Bayesian model selection with applications in social science
Wetzels, R.M.

Citation for published version (APA):
Wetzels, R. M. (2012). Bayesian model selection with applications in social science

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
1 Introduction

Model selection is arguably one of the most important tools in academia. In the social sciences, for instance, researchers often wish to test competing hypotheses, theories, or models. For example, does the learning curve follow an exponential or a power law? Is the ability to inhibit a response better predicted by activation in the subthalamic nucleus or by activation in the anterior cingulate cortex? Are people more creative when they are inside or outside a big box? Can women predict the presence of porn but not the presence of towels and flowers?

The question at hand is which theory is more plausible, or better supported by the data. This question can be addressed by the use of model selection methods. These methods allow researchers to compare models or theories to each other, evaluate these models in a principled manner, and compute which one is more likely after having conducted an experiment. The most popular form of model selection is null hypothesis testing – the topic of this thesis.

Null hypothesis tests are central to the psychological literature. More specifically, frequentist null hypothesis tests are central to the psychological literature. This is understandable, as these tests have been thoroughly studied, are well developed, and yield convenient yes-no decisions. However, these frequentist tests have well-known drawbacks, the negative impact of which is exacerbated by the fact that their use has become ritualized and the end-results are easily misinterpreted and misused (for examples see Cohen, 1994).

In an attempt to present and promote an alternative to the frequentist approach, this thesis focuses on a different statistical philosophy, the Bayesian philosophy. Within this statistical philosophy, we focus on a Bayesian approach to null hypothesis testing. In the remainder of the introduction, we first provide an example of a well-known hypothesis test, the $t$ test. Then we briefly discuss various measures that are used to quantify evidence. Next, the Bayes factor is discussed. This is the most common Bayesian measure of evidence and is used throughout this thesis. Finally, we sketch the outline of this thesis.

1.1 Hypothesis Testing, an Example: The $t$ Test

The $t$ test is one of the most popular hypothesis tests in academia. It is used to investigate if there is a significant difference between two group means. In frequentist statistics, this is accomplished by constructing a null hypothesis and assessing whether the data disprove it. The use of the $t$ test is illustrated by the following example.

Suppose we are interested in investigating whether consumption of alcohol is related to the onset of depression. To study this question empirically, two groups of participants are constructed. One group consists of participants who frequently drink alcohol and another group consists of participants who never drink alcohol. Both groups fill out a depression questionnaire resulting in a depression score for each participant (see Table 1.1).

Figure 1.1 shows that the mean score on the depression scale for the alcohol group ($\bar{X}_A = -0.5$) is lower than the mean score of the non-alcohol group ($\bar{X}_{NA} = 0.5$). However, the distribution of depression scores of both groups overlap. To test whether these means are statistically different, the $t$ test is used. This test yields a so-called $p$ value that determines whether or not the effect can be called significant. Usually, an
1. Introduction

<table>
<thead>
<tr>
<th>Depression Scores</th>
<th>Alcohol Group</th>
<th>Non-Alcohol Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.5</td>
<td>1.1</td>
<td></td>
</tr>
<tr>
<td>0.0</td>
<td>0.8</td>
<td></td>
</tr>
<tr>
<td>0.2</td>
<td>0.3</td>
<td></td>
</tr>
<tr>
<td>0.4</td>
<td>0.4</td>
<td></td>
</tr>
</tbody>
</table>

... ...

Table 1.1: The depression scores of the alcohol group and the non-alcohol group.

![Histogram](image.jpg)

Figure 1.1: A histogram of depression scores for the alcohol group and the non-alcohol group. The mean score of the alcohol group is $\bar{X}_A = -0.5$, the mean score for the non-alcohol group is $\bar{X}_{NA} = 0.5$.

... ...

1.2 Various Measures of Evidence

In the previous example, we entertained a null hypothesis of equal means and an alternative hypothesis of unequal means. In other words, the null hypothesis is the hypothesis that the two groups have an equal mean and the alternative hypothesis states that the two groups do not have an equal mean.

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 \neq \mu_2.$$
As mentioned in the last section, the most popular test for this scenario is the frequentist $t$ test. The data are used to calculate a $t$ statistic; and this statistic combined with the degrees of freedom yields a $p$ value. The resulting $p$ value is defined as the probability of obtaining a test statistic (in this case the $t$ statistic) at least as extreme as the one that was observed in the experiment, given that the null hypothesis is true and the sample is generated according to a specific intended procedure. Obviously, the definition of the $p$ value is difficult to interpret.

In this dissertation we point out that there are several problems concerning the use of $p$ value hypothesis testing, such as the inability to gather evidence in favor of the null hypothesis, the asymmetry between the null hypothesis and the alternative hypothesis, the fallacy of the transposed conditional, and the consequences of optional stopping.

There are alternatives to the $p$ value when evaluating hypotheses or models. One could for example estimate the size of an observed effect. Another way of analyzing the data would be to compare the two hypotheses to each other. This could for example be done by computing an information criterion (such as AIC or DIC), or by computing the Bayes factor. The information criteria and the Bayes factor take a different perspective than null hypothesis significance testing. These alternative methods treat the alternative and the null hypothesis alike whereas the $p$ value only considers the null hypothesis. Another difference is that the information criteria and the Bayes factor implement an Occam’s razor, meaning that they strike a compromise between model fit and model complexity. Consequently, if two models fit the data equally well, the least complex model is preferred. This thesis focuses on the Bayes factor as an alternative to $p$ value hypothesis testing.

### 1.3 The Bayes Factor

In Bayesian statistics, uncertainty (or degree of belief) is quantified by probability distributions over parameters. This makes the Bayesian approach fundamentally different from the classical “frequentist” approach, which relies on sampling distributions of data (J. O. Berger & Delampady, 1987; J. O. Berger & Wolpert, 1988; D. V. Lindley, 1972; Jaynes, 2003).

An important aspect of Bayesian statistical practice is defining the so-called prior distribution. In some cases, this prior distribution reflects the information about the parameters before the data is observed. In other cases, this prior distribution reflects as little information as possible. The choice for a particular prior distribution often depends on the type of problem that is being considered. In this thesis, we focus on a class of prior distributions that are called uninformative, objective, or default.

Another important part of Bayesian statistical inference is the likelihood function that describes the information contained in the data. Using the prior and the likelihood, the posterior distribution is found by using Bayes’ rule:

$$f(\theta_\gamma | Y) = \frac{f(Y | \theta_\gamma)p(\theta_\gamma)}{\int_{\Theta} f(Y | \theta_\gamma)p(\theta_\gamma)\, d\theta_\gamma},$$

where $f(Y | \theta_\gamma)$ is the likelihood function of the data $Y$, $p(\theta_\gamma)$ is the prior distribution of the model parameters $\theta_\gamma$, and $f(\theta_\gamma | Y)$ is the posterior distribution of the model parameters under the model $M_\gamma$. This posterior distribution reflects all the information in the data about the model parameters $\theta_\gamma$.

Within the Bayesian framework, one may quantify the evidence for one hypothesis relative to another. The Bayes factor is the most commonly used (although certainly not the only possible) Bayesian measure for doing so (Jeffreys, 1961; Kass & Raftery, 1995).
1. Introduction

The Bayes factor is the probability of the data under one hypothesis or model, relative to the other; it is a weighted average likelihood ratio that indicates the relative plausibility of the data under the two competing hypotheses.

An alternative—but formally equivalent—conceptualization of the Bayes factor is as a measure of the change from prior model odds to posterior model odds, brought about by the observed data. This change is often interpreted as the weight of evidence (Good, 1983, 1985). Before seeing the data \( Y \), the two models \( M_0 \) and \( M_1 \) are assigned prior probabilities \( p(M_0) \) and \( p(M_1) \). The ratio of the two prior probabilities defines the prior odds. When the data \( Y \) are observed, the prior odds are updated to posterior odds, which is defined as the ratio of the posterior probabilities \( p(M_0 | Y) \) and \( p(M_1 | Y) \):

\[
\frac{p(M_1 | Y)}{p(M_0 | Y)} = \frac{p(Y | M_1)}{p(Y | M_0)} \times \frac{p(M_1)}{p(M_0)},
\]

or

\[
\text{Posterior Odds} = \text{Bayes Factor} \times \text{Prior Odds}.
\]

Equation 1.1 and 1.2 show that the change from prior odds to posterior odds is quantified by \( p(Y | M_1)/p(Y | M_0) \), the Bayes factor \( BF_{10} \).

Under either conceptualization, the Bayes factor has an appealing and direct interpretation as an odds ratio. For example, \( BF_{10} = 2 \) implies that the data are twice as likely to have occurred under \( M_1 \) than under \( M_0 \). Jeffreys (1961), proposed a set of verbal labels to categorize the Bayes factor according to its evidential impact. This set of labels, presented in Table 1.2, facilitates scientific communication but should only be considered an approximate descriptive articulation of different standards of evidence (Kass & Raftery, 1995).

<table>
<thead>
<tr>
<th>Bayes factor</th>
<th>Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt; 100</td>
<td>Decisive evidence for ( M_1 )</td>
</tr>
<tr>
<td>30 – 100</td>
<td>Very Strong evidence for ( M_1 )</td>
</tr>
<tr>
<td>10 – 30</td>
<td>Strong evidence for ( M_1 )</td>
</tr>
<tr>
<td>3 – 10</td>
<td>Substantial evidence for ( M_1 )</td>
</tr>
<tr>
<td>1 – 3</td>
<td>Anecdotal evidence for ( M_1 )</td>
</tr>
<tr>
<td>1</td>
<td>No evidence</td>
</tr>
<tr>
<td>1/3 – 1</td>
<td>Anecdotal evidence for ( M_0 )</td>
</tr>
<tr>
<td>1/10 – 1/3</td>
<td>Substantial evidence for ( M_0 )</td>
</tr>
<tr>
<td>1/30 – 1/10</td>
<td>Strong evidence for ( M_0 )</td>
</tr>
<tr>
<td>1/100 – 1/30</td>
<td>Very Strong evidence for ( M_0 )</td>
</tr>
<tr>
<td>&lt; 1/100</td>
<td>Decisive evidence for ( M_0 )</td>
</tr>
</tbody>
</table>

Table 1.2: Evidence categories for the Bayes factor \( BF_{10} \) (Jeffreys, 1961). We replaced the label “worth no more than a bare mention” with “anecdotal”. Note that, in contrast to \( p \) values, the Bayes factor can quantify evidence in favor of the null hypothesis.

1.4 Outline

This thesis consists of two parts. In the first part, we present Bayesian alternatives to often-used frequentist null hypothesis tests, and we discuss the potential benefits of these tests over their frequentist counterparts. In the second part, we demonstrate how social science can benefit from the adoption of Bayesian methods.
Part I: Bayesian Model Selection: Theoretical

In the first chapter of the first part, Chapter two, we propose a sampling based Bayesian t test. This Savage-Dickey (SD) t test is inspired by the Jeffreys-Zellner-Siow (JZS) t test. The SD test retains the key concepts of the JZS test but is applicable to a wider range of statistical problems. The SD test allows researchers to test order-restrictions and applies to two-sample situations in which the different groups do not share the same variance.

In Chapter three we show how the so-called encompassing prior (EP) approach – which was used to facilitate Bayesian model selection for nested models with inequality constraints– naturally extends to exact equality constraints by considering the ratio of the heights for the posterior and prior distributions at the point that is subject to test (i.e., the Savage-Dickey density ratio). The EP approach generalizes the Savage-Dickey ratio method, and can accommodate both inequality and exact equality constraints. The general EP approach is found to be a computationally efficient procedure to calculate Bayes factors for nested models. However, the EP approach to exact equality constraints is vulnerable to the Borel-Kolmogorov paradox, the consequences of which warrant careful consideration.

In Chapter four, we propose a default Bayesian hypothesis test for the presence of a correlation or a partial correlation. The test is a direct application of Bayesian techniques for variable selection in regression models. We illustrate the use of the Bayesian correlation test with three examples from the psychological literature.

Then, in Chapter five, we present a Bayesian hypothesis test for Analysis of Variance (ANOVA) designs. We illustrate the effect of various g-priors on the ANOVA hypothesis test. The Bayesian test for ANOVA designs is useful for empirical researchers and for students; both groups will get a more acute appreciation of Bayesian inference when they can apply it to practical statistical problems such as ANOVA. We illustrate the use of the test with two examples, and we provide R code that makes the test easy to use.

Part II: Bayesian Model Selection: Applied

The second part of this thesis discusses how Bayesian methods can be useful for social science empirical research.

Statistical inference in psychology has traditionally relied heavily on p value significance testing. This approach to drawing conclusions from data, however, has been widely criticized, and two types of remedies have been advocated. The first proposal is to supplement p values with complementary measures of evidence such as effect sizes. The second is to replace inference with Bayesian measures of evidence such as the Bayes factor. In Chapter six, we provide a practical comparison of p values, effect sizes, and default Bayes factors as measures of statistical evidence, using 855 recently published t tests in psychology. Our comparison yields two main results: First, although p values and default Bayes factors almost always agree about what hypothesis is better supported by the data, the measures often disagree about the strength of this support; for 70% of the data sets for which the p value falls between .01 and .05, the default Bayes factor indicates that the evidence is only anecdotal. Second, effect sizes can provide additional evidence to p values and default Bayes factors. We conclude that the Bayesian approach is comparatively prudent, preventing researchers from overestimating the evidence in favor of an effect.

The next chapter, Chapter seven, is a response to a controversial article claiming evidence that people can see into the future. Does psi exist? In the controversial article, Dr. Bem conducted nine studies with over a thousand participants in an attempt to demonstrate that future events retroactively affect people’s responses. Here we discuss
several limitations of Bem’s experiments on psi; in particular, we show that the data analysis was partly exploratory, and that one-sided \( p \) values may overstate the statistical evidence against the null hypothesis. We reanalyze Bem’s data using a default Bayesian \( t \) test and show that the evidence for psi is weak to nonexistent. We argue that in order to convince a skeptical audience of a controversial claim, one needs to conduct strictly confirmatory studies and analyze the results with statistical tests that are conservative rather than liberal. We conclude that Bem’s \( p \) values do not indicate evidence in favor of precognition; instead, they indicate that experimental psychologists need to change the way they conduct their experiments and analyze their data.

In the final chapter of this thesis, Chapter eight, we discuss an agenda for purely confirmatory research. The veracity of substantive research claims hinges on the way experimental data are collected and analyzed. Here we emphasize two uncomfortable facts that threaten the core of our scientific enterprise. First, psychologists generally do not commit themselves to a method of data analysis before they see the actual data. It then becomes tempting to fine-tune the analysis to the data in order to obtain a desired result, a procedure that invalidates the interpretation of the common statistical tests. The extent of fine-tuning varies widely across experiments and experimenters but is almost impossible for reviewers and readers to gauge. Second, \( p \) values overestimate the evidence against the null hypothesis and disallow any flexibility in data collection. We propose that researchers pre-register their studies and indicate in advance the analyses they intend to conduct. Only these analyses deserve the label “confirmatory”, and only for these analyses are the common statistical tests valid. All other analyses should be labeled “exploratory”. We also propose that researchers interested in hypothesis tests use Bayes factors rather than \( p \) values. Bayes factors allow researchers to monitor the evidence as the data come in, and stop whenever they feel a point has been proven or disproven.

Part III: Appendices

Bayesian methods can also be useful without the focus on Bayes factors. In order to illustrate how to conduct Bayesian inference in psychology more generally, we include two chapters that are focused on Bayesian evaluation of mathematical models without computing the Bayes factor.

In Appendix A, we explore the statistical properties of the Expectancy Valence model. We first demonstrate the difficulty of applying the model on the level of a single participant, we then propose and implement a Bayesian hierarchical estimation procedure to coherently combine information from different participants, and we finally apply the Bayesian estimation procedure to data from an experiment designed to provide a test of specific influence.

Over the last decade, the popularity of Bayesian data analysis in the empirical sciences has greatly increased. This is partly due to the availability of WinBUGS—a free and flexible statistical software package that comes with an array of predefined functions and distributions—allowing users to build complex models with ease. For many applications in the psychological sciences, however, it is highly desirable to be able to define one’s own distributions and functions. This functionality is available through the WinBUGS Development Interface (WBDev). Appendix B illustrates the use of WBDev by means of concrete examples, featuring the Expectancy-Valence model for risky behavior in decision-making, and the shifted Wald distribution of response times in speeded choice.

Next, in Appendix C and D we provide R scripts to compute the Bayes factors for the correlation, the partial correlation and the ANOVA hypothesis test.
Then, in Appendix E we return to the controversial psi study that was reanalyzed in Chapter seven. We study the robustness of the Bayesian $t$ test, that is, we examine the extent to which the default settings yield potentially misleading results. The results show that any other setting would not have changed the qualitative conclusions that were drawn based on the default settings. Hence, our earlier conclusions (based on the default prior) are robust against alternative prior specifications.

Finally, in Appendix F, we present the complete results of the purely confirmatory study on psi mentioned in Chapter eight. All tests yield evidence in favor of the null hypothesis. In other words, all confirmatory studies yielded evidence against the hypothesis that people can look into the future.