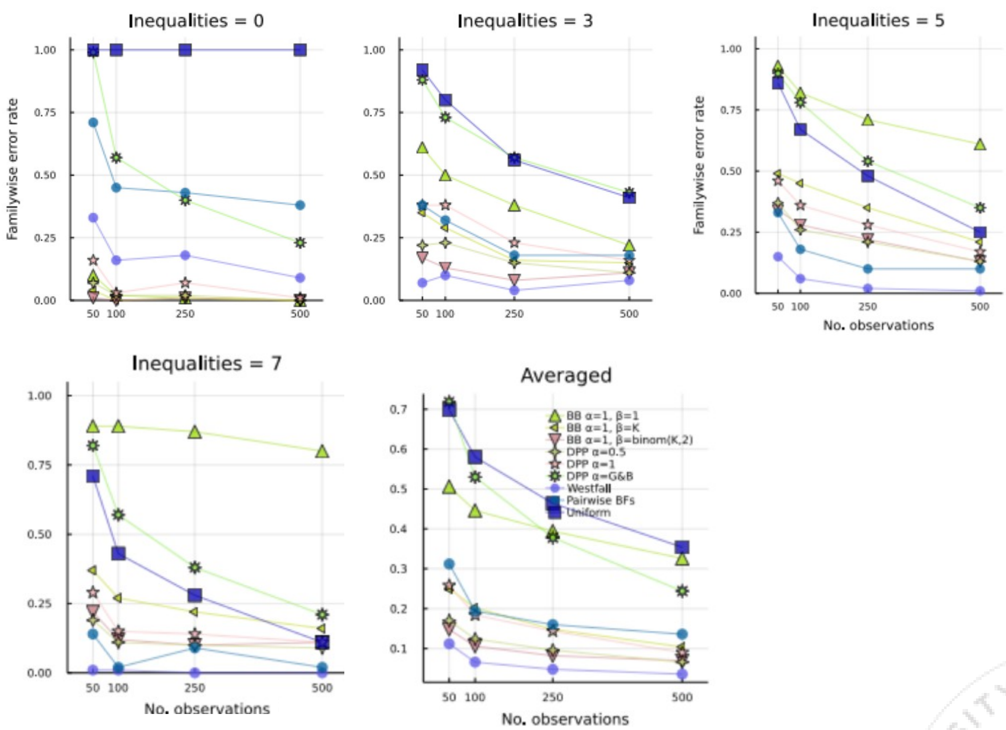

Figure 5.1: Familywise error rate (FWER) of some Bayesian multiple testing procedures for a range of numbers of true inequality constraints in the model. Results

https://doi.org/10.26219/heal.aueb.6544

are shown over increasing sample sizes, by number of true inequalities (0, 3, 5, and 7), and bottom-right panel showing performance averaged over all inequality levels.

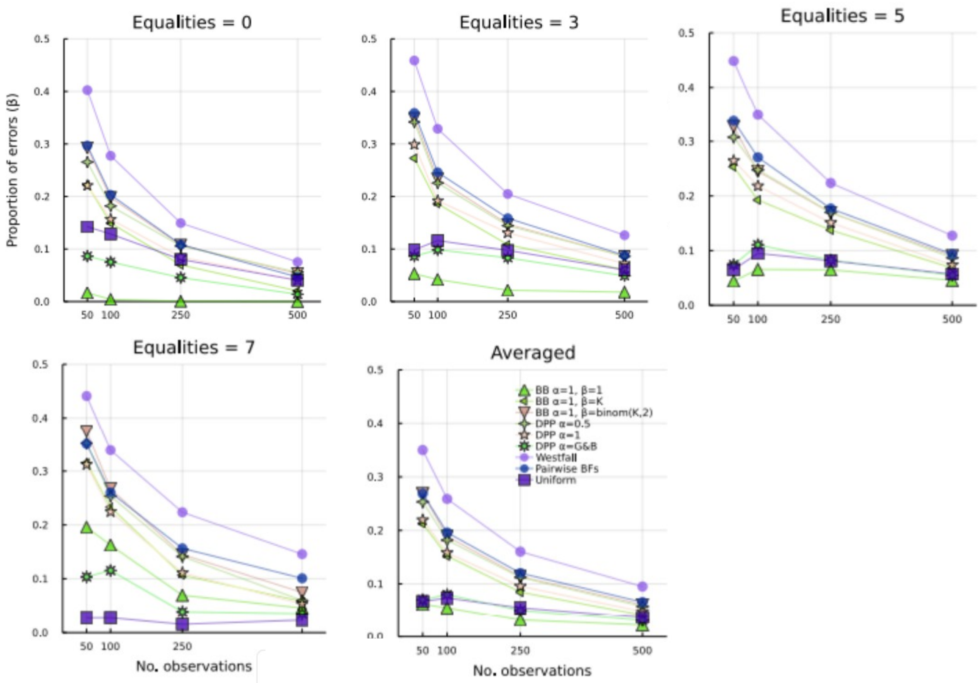

Figure 5.2: Comparison of Bayesian Multiple Testing Procedures Based on Type II Error Rates under Increasing Sample Sizes and Varying Numbers of Equality Constraints in the True Model

B. Code and Data Availability

A part of the simulation scripts and data analysis code developed for this thesis are available in the GitHub repository: https://github.com/alkistimam/Bayesian-Anova-and-Multiplicity-Adjustment-Thesis.git