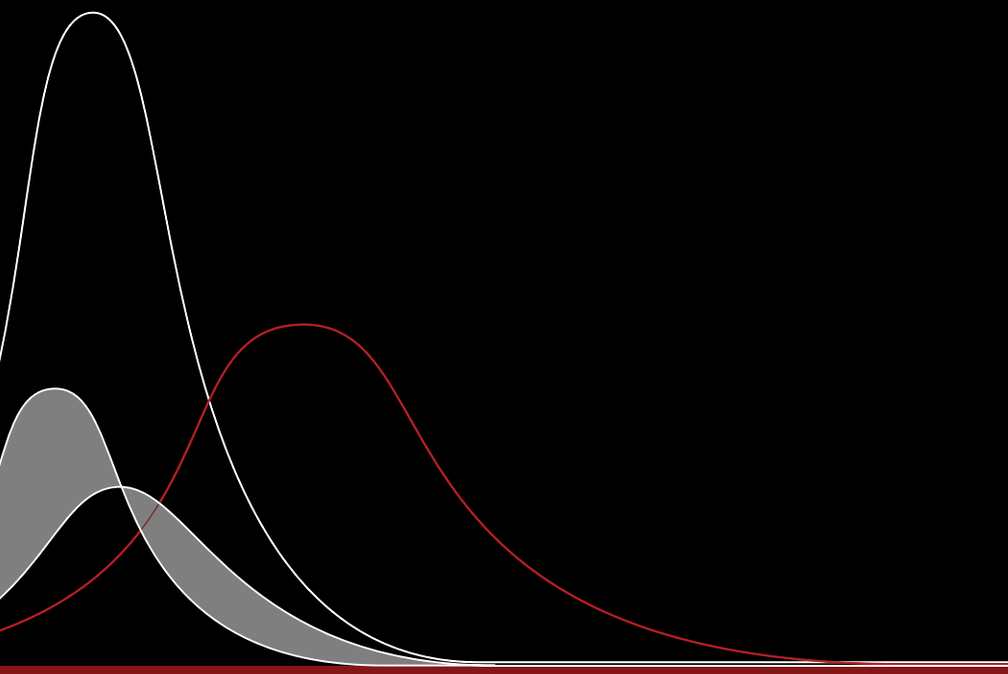
Bayesian Model Selection With Applications in Social Science

Bayesian Model Selection
With Applications in Social Science



Bayesian Model Selection
With Applications in Social Science

Ruud Wetzels

Ruud Wetzels

Ruud Wetzels

Bayesian Model Selection
With Applications in Social Science

Ruud Wetzels

Whilst part of what we perceive
comes through our senses from the object before us,
another part (and it may be the larger part)
always comes out of our own mind.

William James (1890)

ISBN 978-94-6191-404-0
Printed by Ipskamp Drukkers B.V., Enschede

BAYESIAN MODEL SELECTION WITH APPLICATIONS IN SOCIAL SCIENCE

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. D.C. van den Boom
ten overstaan van een door het college voor promoties ingestelde
commissie, in het openbaar te verdedigen in de Agnietenkapel
op woensdag, 26 september 2012, te 10:00 uur

door

Ruud Maria Wetzels
geboren te Heerlen

Promotiecommissie

Promotor:	Prof. Dr. E.–J. Wagenmakers
Copromotor:	Prof. Dr. H.L.J. van der Maas
Overige leden:	Prof. Dr. P.A.L. de Boeck
	Dr. D. Borsboom
	Dr. J.–P. Fox
	Prof. Dr. F. Tuerlinckx
	Dr. W. Vanpaemel

Faculteit der Maatschappij- en Gedragswetenschappen

Contents

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

Depression Scores	
Alcohol Group	Non-Alcohol Group
1.5	1.1
0.0	0.8
0.2	0.3
0.4	0.4
...	...

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

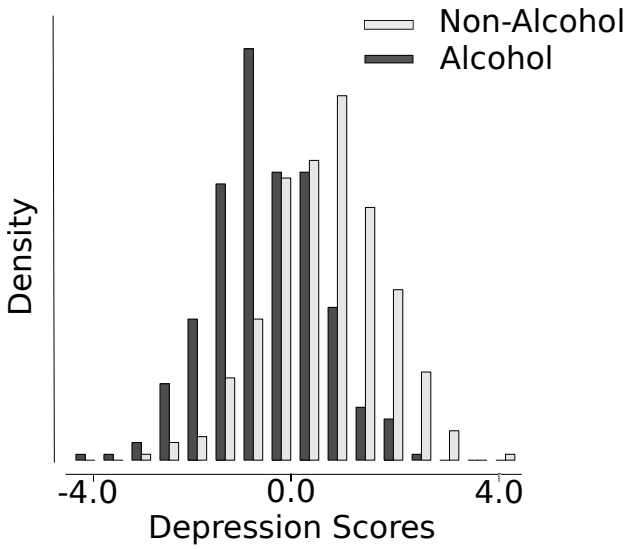


Figure 1.1: A histogram of depression scores for the alcohol group and the non-alcohol group. The mean score of the alcohol group is $\bar{X}_A = -0.5$, the mean score for the non-alcohol group is $\bar{X}_{NA} = 0.5$.

effect is deemed significant when the p value is lower than .05. In the current example, $p = .023$, and hence one can conclude that the effect is significant and reject the null hypothesis that depression scores are unrelated to alcohol consumption.

Although the above result appears to be clear-cut, the p value hypothesis test has various well-known problems, some of which we will list in the next section.

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:

$$f(\boldsymbol{\theta}_\gamma \mid \boldsymbol{Y}) = \frac{f(\boldsymbol{Y} \mid \boldsymbol{\theta}_\gamma) p(\boldsymbol{\theta}_\gamma)}{\int_\Theta f(\boldsymbol{Y} \mid \boldsymbol{\theta}_\gamma) p(\boldsymbol{\theta}_\gamma) d\boldsymbol{\theta}_\gamma},$$

where $f(\boldsymbol{Y} \mid \boldsymbol{\theta}_\gamma)$ is the likelihood function of the data $\boldsymbol{Y}$, $p(\boldsymbol{\theta}_\gamma)$ is the prior distribution of the model parameters $\boldsymbol{\theta}_\gamma$, and $f(\boldsymbol{\theta}_\gamma \mid \boldsymbol{Y})$ is the posterior distribution of the model parameters under the model $\mathcal{M}_\gamma$. This posterior distribution reflects all the information in the data about the model parameters $\boldsymbol{\theta}_\gamma$.

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).

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 $\boldsymbol{Y}$, the two models $\mathcal{M}_0$ and $\mathcal{M}_1$ are assigned prior probabilities $p(\mathcal{M}_0)$ and $p(\mathcal{M}_1)$. The ratio of the two prior probabilities defines the *prior odds*. When the data $\boldsymbol{Y}$ are observed, the prior odds are updated to *posterior odds*, which is defined as the ratio of the posterior probabilities $p(\mathcal{M}_0 \mid \boldsymbol{Y})$ and $p(\mathcal{M}_1 \mid \boldsymbol{Y})$:

$$\frac{p(\mathcal{M}_1 \mid \boldsymbol{Y})}{p(\mathcal{M}_0 \mid \boldsymbol{Y})} = \frac{p(\boldsymbol{Y} \mid \mathcal{M}_1)}{p(\boldsymbol{Y} \mid \mathcal{M}_0)} \times \frac{p(\mathcal{M}_1)}{p(\mathcal{M}_0)}, \tag{1.1}$$

or

$$\text{Posterior Odds} = \text{Bayes Factor} \times \text{Prior Odds}. \tag{1.2}$$

Equation 1.1 and 1.2 show that the change from prior odds to posterior odds is quantified by $p(\boldsymbol{Y} \mid \mathcal{M}_1)/p(\boldsymbol{Y} \mid \mathcal{M}_0)$, 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 $\mathcal{M}_1$ than under $\mathcal{M}_0$. 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).

Bayes factor			Interpretation
	$>$	100	Decisive evidence for $\mathcal{M}_1$
30	–	100	Very Strong evidence for $\mathcal{M}_1$
10	–	30	Strong evidence for $\mathcal{M}_1$
3	–	10	Substantial evidence for $\mathcal{M}_1$
1	–	3	Anecdotal evidence for $\mathcal{M}_1$
	1		No evidence
1/3	–	1	Anecdotal evidence for $\mathcal{M}_0$
1/10	–	1/3	Substantial evidence for $\mathcal{M}_0$
1/30	–	1/10	Strong evidence for $\mathcal{M}_0$
1/100	–	1/30	Very Strong evidence for $\mathcal{M}_0$
	$<$	1/100	Decisive evidence for $\mathcal{M}_0$

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.

Part I

Bayesian Model Selection: Theoretical

2 How to Quantify Support For and Against the Null Hypothesis: A Flexible WinBUGS Implementation of a Default Bayesian t test

Abstract

We propose a sampling based Bayesian t test that allows researchers to quantify the statistical evidence in favor of the null hypothesis. This Savage-Dickey (SD) t test is inspired by the Jeffreys-Zellner-Siow (JZS) t test recently proposed by Rouder, Speckman, Sun, Morey, and Iverson (2009). 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.

An excerpt of this chapter has been published as:
Wetzels, R., Raaijmakers, J. G. W., Jakab, E., & Wagenmakers, E.-J. (2009). How to Quantify Support For and Against the Null Hypothesis: A Flexible WinBUGS Implementation of a Default Bayesian t-test. *Psychonomic Bulletin & Review, 16*, 752–760.

> *Never use the unfortunate expression "accept the null hypothesis"* – Wilkinson and the Task Force on Statistical Inference (1999, p. 599).

2.1 Introduction

Popular theories are difficult to overthrow. Consider, for instance, the following hypothetical sequence of events. First, Dr. John proposes a seasonal memory model (SMM). The model is intuitively attractive and quickly gains in popularity. Dr. Smith, however, remains unconvinced and decides to put one of SMMs predictions to the test. Specifically, SMM predicts that the increase in recall performance due to the intake of glucose is more pronounced in summer than in winter. Dr. Smith conducts the relevant experiment using a within subjects design and finds the exact opposite, although the result is not significant. More specifically, Dr. Smith finds that with $n = 41$ the t value equals 0.79, which corresponds to a two-sided p value of .44 (see Table 2.1).

Clearly, Dr. Smith's data do not support SMMs prediction that the glucose-driven increase in performance is larger in summer than in winter. Instead, the data seem to suggest that the null hypothesis is plausible, and that no difference between summer and winter is evident. Dr. Smith submits his findings to the *Journal of Experimental Psychology: Learning, Memory, and the Seasons*. Three months later, Dr. Smith receives the reviews, and one of them is from Dr. John. This review includes the following comment:

> "From a null result, we cannot conclude that no difference exists, merely that we cannot reject the null hypothesis. Although some have argued that with enough data we can argue for the null hypothesis, most agree that this is only a reasonable thing to do in the face of a sizeable amount [sic] of data [which] has been collected over many experiments that control for all concerns. These conditions are not met here. Thus, the empirical contribution here does not enable readers to conclude very much, and so is quite weak (...).[1]

Table 2.1: Increase in recall performance due to intake of glucose in summer and winter, $t = 0.79$, $p = .44$ (NB: hypothetical example).

Season	N	Mean	SD
Winter	41	0.11	0.15
Summer	41	0.07	0.23

In this article, we outline a statistical method that allows Dr. Smith to quantify the evidence for the null hypothesis versus the SMM hypothesis. More generally, this method is appropriate for a test between two hypotheses, where one is nested in the other. Our work is inspired by the automatic Jeffreys-Zellner-Siow (JZS) Bayesian t test that was recently proposed by Rouder et al. (2009). Although the JZS test is able to quantify support in favor of the null hypothesis, it does not help Dr. Smith. This is because the prediction of SMM (i.e., the alternative hypothesis) is directional, one-sided, or order-restricted (e.g., Hoijtink, Klugkist, & Boelen, 2008; Klugkist, Laudy, & Hoijtink, 2005). In other words, SMM does not merely predict that the increase in recall performance differs from summer to winter, but it makes the more specific prediction that the increase

[1] This quote is taken from an actual review.

in recall performance is *larger* in summer than it is in winter. The JZS test does not directly apply to this scenario. In addition, the JZS two-sample test assumes that both groups share the same variance. When this assumption is violated, the test may no longer be reliable, a phenomenon that statisticians have studied extensively (i.e., the Behrens-Fisher problem, Kim & Cohen, 1998). To address these limitations, we have developed a flexible sampling based alternative to the JZS test. This alternative procedure, which we name the Savage-Dickey (SD) test, retains the key concepts of the JZS test but applies to a wider range of statistical problems. The computer code for the SD test and step-by-step procedures for implementing the program can be found on the first author's website, `http://www.ruudwetzels.com`.

The outline of this article is as follows. First we provide the necessary Bayesian background, and then we discuss the statistical details of Rouder et al.'s JZS test. Next we explain our own procedure, the SD test, and demonstrate by simulation that it mimics the JZS test—both for the one-sample and two-sample case. Subsequently, we outline two ways in which the SD test extends the JZS test. First, the SD test enables researchers such as Dr. Smith to test order-restricted hypotheses (i.e., one-sided t test). Second, the SD test can deal with two-sample situations in which the different groups do not share the same variance.

2.2 Bayesian Hypothesis Testing

In order to keep this article self-contained, we briefly recapitulate the basic principles of Bayesian hypothesis testing (for details see O'Hagan & Forster, 2004; Kass & Raftery, 1995; I. J. Myung & Pitt, 1997; Wasserman, 2000). First we explain the concept of *Bayes factors* and then we discuss Rouder et al.'s JZS test on which our method is based.

Bayes factors

In Bayesian inference, competing hypotheses (i.e., statistical models) are assigned probabilities. For instance, assume that you entertain two hypotheses, a null hypothesis H_0 and an alternative hypothesis H_1. Before seeing the data D, these hypotheses have *prior* probabilities $p(H_0)$ and $p(H_1)$. The ratio of these two probabilities defines the *prior odds*. When the data D come in, the prior odds are updated to *posterior odds*, which is defined as the ratio of posterior probabilities $p(H_0|D)$ and $p(H_1|D)$:

$$\frac{p(H_0|D)}{p(H_1|D)} = \frac{p(D|H_0)}{p(D|H_1)} \times \frac{p(H_0)}{p(H_1)}. \tag{2.1}$$

Equation 2.1 shows that the change from prior odds to posterior odds is quantified by $p(D|H_0)/p(D|H_1)$, the so-called *Bayes factor*. Thus, Equation 2.1 reads:

$$\text{Posterior odds} = \text{Bayes factor} \times \text{Prior odds}. \tag{2.2}$$

When the Bayes factor is, say, 14, this indicates that the data are 14 times more likely to have occurred under H_0 than under H_1, irrespective of the prior probabilities that you may assign to H_0 and H_1. When H_0 and H_1 are equally likely *a priori*, however, a Bayes factor of 14 translates directly to posterior probability—here this means that after seeing the data, H_0 is 14 times more likely than is H_1. Alternatively, one may state that the

posterior probability in favor of H_0 equals $14/15 \approx 0.93$, and the posterior probability in favor of H_1 is its complement, that is, $p(H_1|D) = 1 - p(H_0|D) \approx 0.07$.[2]

One of the attractions of the Bayes factor is that it follows the principle of parsimony: when two models fit the data equally well, the Bayes factor prefers the simple model over the more complex one (J. O. Berger & Jefferys, 1992; I. J. Myung & Pitt, 1997). This fact can be appreciated by considering how the components of the Bayes factor are calculated. Specifically, both $p(D|H_0)$ and $p(D|H_1)$ are derived by averaging the likelihood over the prior:

$$p(D|H) = \int_{\theta \in \Theta_H} f_H(D|\theta) p_H(\theta) d\theta, \tag{2.3}$$

where Θ_H denotes the parameter space under the hypothesis of interest H, f_H is the likelihood, and p_H denotes the prior distribution on the model parameters θ. Note that a complex model has a relatively large parameter space—a complex model tends to have many parameters, some of which may furthermore have a complicated functional form. Because of its large parameter space, a complex model has to spread out its prior probability quite thinly over the parameter space. As a result, the occurrence of any particular event will not greatly add to that model's credibility. A prior that is very spread out will occupy a relatively large part of the parameter space in which the likelihood for the observed data is almost zero, and this decreases the average likelihood $p(D|H)$ (I. J. Myung & Pitt, 1997).

Rouder et al.'s default Bayesian JZS t test

Consider the one-sample t test. We assume that the data are Normally distributed with unknown mean μ and unknown variance σ^2. The null hypothesis states that the mean is equal to zero, that is, $H_0 : \mu = 0$. The alternative hypothesis states that the mean is not equal to zero, that is, $H_1 : \mu \neq 0$. Denote by BF_{01} the Bayes factor in favor of H_0 over H_1. From Equation 2.3, the separate components of BF_{01} are given by:

$$p(D|H_0) = \int_0^\infty f_0(D|\mu = 0, \sigma^2) p_0(\mu = 0, \sigma^2) d\sigma^2 \tag{2.4a}$$

$$p(D|H_1) = \int_{-\infty}^\infty \int_0^\infty f_1(D|\mu, \sigma^2) p_1(\mu, \sigma^2) d\sigma^2 d\mu. \tag{2.4b}$$

These equations feature priors on the model parameters (i.e., p_0 and p_1). Rouder et al. (2009) followed Jeffreys (1961) and proposed a prior on effect size $\delta = \mu/\sigma$ instead of on the mean μ. Specifically, Rouder et al. (2009) defined a Cauchy prior on δ with location parameter 0 and scale parameter 1 (i.e., a t distribution with one degree of freedom), and a Jeffreys' prior (Jeffreys, 1961) on the variance:

$$\delta \sim \text{Cauchy}(0,1), \tag{2.5}$$

$$p(\sigma^2) \propto 1/\sigma^2, \tag{2.6}$$

where $\propto$ denotes "is proportional to". This completes the specification of H_0 and H_1. Rouder et al. (2009) then derived the following equation for the JZS Bayes factor:

[2] The absolute posterior model probabilities hold only when H_0 and H_1 are the sole two models under consideration.

$$BF_{01} = \frac{(1+\frac{t^2}{\nu})^{-(\nu+1)/2}}{\int_0^\infty (1+Ng)^{-1/2}(1+\frac{t^2}{(1+Ng)\nu})^{-(\nu+1)/2}(2\pi)^{-1/2}g^{-3/2}\exp -1/(2g)dg}, \quad (2.7)$$

where t is the t statistic for the one-sided t test, N is the number of observations, $\nu = N-1$ equals the degrees of freedom and g represents Zellner's g-prior (for a detailed explanation see Liang, Paulo, Molina, Clyde, & Berger, 2008; Zellner, 1986; Zellner & Siow, 1980).

In order to apply this Bayesian t test to two-sample designs, Equation 2.7 needs to be adjusted in three ways: (1) replace the one-sample t value with the two-sample t value; (2) calculate N as $N_X N_Y/(N_X + N_Y)$, where X and Y denote the separate groups; and (3) calculate ν as $N_X + N_Y - 2$.

Now recall the data collected by Dr. Smith (see Table 2.1). Dr. Smith used a within-subject design, and hence a one-sample t test on the difference scores is appropriate. From the Bayes factor calculator provided on Rouder's website[3] we obtain a Bayes factor of 6.08—this means that the data are about 6 times more likely under the null hypothesis than under the alternative hypothesis. When we assume that both hypotheses are equally likely *a priori*, we can compute $p(H_0|D)$, the posterior probability for the null hypothesis, as $6.08/7.08 \approx .86$.

Unfortunately, the test developed by Rouder and colleagues does not apply to the problem that confronts Dr. Smith. As mentioned earlier, the SMM predicts that the effect will go in a specific direction—a direction other than the one that is observed in Dr. Smith's experiment. In order to calculate the Bayes factors that are appropriate for a one-sided test, we have developed a sampling based alternative test.[4]

2.3 SD: An MCMC Sampling Based t Test

Calculation of the Savage-Dickey (SD) t test involves four steps. The associated computer programs can be found on the first author's website.

Step 1. Rescaling the Data

Prior to the analyses, we rescale the data such that one group has mean 0 and standard deviation 1. This scaling does not affect the test statistic. For the data from Dr. Smith, for instance, the "summer mean" of 0.07 is subtracted from all observations, both in the winter condition and in the summer condition. Next, all observations are divided by the "summer standard deviation". The main advantage of this rescaling procedure is that the prior distributions for the parameters hold regardless of the scale of measurement: for our Bayesian SD test, it does not matter whether, say, response times are measured in seconds or in milliseconds.

Step 2. Defining Prior Distributions

We follow Rouder et al. and use a Cauchy(0,1) prior for effect size δ. For the standard deviation σ we use a half-Cauchy(0,1) (Gelman & Hill, 2007), that is, a Cauchy(0,1) distribution that is defined only for positive numbers. This choice for σ is reasonably

[3] `http://pcl.missouri.edu/bayesfactor`.

[4] There may or may not be an analytical solution to the order-restricted problem, and here we do not attempt to derive such a solution. Instead, the goal is to illustrate the flexibility of the SD test using the order-restricted hypothesis test as an example.

uninformative, but—in contrast to Jeffrey's prior in Equation 2.6—the distribution is still proper (i.e., the area under the distribution is finite).[5] For the two-sample t test, we specify a Cauchy(0,1) prior for the grand mean μ.

Step 3. Obtaining Posteriors using WinBUGS

The WinBUGS program[6] (D. J. Lunn, Thomas, Best, & Spiegelhalter, 2000) uses built–in Markov chain Monte Carlo techniques (MCMC; Gamerman & Lopes, 2006) to obtain samples from posterior distributions. After specifying the SD model in WinBUGS, the posterior distribution for effect size δ can be approximated to any desired degree of accuracy by increasing the number of samples. Because the SD model is relatively simple, we can draw as many as one million samples in a matter of minutes.

Step 4. Calculating Bayes factors using the Savage-Dickey Density Ratio

To obtain the Bayes factor, we use a method that is simple, intuitive, and flexible; the Savage-Dickey Density Ratio Method (S–D, e.g. Dickey & Lientz, 1970,O'Hagan & Forster, 2004, pp.174–177, Verdinelli & Wasserman, 1995). This method applies only to nested model comparisons but it greatly simplifies the computation of the Bayes factor: the only information that is required is the height of the prior and the posterior distributions for the parameter of interest (i.e., δ) under the alternative hypothesis H_1 at the point that is subject to test. The reader who is not interested in the mathematical derivation may safely skip to Equation 2.10.

Let δ be the parameter of interest and σ the nuisance parameter. We assume, as is reasonable in many cases, that the conditional density for δ is continuous at $\delta = 0$, such that $\lim_{\delta \to 0} p(\sigma^2|H_1, \delta) = p(\sigma^2|H_0)$. This means that the prior for the nuisance parameter in the complex model, conditional on $\delta \to 0$, equals the prior for the nuisance parameters in the simple model for which $\delta = 0$ by definition. We can then write $p(\sigma^2|H_1, \delta = 0) = p(\sigma^2|H_0)$, an equality that holds automatically when the prior distributions are specified to be independent.

The foregoing allows us to simplify the marginal likelihood for H_0 as follows:

$$\begin{aligned} p(D|H_0) &= \int_0^\infty f(D|H_0, \sigma^2) p(\sigma^2|H_0) d\sigma^2 \\ &= \int_0^\infty f(D|H_1, \sigma^2, \delta = 0) p(\sigma^2|H_1, \delta = 0) d\sigma^2 \\ &= p(D|H_1, \delta = 0). \end{aligned} \tag{2.8}$$

We now apply Bayes' rule to the results of Equation 2.8 and obtain

$$p(D|H_0) = p(D|H_1, \delta = 0) = \frac{p(\delta = 0|H_1, D) p(D|H_1)}{p(\delta = 0|H_1)}. \tag{2.9}$$

Dividing both sides of Equation 2.9 by $p(D|H_1)$ results in

$$BF_{01} = \frac{p(D|H_0)}{p(D|H_1)} = \frac{p(\delta = 0|H_1, D)}{p(\delta = 0|H_1)}. \tag{2.10}$$

[5] This is helpful as WinBUGS does not allow the specification of improper priors. In any case, because sigma is a nuisance parameter in this model, the prior for sigma has a negligible effect on the calculation of the Bayes factor.

[6] WinBUGS is easy to learn and is supported by a large community of active researchers, see `http://www.mrc-bsu.cam.ac.uk/bugs/`.

This result is generally known as the Savage-Dickey density ratio (Dickey & Lientz, 1970; O'Hagan & Forster, 2004) and it shows that the Bayes factor equals the ratio of the posterior and prior ordinate under H_1 at the point of interest (i.e., $\delta = 0$). Note that there is no need to integrate out any model parameters, that the only distribution that matters is the one for the parameter of interest δ, and that the only hypothesis that needs to be considered is H_1. These are considerable simplifications compared to the standard procedure (cf. Equation 2.4).

Thus, Equation 2.10 shows that all that is required to compute the Bayes factor is the height of the prior and posterior distributions for δ at $\delta = 0$. The height of the prior distribution at $\delta = 0$ can be immediately computed from the Cauchy(0,1) distribution. The height of the posterior distribution at $\delta = 0$ can be easily estimated from the MCMC samples, for instance by applying a nonparametric density estimator (e.g., Stone, Hansen, Kooperberg, & Truong, 1997) or a Normal approximation to the posterior (i.e., parametric density estimation). The Normal approximation is motivated by the Bayesian Central Limit Theorem (Carlin & Louis, 2000, pp. 122–124) which states that under general regularity conditions, all posterior distributions tend to a Normal distribution as the number of observations grows large.

Our experience with the SD test suggests that the difference between nonparametric and parametric estimation is negligible. In the work reported here, we choose to use the Normal approximation because it is computationally more efficient. However, it is prudent to always plot the posterior distributions and check whether the posterior ordinate at $\delta = 0$ is estimated correctly. For practical applications, we also advise the user to use both the nonparametric and the parametric estimator and confirm that they yield approximately the same result.

2.4 The One-Sample SD t Test: Comparison to Rouder et al.

The one-sample t test is used to test whether the population mean of one particular sample of observations is equal to zero or not. In experimental psychology, the one-sample t test is often used for within-subjects designs, in which the scores for two conditions can be reduced to a single difference score.

In order to clarify the structure of the one-sample t test we use graphical model notation (e.g., Gilks, Thomas, & Spiegelhalter, 1994; Lauritzen, 1996; Lee, 2008; Spiegelhalter, 1998). In this notation, nodes represent variables of interest, and the graph structure is used to indicate dependencies between the variables, with children depending on their parents. Double borders indicate that the variable under consideration is deterministic (i.e., they are calculated without noise from other variables) rather than stochastic. Finally, observed variables are shaded and unobserved variables are not shaded. The graphical model for the one-sample t test is shown in Figure 2.1.

In the graphical model, X represents the observed data, distributed according to a Normal distribution with mean μ_X and a variance σ_X^2. Because $\delta = \mu_X/\sigma_X$, μ_X is given by $\mu_X = \delta \times \sigma_X$. The null hypothesis puts all prior mass for δ on a single point, that is, $H_0 : \delta = 0$, whereas the alternative hypothesis assumes that δ is Cauchy(0,1) distributed, $H_1 : \delta \sim$ Cauchy(0,1). It is relatively straightforward to implement this graphical model in WinBUGS, obtain samples from the posterior distribution for δ, and carry out the Savage-Dickey test.

Because our SD t test is based on a sampling-based procedure that relies on the convergence of a stochastic process, it is desirable to verify whether the results of the SD test coincide with those from the JZS test, which is based on an analytical solution.

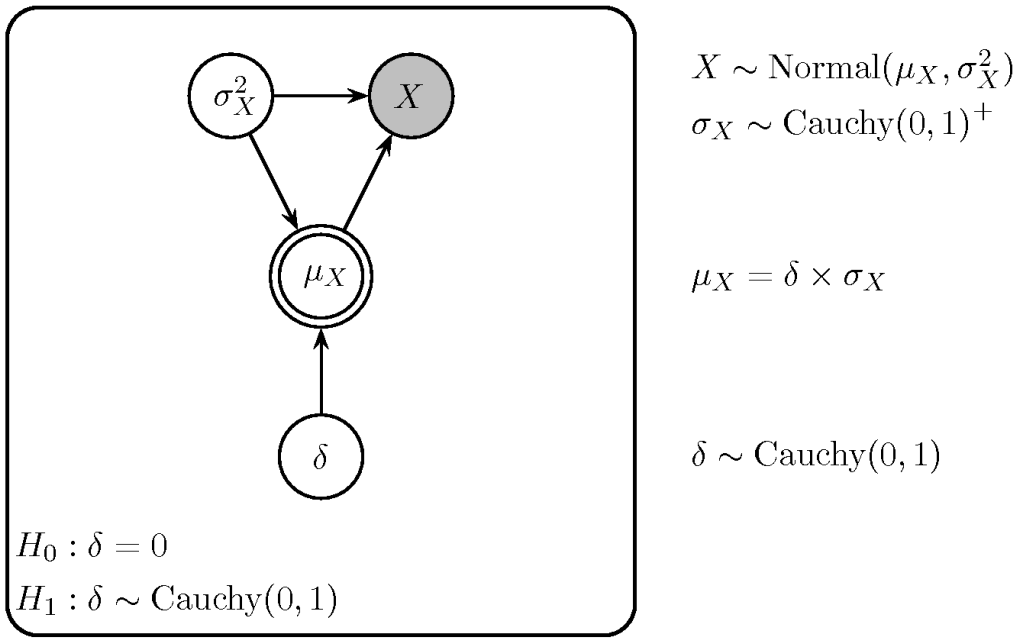


Figure 2.1: Graphical model for the SD one-sample t test. Cauchy(0,1)$^+$ denotes the half-Cauchy(0,1) defined for positive numbers only.

This verification was carried out by means of a simulation study, the results of which are shown in Figure 2.2. We simulated 100 data sets by systematically increasing the difference between the group means to yield a set of 100 different t values. For each of the 100 data sets we then compared the Bayes factor calculated by the JZS-test to the SD Bayes factor. For all panels, the x-axis gives the t-statistic, and the y-axis gives the associated posterior probability for the null hypothesis, $p(H_0|D)$, derived from the Bayes factor under the assumption that H_0 and H_1 are equally likely *a priori*. Each panel shows the overlap between the JZS test and the SD test for a specific sample size (i.e., $N \in \{20, 40, 80, 160\}$, based on 100 simulated data sets. The results demonstrate that for the one-sample scenario, the SD test closely mimics the JZS test.

2.5 The Two-Sample SD t Test: Comparison to Rouder et al.

The two-sample t test is used to test whether the population means of two independent samples of observations are equal to each other or not. In experimental psychology, the two-sample t test is often used for between-subjects designs.

The graphical model for the two-sample t test is shown in Figure 2.3. The graphical model shows that X and Y represent the two groups of observed data. Both X and Y are distributed according to a Normal distribution with shared variance σ^2. The mean of X is given by $\mu + \alpha/2$, and the mean of Y is given by $\mu - \alpha/2$.

Because $\delta = \alpha/\sigma$, α is given by $\alpha = \delta \times \sigma$. As for the one-sample scenario, the null hypothesis puts all prior mass for δ on a single point, that is, $H_0 : \delta = 0$, whereas the alternative hypothesis assumes that δ is Cauchy(0,1) distributed, $H_1 : \delta \sim$ Cauchy(0,1).

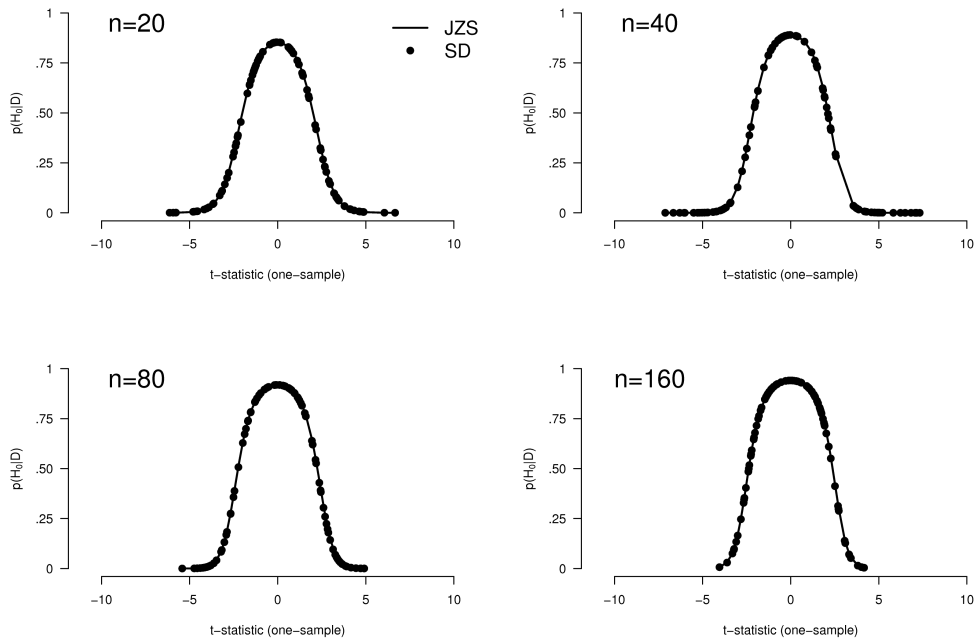


Figure 2.2: Comparison between the one-sample SD values and JZS values, for various sample sizes. The black dots represent the SD values and the solid line represents the JZS values.

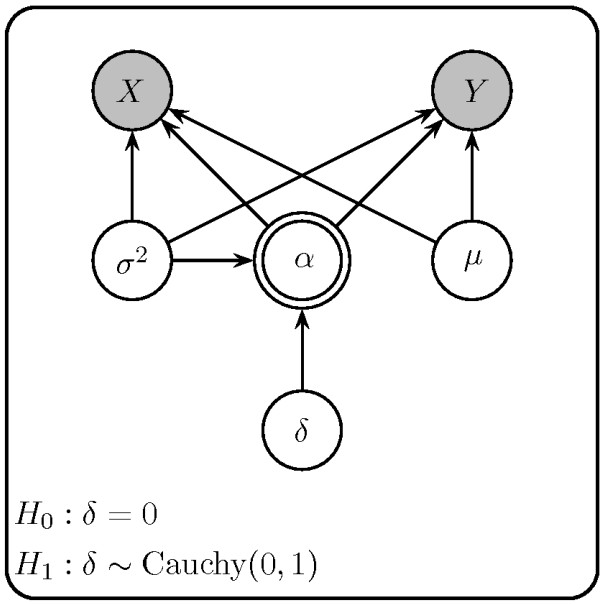
To compare this SD test to Rouder et al.'s JZS test we conducted a simulation study, identical to the one-sample scenario in all respects except for the number of groups. The results of this simulation study are shown in Figure 2.4. The results demonstrate that for the two-sample scenario, the SD test closely mimics the JZS test.

2.6 Extension 1: Order-Restrictions

Recall once again the experiment by Dr. Smith (see Table 2.1). The SMM predicted that the effect of glucose would be larger in summer than in winter. We now show how the SD test can be used to test such order-restricted hypotheses, allowing Dr. Smith to quantify exactly the extent to which the data support the null hypothesis versus the alternative SMM hypothesis.

The top panel of Figure 2.5 shows the unrestricted prior and posterior distributions for δ for the data from Dr. Smith. Negative values of δ indicate that the effect of glucose is larger in summer than in winter. From the Savage-Dickey method we can compute the Bayes factor in favor of $H_0 : \delta = 0$ versus the unrestricted alternative $H_1 : \delta \neq 0$, instantiated as $\delta \sim \text{Cauchy}(0,1)$. Note that the result—$BF_{01} = 6.08$—is identical to the Bayes factor that is obtained from the JZS test: the data are about six times more likely under H_0 than under H_1.

The middle panel of Figure 2.5 shows the SD test that applies to the prediction that Dr.



$$X \sim \text{Normal}(\mu + \alpha/2, \sigma^2)$$
$$Y \sim \text{Normal}(\mu - \alpha/2, \sigma^2)$$

$$\sigma \sim \text{Cauchy}(0, 1)^+$$
$$\mu \sim \text{Cauchy}(0, 1)$$
$$\alpha = \sigma \times \delta$$

$$\delta \sim \text{Cauchy}(0, 1)$$

Figure 2.3: Graphical model for the SD two-sample t test. Cauchy(0,1)$^+$ denotes the half-Cauchy(0,1) defined for positive numbers only.

Smith seeks to test, that is, $H_0 : \delta = 0$ versus the order-restricted hypothesis $H_1 : \delta < 0$, instantiated as $\delta \sim$ Cauchy(0,1)$^-$, a half-Cauchy(0,1) distribution that is defined only for negative numbers. In order to calculate the height of the order-restricted posterior distribution at $\delta = 0$, we focus solely on that part of the unrestricted posterior for which $\delta < 0$. After renormalizing, we obtain a truncated but proper posterior distribution that ranges from $\delta = -\infty$ to $\delta = 0$. Figure 2.5 shows both the half-Cauchy(0,1) prior (solid line) and the truncated posterior (dashed line). The Savage-Dickey ratio at $\delta = 0$ yields a Bayes factor of $BF_{01} = 13.75$. This means that the data are almost 14 times more likely under H_0 than under the order-restricted H_1 that is associated with SMM. When H_0 and H_1 are equally likely *a priori*, the posterior probability in favor of the null hypothesis is about $13.75/14.75 \approx .93$, which is considered "positive evidence" for the null hypothesis (Raftery, 1995; Wagenmakers, 2007).

For completeness, the bottom panel of Figure 2.5 shows the SD test for the alternative order-restriction. In this case, we seek to test $H_0 : \delta = 0$ versus $H_1 : \delta > 0$, instantiated as $\delta \sim$ Cauchy$(0, 1)^+$, a half-Cauchy(0,1) distribution that is defined only for positive numbers. The Savage-Dickey density ratio yields a Bayes factor of $BF_{01} = 3.91$, which indicates that the data are almost 4 times more likely under H_0 than under H_1.

2.7 Extension 2: Variances Free to Vary in the Two-Sample t Test

For the two-sample scenario, the JZS test assumes that the separate samples share a common unknown variance. When this assumption is false and both groups have unequal numbers of observations, results of the JZS t test should be interpreted with care.

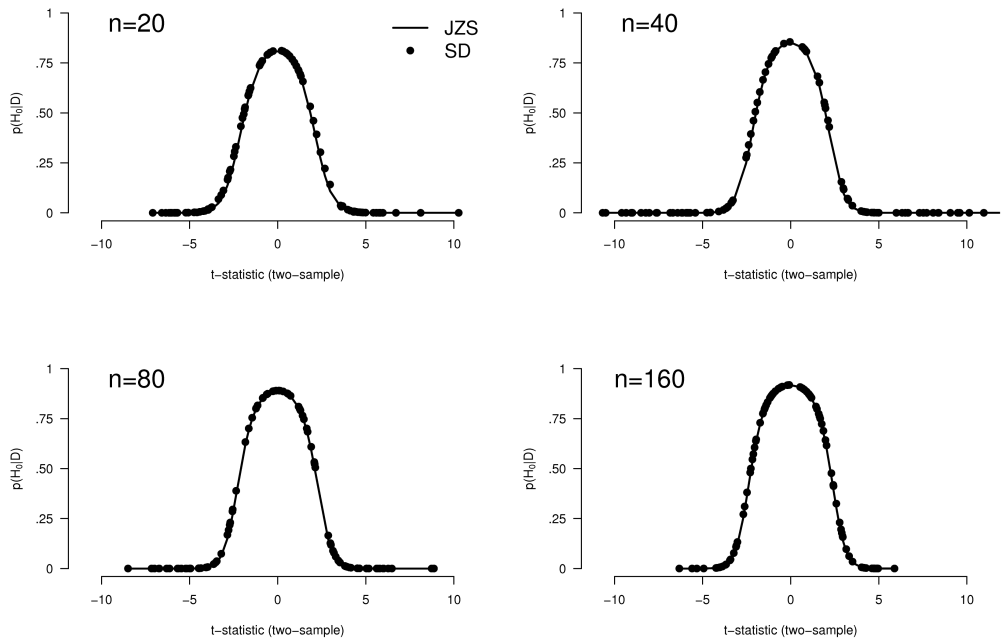


Figure 2.4: Comparison between the two-sample SD values and JZS values, for various sample sizes. The black dots represent the SD values and the solid line represents the JZS values.

This complication (i.e., testing for the difference of two Normal means with unequal variances) is known as the Behrens-Fisher problem, and it is one of the oldest problems in statistics. Within the paradigm of p value hypothesis testing, several solutions to the Behrens-Fisher problem have been proposed (Kim & Cohen, 1998). These solutions (i.e., corrections for unequal variances) have been implemented in popular statistical software packages such as SPSS and R.

In order to address the Behrens-Fisher problem, we adjusted the SD test in two ways. First, as illustrated in Figure 2.6, each of the two groups now has its own variance. Second, the previous relation $\alpha = \delta \times \sigma$ no longer holds, as we now have two σ parameters. We use a standard solution and calculate the pooled standard deviation (Hedges, 1981):

$$\alpha = \delta \times \sqrt{\frac{(\sigma_1^2 \times (n_1 - 1)) + (\sigma_2^2 \times (n_2 - 1))}{n_1 + n_2 - 2}}. \tag{2.11}$$

After implementing these changes, calculation of the Bayes factor proceeds in the same fashion as before.

To illustrate the behavior of the separate variance SD Bayes factors, we follow Moreno, Bertolino, and Racugno (1999) and apply the tests to hypothetical data from Box and Tiao (1973, p. 107). These data have the following properties: $n_1 = 20$, $\text{var}_1 = 12$, $n_2 = 12$, and $\text{var}_2 = 40$. As can be seen from Table 2.2, the support for the null hypothesis decreases as the difference in group means increases. The separate variance

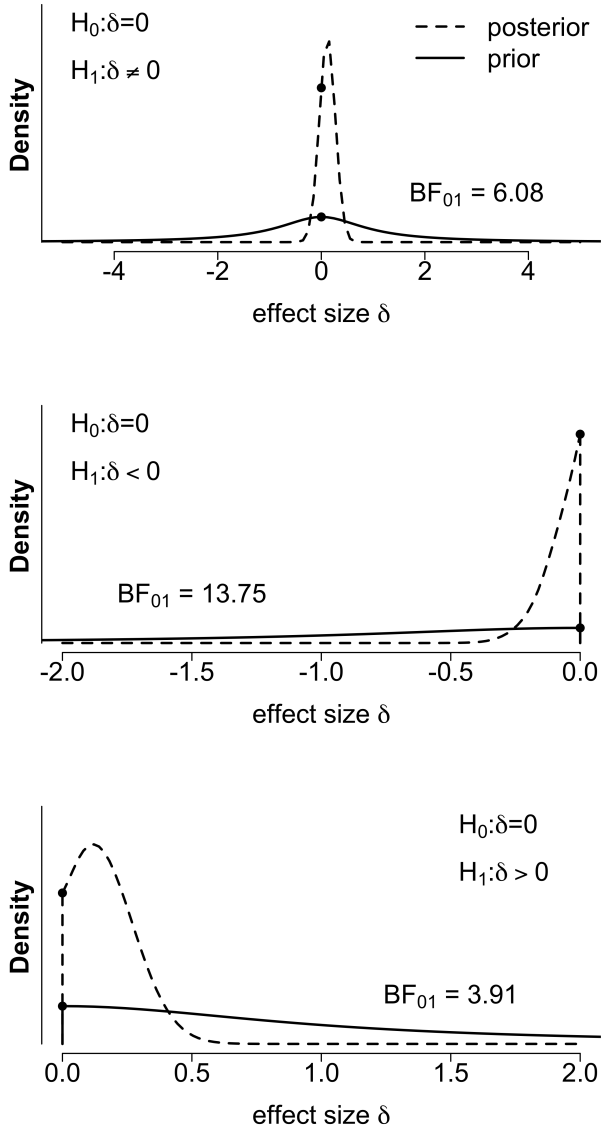
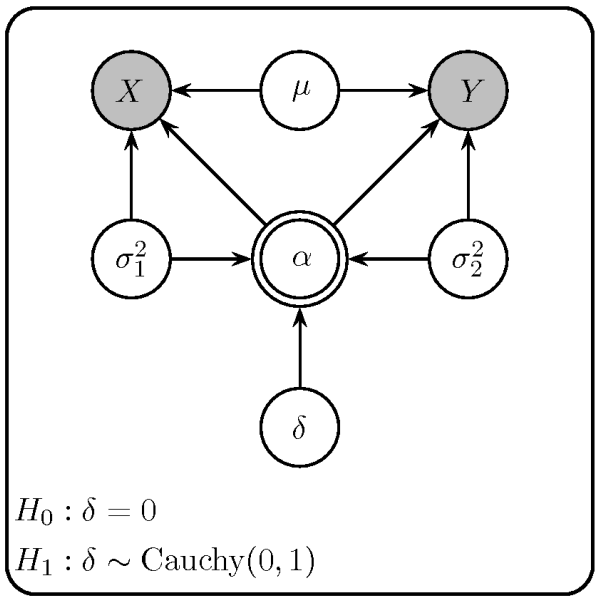


Figure 2.5: The prior and posterior distributions of effect size δ, based on the data from Dr. Smith (Table. 1). The top panel illustrates the unrestricted SD test, the middle panel illustrates the order-restricted test associated with the SMM, and the bottom panel illustrates the SD test for the alternative order-restriction. The dots mark the height of the prior and posterior distributions at δ=0.



$$X \sim \text{Normal}(\mu + \alpha/2, \sigma_1^2)$$
$$Y \sim \text{Normal}(\mu - \alpha/2, \sigma_2^2)$$
$$\mu \sim \text{Cauchy}(0, 1)$$
$$\sigma_1 \sim \text{Cauchy}(0, 1)^+$$
$$\sigma_2 \sim \text{Cauchy}(0, 1)^+$$
$$\delta \sim \text{Cauchy}(0, 1)$$

Figure 2.6: Graphical model for Rouder's default Bayesian two-sided t test with unequal variances.

SD test tends to favor the null hypothesis more than does the shared variance SD test, although the difference is small. The intrinsic Bayes factor (i.e., a default Bayes factor that uses minimal training samples and uninformative priors, J. O. Berger & Pericchi, 1996; Moreno et al., 1999) supports the null hypothesis the most. A more detailed treatment of the Behrens-Fisher problem is beyond the scope of the present article—we include it here only to highlight the flexibility of the SD test.

Table 2.2: Comparison of SD Bayes factors to the intrinsic Bayes factor for hypothetical data reported in Box and Tiao (1973, p. 107) and analyzed in Moreno et al. (1999). Note. $BF_{01}^{\text{SD1}\sigma}$ denotes the SD Bayes factor using a shared variance, $BF_{01}^{\text{SD2}\sigma}$ denotes the SD Bayes factor using two separate variances, and BF_{01}^{I} denotes the intrinsic Bayes factor reported by Moreno et al..

$\bar{X}$-$\bar{Y}$	$BF_{01}^{\text{SD1}\sigma}$	$BF_{01}^{\text{SD2}\sigma}$	BF_{01}^{I}
0.00	3.93	3.36	5.00
2.20	2.08	2.16	2.86
4.22	0.45	0.81	0.76
5.00	0.21	0.51	0.40
10.0	<0.02	<0.02	<0.02

2.8 Summary and Conclusion

In this paper we developed a "Savage-Dickey" Bayesian t test that extends the Bayesian JZS t test recently proposed by Rouder et al. (2009). Our sampling-based SD test can handle order-restrictions and addresses the situation in which two groups have unequal variance.

One of the advantages of the SD test is its flexibility—for instance, it would be trivial to replace the default priors with priors that are informed by previous experiments or detailed expert knowledge about the problem at hand. We chose to use the Cauchy(0,1) prior for effect size δ, as proposed by Rouder et al., but many more prior distributions are possible. For example, Killeen (2007) argues that, based on extensive research in social psychology (Richard, Bond, & Stokes-Zoota, 2003), the distribution of effect sizes is Normally distributed with variance 0.3.

Another advantage of the SD test, and Bayesian methods in general, is that they allow for *sequential inference*. As stated by Edwards, Lindman, and Savage (1963, p. 193), "the rules governing when data collection stops are irrelevant to data interpretation. It is entirely appropriate to collect data until a point has been proven or disproven, or until the data collector runs out of time, money, or patience". More concretely, this means that one can apply the SD t test and monitor the resulting Bayes factor after every new participant, stopping data collection whenever the evidence is sufficiently compelling. Note that within the paradigm of p-value hypothesis testing, such practice amounts to cheating; with enough time, money, and patience, "optional stopping" is guaranteed to yield a significant result (for a discussion see Wagenmakers, 2007).

Here we have limited ourselves to the t test. Nevertheless, the Savage-Dickey idea is quite general and it can facilitate Bayesian hypothesis testing for a wide range of relatively complex mathematical process models such as the Expectancy-Valence model for the Iowa Gambling Task (Busemeyer & Stout, 2002; Wetzels, Vandekerckhove, Tuerlinckx, & Wagenmakers, in press), the Ratcliff diffusion model for response times and accuracy (Vandekerckhove, Tuerlinckx, & Lee, 2008; Wagenmakers, 2009), models of categorization such as ALCOVE (J. K. Kruschke, 1992) or GCM (Nosofsky, 1986), multinomial processing trees (Batchelder & Riefer, 1999), the ACT-R model (Weaver, 2008), and many more. Another exciting possibility is to apply the Savage-Dickey method to facilitate Bayesian hypothesis testing in hierarchical models (i.e., models with random effects for subjects or items) such as those advocated by Rouder and others (Rouder, Lu, Morey, Sun, & Speckman, 2008; Rouder & Lu, 2005; Rouder et al., 2007; Shiffrin, Lee, Kim, & Wagenmakers, 2008).

For example, one might wish to study the effect of an antidepressant on the parameters of the Ratcliff diffusion model. Specifically, the hypothesis of interest may hold that the antidepressant decreases response caution a. This means that $H_0 : \delta = 0$ and $H_1 : \delta > 0$, where δ indicates the difference in response caution ($\delta = a_{off} - a_{on}$) between people that are either on or off medication. Standard approaches for computing the Bayes factor require that one integrates out all the other parameters of the diffusion model (i.e., drift rate, non-decision time, starting point, the probability of a response contaminant, and the across-trial variabilities), separately for H_0 and H_1. In contrast, the Savage-Dickey approach only requires one to estimate the height of the posterior distribution at $\delta = 0$—a considerable simplification.

In closing, we agree with Rouder et al. (2009) that many scientific hypotheses are formulated in terms of invariances, and that invariances can be formulated in terms of statistical null hypotheses (Wagenmakers, Lee, Lodewyckx, & Iverson, 2008). To quantify the statistical evidence in favor of such substantive null hypotheses, we need to move away

from p value hypothesis testing (with which one can only "fail to reject" a null hypothesis) and move toward Bayesian hypothesis testing. In this paper, we have discussed a related problem of considerable scientific importance: a substantive hypothesis (i.e., the SMM) makes a specific prediction, and falsification of the theory requires that one is able to quantify the support in favor of the null hypothesis.

We believe that Bayesian hypothesis testing not only provides a coherent framework to quantify knowledge and uncertainty, but that it also addresses the kinds of questions that experimental psychologists would like to see answered. Bayesian t tests such as Rouder et al.'s JZS test and our SD test are the first steps towards a more rational and informative method for testing statistical hypotheses in psychology.

3 An Encompassing Prior Generalization of the Savage-Dickey Density Ratio

Abstract

An encompassing prior (EP) approach to facilitate Bayesian model selection for nested models with inequality constraints has been previously proposed. In this approach, samples are drawn from the prior and posterior distributions of an encompassing model that contains an inequality restricted version as a special case. The Bayes factor in favor of the inequality restriction then simplifies to the ratio of the proportions of posterior and prior samples consistent with the inequality restriction. This formalism has been applied almost exclusively to models with inequality or "about equality" constraints. It is shown that the EP approach 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.

An excerpt of this chapter has been published as:
Wetzels, R., Grasman, R.P.P.P, & Wagenmakers, E.-J. (2009). An Encompassing Prior Generalization of the Savage-Dickey Density Ratio. *Computational Statistics & Data Analysis, 54*, 2094–2102.

3.1 Introduction

In this article we focus on Bayesian model selection for nested models. Consider, for instance, a parameter vector $\theta = (\psi, \phi) \in \Theta \subseteq \Psi \times \Phi$ and suppose we want to compare an encompassing model M_e to a restricted version $M_1 : \psi = \psi_0$. Then, after observing the data D, the Bayes factor in favor of M_1 is

$$BF_{1e} = \frac{p(D \mid M_1)}{p(D \mid M_e)} = \frac{\int p(D|\psi = \psi_0, \phi)p(\psi = \psi_0, \phi)d\phi}{\iint p(D \mid \psi, \phi)p(\psi, \phi)d\psi d\phi}.$$

Thus, the Bayes factor is the ratio of the marginal likelihoods of two competing models; alternatively, the Bayes factor can be conceptualized as the change from prior model odds $p(M_1)/p(M_e)$ to posterior model odds $p(M_1 \mid D)/p(M_e \mid D)$ (Kass & Raftery, 1995). The Bayes factor quantifies the evidence that the data provide for one model versus another, and as such it represents "the standard Bayesian solution to the hypothesis testing and model selection problems" (Lewis & Raftery, 1997, p. 648).

Unfortunately, for most models the Bayes factor cannot be obtained in analytic form. Several methods have been proposed to estimate the Bayes factor numerically (see Gamerman and Lopes (2006, Chap. 7) for a description of 11 such methods). Nevertheless, calculation of the Bayes factor often remains a computationally complicated task.

Here we first describe an encompassing prior (EP) approach that was recently proposed by Hoijtink, Klugkist, and colleagues (Klugkist, Kato, & Hoijtink, 2005; Klugkist, Laudy, & Hoijtink, 2005; Hoijtink et al., 2008). The EP approach applies to nested models and virtually eliminates the computational complications inherent in most other methods. Next we show that the EP approach is a generalization of the Savage-Dickey density ratio. Finally, we discuss the Borel-Kolmogorov paradox and examine the implications of this paradox for the EP approach.

3.2 Bayes Factors from the Encompassing Prior Approach

For concreteness, consider two Normally distributed random variables with means μ_1 and μ_2, and common standard deviation σ. We focus on the following hypotheses:

$$M_e : \mu_1, \mu_2; \sigma,$$

$$M_1 : \mu_1 > \mu_2; \sigma,$$

$$M_2 : \mu_1 \approx \mu_2; \sigma,$$

$$M_3 : \mu_1 = \mu_2; \sigma.$$

In the encompassing model M_e, all parameters are free to vary. Models M_1, M_2, and M_3 are nested in M_e and stipulate particular restrictions on the means; specifically, M_1 features an inequality constraint, M_2 features an "about equality" constraint, and M_3 features an exact equality constraint. We now deal with these in turn.

Computing Bayes Factors for Inequality Constraints

Suppose we compare two models, an encompassing model M_e and an inequality constrained model M_1. We denote the prior distributions under M_e by $p(\psi, \phi|M_e)$, where ψ

is the parameter vector of interest (e.g., μ_1 and μ_2 in the earlier example) and ϕ is the parameter vector of nuisance parameters (e.g., σ in the earlier example).

Then, the prior distribution of the parameters under model M_1 can be obtained from $p(\psi, \phi|M_e)$ by restricting the parameter space of ψ:

$$p(\psi, \phi|M_1) = \frac{p(\psi, \phi|M_e) I_{M_1}(\psi, \phi)}{\iint p(\psi, \phi|M_e) I_{M_1}(\psi, \phi) d\psi d\phi}. \tag{3.1}$$

In Equation 3.1, $I_{M_1}(\psi, \phi)$ is the indicator function of model M_1. This means that $I_{M_1}(\psi, \phi) = 1$ if the parameter values are in accordance with the constraints imposed by model M_1, and $I_{M_1}(\psi, \phi) = 0$, otherwise. Note that this specification of priors is only valid under the assumption that the nuisance parameters in M_e and M_1 fulfill exactly the same role (for a debate see Consonni and Veronese (2008); Del Negro and Schorfheide (2008)).

Under the above specification of priors, Klugkist and Hoijtink (2007) showed that the Bayes factor BF_{1e} can be easily obtained by drawing values from the posterior and prior distribution for M_e:

$$BF_{1e} = \frac{\frac{1}{m}\sum_{i=1}^{m} I_{M_1}(\psi^{(i)}, \phi^{(i)}|D, M_e)}{\frac{1}{n}\sum_{j=1}^{n} I_{M_1}(\psi^{(j)}, \phi^{(j)}|M_e)}, \tag{3.2}$$

where m represents the total number of MCMC samples for the posterior of ψ, and n represents the total number of MCMC samples for the prior of ψ. The numerator represents the proportion of M_e's *posterior* samples for ψ that obey the constraint imposed by M_1, and the denominator represents the proportion of M_e's *prior* samples for ψ that obey the constraint imposed by M_1.

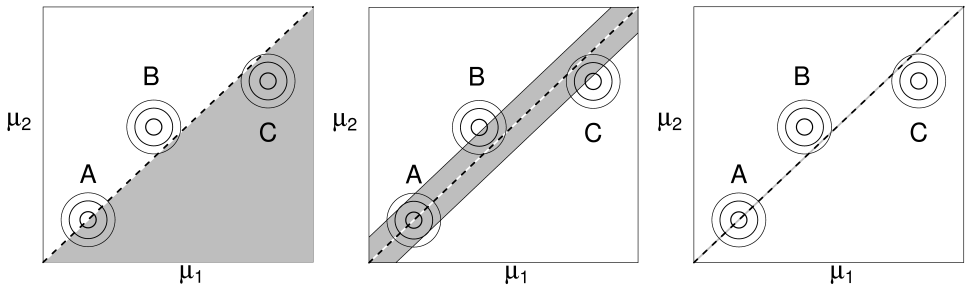
To illustrate, consider again our initial example in which $M_e : \mu_1, \mu_2; \sigma$ and $M_1 : \mu_1 > \mu_2; \sigma$. Figure 3.1a shows the joint parameter space for μ_1 and μ_2; for illustrative purposes, we assume that the joint prior is uniform across the parameter space. In Figure 3.1a, half of the prior samples are in accordance with the constraints imposed by M_1. Figure 3.1a also shows three possible encompassing posterior distributions: A, B, and C. In case A, half of the posterior samples are in accordance with the constraint, and this yields $BF_{1e} = 1$. In case B, very few samples are in accordance with the constraint, and this yields a Bayes factor BF_{1e} that is close to zero (i.e., very large support against M_1). In case C, almost all samples are in accordance with the constraint, and this yields a Bayes factor BF_{1e} that is close to 2.

Bayes Factors for About Equality Constraints

In the EP approach, the Bayes factor for about equality constraints can be calculated in the same manner as for inequality constraints. To illustrate, consider our example in which $M_e : \mu_1, \mu_2; \sigma$ and $M_2 : \mu_1 \approx \mu_2; \sigma$. Figure 3.1b shows as a gray area the proportion of prior samples that are in accordance with the constraints imposed by M_2, which in this case equals about .20. Note that $\mu_1 \approx \mu_2$ means $|\mu_1 - \mu_2| < \varepsilon$. The choice for ε defines the size of the parameter space that is allowed by the constraint.

Now consider the three possible encompassing posterior distributions shown in Figure 3.1b. In case A, about 80% of the posterior samples are in accordance with the constraint, and this yields a Bayes factor $BF_{2e} = .8/.2 = 4$. In case B and C, slightly less than half of the samples, about 40%, are in accordance with the constraint, and this yields a Bayes factor $BF_{2e} = .4/.2 = 2$.

As before, the Bayes factors are calculated with relative ease—all that is required are prior and posterior samples from the encompassing model M_e.



(a) $M_1 : \mu_1 > \mu_2$ (b) $M_2 : \mu_1 \approx \mu_2$ (c) $M_3 : \mu_1 = \mu_2$

Figure 3.1: The encompassing prior approach for inequality, about equality, and exact equality constraints. For illustrative purposes, we assume that the encompassing prior is uniform over the parameter space. The gray area represents the part of the encompassing parameter space that is in accordance with the constraints imposed by the nested model. The circles A, B and C represent three different encompassing posterior distributions. Note that the lower and upper bound for μ_1 and μ_2 are the same.

Bayes Factors for Exact Equality Constraints

In some situations, any difference between μ_1 and μ_2 is deemed relevant, and this requires a test for exact equality. For instance, one may wish to test whether a chemical compound adds to the effectiveness of a particular medicine. In such experimental studies, an exact null effect is a priori plausible. However, it may appear that the EP approach does not extend to exact equality constraints in a straightforward fashion.

To illustrate, consider our example in which $M_e : \mu_1, \mu_2; \sigma$ and now $M_3 : \mu_1 = \mu_2; \sigma$. Figure 3.1c shows that the only values allowed by the constrained model M_3 are those that fall exactly on the diagonal. As μ_1 and μ_2 are continuous variables, the proportion of prior and posterior samples that obey this constraint is zero. Therefore, the EP Bayes factor is $0/0$, which has led several researchers to conclude that the EP Bayes factor is not defined for exact equality constraints (Rossell, Baladandayuthapani, & Johnson, 2008, pp. 111-112; J. I. Myung, Karabatsos, & Iverson, 2008, p. 317; Klugkist, 2008, p. 71). The next two sections investigate in what sense the EP Bayes factor can be defined for exact equality constraints, and its relation to the Savage-Dickey density ratio. Difficulties that arise because of the Borel-Kolmogorov paradox are discussed in the subsequent sections.

Bayes factors for exact equality constraints: An iterative method

In order to estimate the EP Bayes factor for exact equality constrained models, Laudy (2006, p. 115) and Klugkist (2008) proposed an iterative procedure. In the context of a test between $M_e : \mu_1, \mu_2; \sigma$ and $M_3 : \mu_1 = \mu_2; \sigma$, the procedure comprises the following steps:

Step 1: Choose a small value ε_1 and define $M_{3.1} : |\mu_1 - \mu_2| < \varepsilon_1$;

Step 2: Compute the Bayes factor $BF_{(3.1)e}$ using Equation 3.2;

Step 3: Define $\varepsilon_2 < \varepsilon_1$ and $M_{3.2} : |\mu_1 - \mu_2| < \varepsilon_2$;

Step 4: Sample from the constrained ($|\mu_1 - \mu_2| < \varepsilon_1$) prior and posterior and compute the Bayes factor $BF_{(3.2)(3.1)}$;

Repeat steps 3 and 4, with each $\varepsilon_{n+1} < \varepsilon_n$, until $BF_{n+1,n} \approx 1$. Then the required Bayes factor BF_{3e} can be calculated by multiplication:

$$BF_{3e} = BF_{(3.1)e} \times BF_{(3.2)(3.1)} \times, \ldots, \times BF_{n(n-1)}. \tag{3.3}$$

In the limit (i.e., when $\varepsilon_n \to 0$), this method yields the Bayes factor for the exact equality model M_3 versus the encompassing model M_e. Although this iterative method solves the problem of having no samples that obey an exact equality constraint, the method is only approximate and potentially time consuming.

Bayes factors for exact equality constraints: A one-step method—equivalence to the Savage-Dickey density ratio

The iterative procedure turns out to be identical to the Savage-Dickey density ratio method, a one-step method that is both principled and fast. In order to understand this intuitively, Figure 3.2 shows a fictitious prior and posterior distribution for $\mu_1 - \mu_2$, obtained under the encompassing model M_e. The surface of the dashed areas equals the proportion of the prior and posterior distribution that is consistent with the constraint $|\mu_1 - \mu_2| < \varepsilon$. In the EP approach, the Bayes factor is obtained by integrating the posterior and prior distribution over the area defined by the constraint. However, it is clear that as $\varepsilon \to 0$, the area of both regions equals 0.

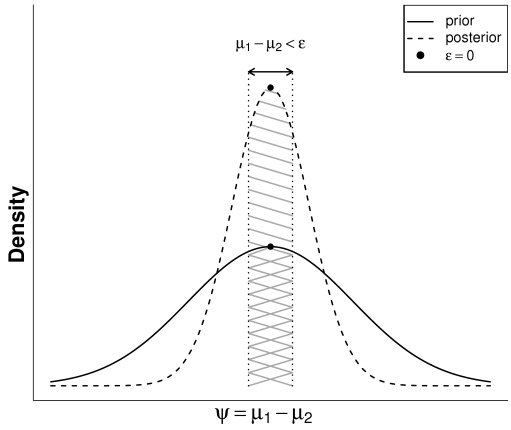


Figure 3.2: The encompassing prior approach for exact equality constraints is the Savage-Dickey density ratio. The top dot represents the value of the posterior distribution at $\mu_1 = \mu_2$ and the bottom dot represents the value of the prior distribution at $\mu_1 = \mu_2$. The ratio of the heights of both densities equals the Bayes factor. Note that the posterior of ψ does not have to be centered around zero.

The Bayes factor is given by the ratio of the two integrals. Hence, the Bayes factor for the equality constraint in the EP approach is the limit

$$BF_{3e} = \lim_{\epsilon \to 0} \frac{\int_{-\epsilon/2}^{\epsilon/2} p(\psi_0 + \psi \mid D, M_e) d\psi}{\int_{-\epsilon/2}^{\epsilon/2} p(\psi_0 + \psi \mid M_e) d\psi}.$$

Here we generically formulated the hypothesis in terms of the parameter ψ. In the example hypothesis $H_0 : \mu_1 = \mu_2$ this corresponds to defining $\psi = \mu_1 - \mu_2$ and $\psi_0 = 0$.

We also marginalized over any nuisance parameters not of interest; σ in our example, i.e., $p(\mu_1, \mu_2 \mid D, M_e) = \int_{-\infty}^{\infty} p(\mu_1, \mu_2, \sigma \mid D, M_e) d\sigma$. Then, in the example, $\psi = \mu_1 - \mu_2$ has marginal posterior density $p(\psi|D, M_e) = \int p(\mu_1, \mu_1 - \psi|D, M_e) d\mu_1$ (see e.g., Miller & Miller, 2004, pp. 246). These integrals can be evaluated analytically, with quadratures, or can be approximated using MCMC sampling (Gamerman & Lopes, 2006). To calculate the Bayes factor, only the marginal posterior density of interest needs to be considered.

Clearly the limit above approaches the form 0/0 and so l'Hôpital's 0/0 rule can be employed to obtain

$$BF_{3e} = \lim_{\epsilon \to 0} \frac{p(\psi_0 + \epsilon/2 \mid D, M_e)/2 + p(\psi_0 - \epsilon/2 \mid D, M_e)/2}{p(\psi_0 + \epsilon/2 \mid M_e)/2 + p(\psi_0 - \epsilon/2 \mid M_e)/2} = \frac{p(\psi_0 \mid D, M_e)}{p(\psi_0 \mid M_e)}, \tag{3.4}$$

where ψ_0 represents the point of exact equality specified by the constrained model; in our example, $M_3 : \psi_0$ means $\mu_1 - \mu_2 = 0$.

Equation 3.4 shows that the Bayes factor BF_{3e} simplifies to the ratio of the height of the marginal posterior and the height of the marginal prior at the point of interest, if the limiting processes in the numerator and the denominator are chosen to be equal. This result is known as the Savage-Dickey density ratio (Dickey & Lientz, 1970; O'Hagan & Forster, 2004; Dickey, 1971; Verdinelli & Wasserman, 1995). For the example shown in Figure 3.2, the Bayes factor in favor of the exact equality model, BF_{3e}, is approximately 2.

For completeness, we now sketch the proof that the Savage-Dickey density ratio equals the Bayes factor (cf. O'Hagan & Forster, 2004). As before, let ψ be the parameter of interest and ϕ the nuisance parameter; let M_e be the encompassing model, a restricted version of which is defined as $M_3 : \psi = \psi_0$. The Savage-Dickey density ratio is equal to the Bayes factor if the prior of the nuisance parameter in the restricted model M_3 is defined by conditioning, that is, if $p(\phi|M_3) = p(\phi|\psi = \psi_0, M_e)$ (cf. Equation 3.1).

The foregoing allows us to rewrite the marginal likelihood for M_3:

$$\begin{aligned} p(D|M_3) &= \int p(D|\phi, M_3) p(\phi|M_3) d\phi, \\ &= \int p(D|\phi, \psi = \psi_0, M_e) p(\phi|\psi = \psi_0, M_e) d\phi, \\ &= p(D|\psi = \psi_0, M_e). \end{aligned} \tag{3.5}$$

We now apply Bayes' rule to the end result of Equation 3.5 and obtain

$$p(D|M_3) = \frac{p(\psi = \psi_0|D, M_e) p(D|M_e)}{p(\psi = \psi_0|M_e)}. \tag{3.6}$$

Dividing both sides of Equation 3.6 by $p(D|M_e)$ results in

$$BF_{3e} = \frac{p(D|M_3)}{p(D|M_e)} = \frac{p(\psi = \psi_0|D, M_e)}{p(\psi = \psi_0|M_e)}, \tag{3.7}$$

which shows that the Bayes factor equals the ratio of the posterior and prior ordinate under M_e at the point of interest (i.e., $\psi = \psi_0$).

An example comparing the iterative EP approach to the Savage-Dickey method

In order to illustrate how the results from the two methods converge, we randomly drew 100 samples from a Normal distribution with mean 0.2 and variance 1, and found a

corresponding one-sample t-statistic of 1.945. We then used a Bayesian t-test with a Cauchy(0,1) prior on effect size $\delta = \mu/\sigma$ and a folded Cauchy(0,1) on σ (for details see Rouder et al., 2009) to compute the Bayes factor in favor of $H_0 : \delta = 0$ relative to $H_1 : \delta \sim$ Cauchy(0,1), which yielded $BF_{3e} = 2.011$.

Figure 3.3 compares the behavior of the iterative encompassing prior approach to that of the Savage-Dickey density ratio. The dashed horizontal line shows the result from the Savage-Dickey implementation of the Bayesian t-test (Wetzels, Raaijmakers, Jakab, & Wagenmakers, 2009). The dots show the result from the iterative encompassing prior approach (Equation 3.3), as a function of the size of the smallest interval ε. When $\varepsilon = 0.01$, the iterative EP Bayes factor has converged to the correct Bayes factor.

Note that the iterative EP approach involves the product of multiple Bayes factors (cf. Equation 3.3). In contrast, the Savage-Dickey procedure involves only one Bayes factor. Because the computation of each Bayes factor requires many MCMC samples, the computational demands are likely to be much higher in the iterative EP approach than in the Savage-Dickey approach.

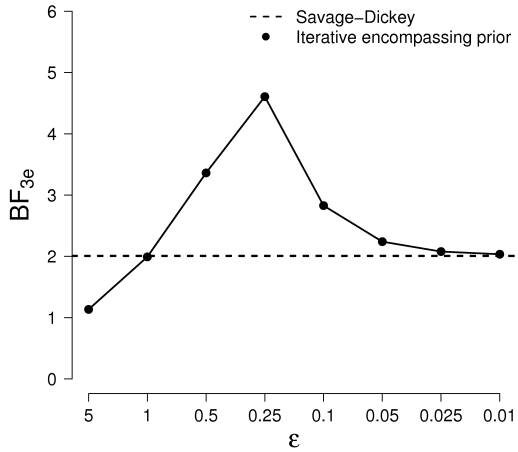


Figure 3.3: A comparison of the Savage-Dickey density ratio and the iterative encompassing prior approach for simulated data. The Bayes factor favoring the null hypothesis, BF_{3e}, is 2.011. The dashed line shows the Savage-Dickey Bayes factor. The dots represent the iterative Bayes factor calculated by systematically decreasing the interval surrounding the exact equality of interest (Equation 3.3).

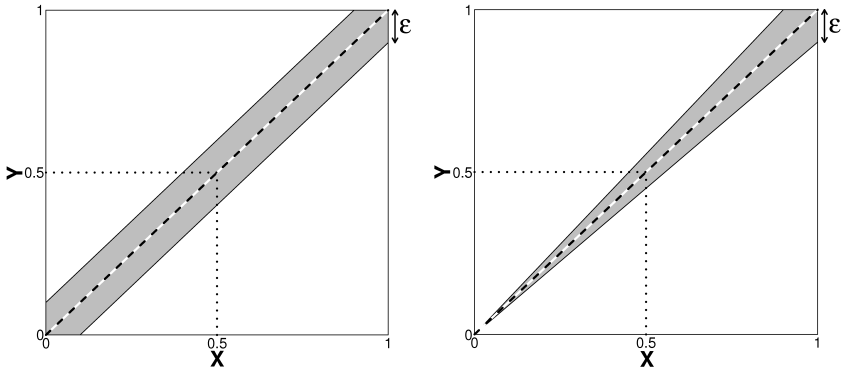
3.3 The Borel-Kolmogorov Paradox

The main drawback of the EP approach to exact equalities is that it is subject to the Borel-Kolmogorov paradox (DeGroot & Schervish, 2002; Jaynes, 2003; D. Lindley, 1997; Proschan & Presnell, 1998; Rao, 1988; Singpurwalla & Swift, 2001). This paradox arises when one conditions on events of probability zero. In the case of exact equality constraints, priors for the constrained model are constructed by conditioning on a null-set, and this gives rise to the Borel-Kolmogorov paradox.

The Borel-Kolmogorov Paradox: An example

Consider the following situation, inspired by an example from D. Lindley (1997). Suppose that a point P is described by its Cartesian coordinates X and Y. Furthermore, suppose that $0 \leq X \leq 1$ and $0 \leq Y \leq 1$, and that P has a uniform distribution on the unit square. Suppose you are told that P lies on the diagonal through the origin, event B. What is your probability that X, associated with that P, and hence also Y, is less then $1/2$ (i.e., event A)?

The paradox lies in the fact that the answer to this question depends on how we parameterize the diagonal. We examine two situations: $Z_1 = X - Y = 0$ (see Figure 3.4a) and $Z_2 = X/Y = 1$ (see Figure 3.4b). Note that because X and Y are continuous, the probability that $X = Y$ is zero. Because conditioning on an event with probability zero is problematic, we consider values of X and Y that lie in the proximity of the line $X = Y$.



(a) X-Y=0 (b) X/Y=1

Figure 3.4: Example of the Borel-Kolmogorov paradox. The shaded areas indicate the acceptable values of X and Y for the two parameterizations. In panel (a), all values of X are equally likely while in panel (b), the wedge shaped area implies that values close to 1 are more likely than those close to 0.

From the geometry of the problem, the associated probability in the first case is

$$P(Y - \epsilon \leq X \leq Y + \epsilon) = 2(1/2 - 1/2(1 - \epsilon)^2) = (2 - \epsilon)\epsilon,$$

while the associated probability for the second case is

$$P(Y - \epsilon Y \leq X \leq Y - \epsilon Y) = 2(1/2 - 1/2 \cdot 1 \cdot (1 - \epsilon)) = \epsilon.$$

Now consider the probabilities that (X, Y) lies in the left lower quadrant of the square (i.e., $X, Y \leq 1/2$) and either that $|X - Y| < \epsilon$ and or that $|X/Y - 1| < \epsilon$. Again from geometry, the probability of the first case is

$$P(|X - Y| < \epsilon \cap X, Y \leq 1/2) = (1 - \epsilon)\epsilon,$$

and for the second case

$$P(|X/Y - 1| < \epsilon \cap X, Y \leq 1/2) = \frac{1}{4}\epsilon.$$

Hence, the corresponding conditional probabilities of the events that (X, Y) lies in the left lower quadrant of the square, given that either the event $|X - Y| \leq \epsilon$ or that the event $|X/Y - 1| \leq \epsilon$ occurred, are respectively

$$P(X, Y \leq 1/2 \,|\, |X - Y| < \epsilon) = \frac{1 - \epsilon}{2 - \epsilon},$$

and

$$P(X, Y \leq 1/2 \,|\, |X/Y - 1| < \epsilon) = \frac{1}{4}.$$

If we now take to the limit $\epsilon \to 0$, both the events $|X - Y| \leq \epsilon$ and $|X/Y - 1| \leq \epsilon$ coincide with the lower left half of the diagonal of the square. However, although the first probability coincides with our intuition that $P(X, Y \leq 1/2 \,|\, X = Y) = 1/2$, the second has it that the probability for this seemingly equal event should be $1/4$!

This example shows that the probability of an event conditioned on a limiting event of zero probability depends on the way in which the limiting event was generated, that is, on the parameterization that was chosen to generate the zero probability event. In effect, conditional probability is not invariant under coordinate transformations of the conditioning variable. This paradox is resolved if one accepts that conditional probability cannot be unambiguously defined with respect to events of zero probability without specifying the limiting process from which it should result (Jaynes, 2003). It is on random variables, not on singular events, that conditioning is unambiguous (see Kolmogorov, 1956, Billingsley, 2008, and Wolpert, 1995).

Implications for the limiting encompassing prior and the Savage-Dickey density ratio

The foregoing implies that in the EP approach to exact equalities, the resulting Bayes factor may depend on the choice of the parameterization, a feature that is clearly undesirable (Dawid & Lauritzen, 2001; Schweder & Hjort, 1996; Wolpert, 1995). Note that the Borel-Kolmogorov paradox does not occur in the case of inequality constraints, where one conditions on an interval, rather than on a single point.

Equation 3.4 shows that the EP Bayes factor for exact equality constraints is equal to the Savage-Dickey ratio. However, the rectangular regions of integration and the use of the same limiting processes in both the numerator and the denominator are arbitrary choices. Different choices of the limiting process can lead to different Bayes factors, as shown next. To this end, let $\gamma_i(\epsilon) \geq 0$ and $\delta_i(\epsilon) > 0$, differentiable in a neighborhood $(0, \epsilon)$, such that $\lim_{\epsilon \to 0} \gamma_i(\epsilon) = \lim_{\epsilon \to 0} \delta_i(\epsilon) = 0$ and $\gamma_i'(0) + \delta_i'(0) \neq 0$, for $i = 1, 2$. Here prime $'$ indicates derivative. Then, without loss of generality, these functions can be chosen to suit any form of the limiting process in the EP process

$$BF_{3e} = \lim_{\epsilon \to 0} \frac{\int_{-\gamma_1(\epsilon)}^{\delta_1(\epsilon)} p(\psi_0 + \psi \,|\, D, M_e) d\psi}{\int_{-\gamma_2(\epsilon)}^{\delta_2(\epsilon)} p(\psi_0 + \psi \,|\, M_e) d\psi}.$$

Intuitively this would seem to go to the same limit as earlier, but in fact it does not, as l'Hôpital's rule shows:

$$\begin{aligned} BF_{3e} &= \lim_{\epsilon \to 0} \frac{p(\psi_0 + \delta_1(\epsilon) \,|\, D, M_e)\delta_1'(\epsilon) + p(\psi_0 - \gamma_1(\epsilon) \,|\, D, M_e)\gamma_1'(\epsilon)}{p(\psi_0 + \delta_2(\epsilon) \,|\, M_e)\delta_2'(\epsilon) + p(\psi_0 - \gamma_2(\epsilon) \,|\, M_e)\gamma_2'(\epsilon)}, \\ &= \frac{p(\psi_0 \,|\, D, M_e)\,[\delta_1'(0) + \gamma_1'(0)]}{p(\psi_0 \,|\, M_e)\,[\delta_2'(0) + \gamma_2'(0)]}. \end{aligned} \tag{3.8}$$

This is the above Savage-Dickey ratio if and only if $\delta_1'(0)+\gamma_1'(0) = \delta_2'(0)+\gamma_2'(0)$. As both $\delta_1'(0)+\gamma_1'(0)$ and $\delta_2'(0)+\gamma_2'(0)$ measure the rate at which the numerator and denominator approach zero, the limit of the EP approach equals the Savage-Dickey ratio if and only if both numerator and denominator approach 0 at the same rate. If the rate at which the numerator and the denominator approach zero is not the same, any desired value of the Bayes factor can be obtained.

In light of the Borel-Kolmogorov paradox, it is important to understand when the Savage-Dickey ratio method is invariant under smooth transformations of the chosen parameterization, especially when nuisance parameters are present. To this end, suppose the chosen set of (absolute continuous) parameters is θ with prior $p(\theta|M_e)$ and posterior $p(\theta|D, M_e)$. Let g be a differentiable invertible transform (a diffeomorfism) with inverse h so that

$$\chi = g(\theta) \qquad \text{and} \qquad h(\chi) = \theta.$$

The implied prior is denoted $\widetilde{p}(\chi|M_e)$ and the implied posterior is denoted $\widetilde{p}(\chi|D, M_e)$. In general, the parameter vector can be partitioned as $\theta = (\psi, \phi)$, where ϕ contains nuisance parameters that are not involved in the evaluation of the null hypothesis. We are interested in evaluating the evidence for the simple hypothesis

$$M_3 : \psi = \psi_0,$$

which, in terms of $\chi = (\nu, \xi)$ can often be cast equivalently as $M_3 : \nu = \nu_0$. We wish to know under what circumstances the Savage-Dickey ratios are equal. That is, we want to determine conditions on g under which the desired equality

$$BF_{3e} = \frac{p(\psi_0|D, M_e)}{p(\psi_0|M_e)} = \frac{\widetilde{p}(\nu_0|D, M_e)}{\widetilde{p}(\nu_0|M_e)}, \tag{3.9}$$

is true. It turns out that this equality holds, as long as the transformation g does not depend on the data D, and as long as the parameters on which M_3 imposes a simple hypothesis transform independently of the nuisance parameters. This follows from the following considerations.

By the "change of variables" rule,

$$\begin{aligned}\widetilde{p}(\chi|D, M_e) &= p(h(\chi)|D, M_e)|h'(\chi)|_+,\\ \widetilde{p}(\chi|M_e) &= p(h(\chi)|M_e)|h'(\chi)|_+,\end{aligned}$$

where $|h'(\chi)|_+$ denotes the absolute value of the determinant of the Jacobian matrix $h'(\chi) = \partial_\chi h(\chi)$ of the transformation h. Partition $h(\chi)$ as $\theta = h(\chi) = (\psi, \phi) = (\psi(\nu, \xi), \phi(\nu, \xi))$. In terms of hypothesis and nuisance parameters these can be expressed as

$$\widetilde{p}(\nu, \xi|M_e) = p(\psi(\nu, \xi), \phi(\nu, \xi)|M_e)\, |\phi_\xi(\nu, \xi)|_+ |\psi_\nu(\nu, \xi) - \phi_\nu(\nu, \xi)\phi_\xi(\nu, \xi)^{-1}\psi_\xi(\nu, \xi)|_+, \tag{3.10}$$

and similarly for $\widetilde{p}(\nu, \xi|D, M_e)$. Here $\phi_\xi(\nu, \xi)$ denotes the matrix of partial derivatives of ϕ with respect to ξ, ψ_ν of ψ with respect to ν, etc.

The implicit function theorem ensures the existence of a function $\nu(\psi, \xi)$, such that $\psi_0 = \psi(\nu(\psi_0, \xi), \xi)$, for all ξ. Then, by the chain rule,

$$\int p(\psi_0, \phi|M_e)d\phi = \int p(\psi(\nu(\psi_0, \xi), \xi), \phi(\nu(\psi_0, \xi), \xi)|M_e)\, |\partial_\xi \phi(\nu(\psi_0, \xi), \xi)|_+ \, d\xi.$$

The Jacobian in the last integral can be expressed as

$$|\partial_\xi \phi(\nu(\psi_0,\xi),\xi)|_+ = |\phi_\nu(\nu(\psi_0,\xi),\xi)\nu_\xi(\psi_0,\xi) + \phi_\xi(\nu(\psi_0,\xi),\xi)|_+.$$

Therefore, if $\nu_\xi(\psi,\xi) \equiv 0$, implying that $\nu(\psi,\xi) = \nu(\psi)$ does not depend on ξ, then

$$\int p(\psi_0,\phi|M_e)d\phi = \int p(\psi(\nu(\psi_0),\xi|M_e),\phi(\nu(\psi_0),\xi))\ |\phi_\xi(\nu(\psi_0),\xi)|_+\ d\xi,$$

which, by equation (3.10) can be expressed as

$$\begin{aligned}\int p(\psi_0,\phi|M_e)d\phi &= \int \widetilde{p}(\nu_0,\xi|M_e)\ |\psi_\nu(\nu_0,\xi) - \phi_\nu(\nu_0,\xi)\phi_\xi(\nu_0,\xi)^{-1}\psi_\xi(\nu_0,\xi)|_+^{-1}\ d\xi,\\ &= \int \widetilde{p}(\nu_0,\xi|M_e)\ |\psi_\nu(\nu_0)|_+^{-1}\ d\xi.\end{aligned}$$

Here we used the fact that $\psi(\nu(\psi_0),\xi) = \psi(\nu(\psi_0,\xi),\xi) = \psi_0$. We also used the fact that $\nu_0 = \nu(\psi_0)$, which is warranted by the above assumption that $\nu_\xi(\psi,\xi) = 0$. Specifically, because $\partial_\xi\psi(\nu(\psi_0,\xi),\xi) = \psi_\nu(\nu(\psi_0,\xi),\xi)\nu_\xi(\psi_0,\xi) + \psi_\xi(\nu(\psi_0,\xi),\xi) = \partial_\xi\psi_0 = 0$, it follows that $\psi_\xi(\nu(\psi_0,\xi),\xi) \equiv 0$ for all ψ_0. This implies that ψ does not depend on ξ, and therefore that $\psi_0 = \psi(\nu(\psi_0,\xi),\xi) = \psi(\nu(\psi_0)) = \psi(\nu_0)$. The latter conclusion that $\psi_\xi \equiv 0$ also yields the second equality.

Consequently, the evidence for M_3 is obtained from the Savage-Dickey ratio

$$\begin{aligned}BF_{3e} &= \frac{p(\psi_0|D,M_e)}{p(\psi_0|M_e)} = \frac{\int p(\psi_0,\phi|D,M_e)d\phi}{\int p(\psi_0,\phi|M_e)d\phi}\\ &= \frac{\int \widetilde{p}(\psi_0,\xi|D,M_e)|\psi_\nu(\nu_0)|_+^{-1}d\xi}{\int \widetilde{p}(\psi_0,\xi|M_e)|\psi_\nu(\nu_0)|_+^{-1}d\xi} = \frac{\widetilde{p}(\psi_0|D,M_e)}{\widetilde{p}(\psi_0|M_e)},\end{aligned}$$

which is (3.9).

In sum, computing the Bayes factor for exact equality constraints is a delicate matter. The iterative EP approach and the Savage-Dickey density ratio can lead to different Bayes factors if the limiting process in the iterative EP approach is not carefully chosen (i.e., the numerator and denominator of Equation 3.8 should approach 0 at the same rate). Moreover, both methods suffer from the Borel-Kolmogorov paradox. However, the Savage-Dickey density ratio is invariant under smooth transformations of the chosen parameterization, as long as the transformation does not depend on the data, and as long as the parameters transform independently of the nuisance parameters.

3.4 Concluding Remarks

Here we have shown that the Savage-Dickey density ratio method is a special case of the encompassing prior (EP) approach proposed by Hoijtink, Klugkist, and colleagues. The EP approach was developed to account for models with inequality constraints; as it turns out, the approach naturally extends to models with exact equality constraints. Consequently, the EP approach offers a unified, elegant, and simple method to compute Bayes factors in nested models.

The main drawback of the EP/Savage-Dickey method for exact equalities is its susceptibility to the Borel-Kolmogorov paradox. We have shown that the SD-ratio yields the same value under different transformations, as long as the parameters, on which M_1 imposes a simple hypothesis, transform independently of the nuisance parameters. It should

be noted that in order to avoid the Borel-Kolmogorov paradox, alternative procedures seek to construct priors not by the usual conditioning, but by method such as marginalization (Kass & Raftery, 1995), Jeffreys conditioning (Dawid & Lauritzen, 2001), reference conditioning (Roverato & Consonni, 2004), Kullback-Leibler projection (Consonni & Veronese, 2008; Dawid & Lauritzen, 2001), and Hausdorff integrals (Kleibergen, 2004).

Unfortunately, these alternative procedures give rise to paradoxes and problems of their own (see Consonni & Veronese, 2008, for a review and a comparison). Presently, there does not appear to be a universally agreed-on method for specifying priors in nested models that is clearly superior to the conditioning procedure inherent in the Hoijtink and Klugkist EP approach.

4 A Default Bayesian Hypothesis Test for Correlations and Partial Correlations

Abstract

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. The test is easy to apply and yields practical advantages that the standard frequentist tests lack; in particular, the Bayesian test can quantify evidence in favor of the null hypothesis and allows researchers to monitor the test results as the data come in. We illustrate the use of the Bayesian correlation test with three examples from the psychological literature. Computer code and example data are provided in the journal archives.

An excerpt of this chapter has been published as:
Wetzels, R., and Wagenmakers, E.-J. (in press). A Default Bayesian Hypothesis Test for Correlations and Partial Correlations. *Psychonomic Bulletin & Review.*

4.1 Introduction

A correlation coefficient indicates how strongly two variables are related. The concept is basic and it comes as no surprise that the correlation coefficient ranks among the most popular statistical tools in any subfield of psychological science. For instance, in experimental psychology the correlation coefficient has been used to quantify the relation between time spent meditating and visual acuity (MacLean et al., 2010); in social psychology it has been used to quantify the relation between income and subjective well-being (Diener, Ng, Harter, & Arora, 2010); and in neuroscience it has been used to quantify the relation between response inhibition and the connectivity of brain networks (Jahfari et al., 2011).

The first correlation coefficient was developed by Francis Galton in 1888 (Stigler, 1989); further work by Francis Edgeworth and Karl Pearson resulted in the correlation measure that is used most frequently today, the "Pearson product-moment correlation coefficient" or r (Pearson, 1920; Rodgers & Nicewander, 1988; Cohen, Cohen, West, & Aiken, 2003; see Stigler, 1986 for a detailed historical overview). The coefficient r is a measure of the linear relation between two variables, where $r = -1$ indicates a perfectly negative linear relation, $r = 1$ indicates a perfectly positive relation, and $r = 0$ indicates the absence of any linear relation.

In this article we focus on the hypothesis test for the Pearson correlation coefficient. This is one of the most common tests in experimental psychology, taught in virtually every undergraduate statistics class. The test that is taught is the standard (i.e., classical, orthodox, or frequentist) hypothesis test, a test that produces a p value for drawing conclusions – the common rule is that when $p < .05$, one can reject the null hypothesis that no relation is present.

Unfortunately, frequentists p value tests have a number of drawbacks (e.g., J. O. Berger & Wolpert, 1988; Edwards et al., 1963; J. K. Kruschke, 2010b; Sellke, Bayarri, & Berger, 2001; Wagenmakers, 2007). First, p values do not allow researchers to quantify evidence in favor of the null hypothesis (Gallistel, 2009; Rouder et al., 2009). Second, p values depend on the sampling plan and hence its users may not stop data collection when an interim result is compelling, nor may they continue data collection when the fixed sample size result is ambiguous (Edwards et al., 1963). Third, the p value overestimates the evidence against the null hypothesis (Sellke et al., 2001; Rouder & Morey, 2011; Wetzels et al., 2011). These drawbacks are not merely theoretical, but have real consequences for the way in which psychologists carry out their experiments and draw conclusions from their data.

An alternative to frequentist tests is provided by Bayesian inference, and in particular the so-called Bayes factor (Jeffreys, 1961; Kass & Raftery, 1995). The Bayes factor computes the probability of the observed data under the null hypothesis vis-a-vis the alternative hypothesis. In contrast to the frequentist p value, the Bayes factor allows researchers to quantify evidence in favor of the null hypothesis (e.g., Wetzels et al., 2009). Moreover, with the Bayes factor "It is entirely appropriate to collect data until a point has been proven or disproven, or until the data collector runs out of time, money, or patience" (Edwards et al., 1963, p. 193). Thus, the Bayes factor altogether eliminates the optional stopping phenomenon, where researchers can bias their results by collecting data until $p < .05$ (e.g., Simmons, Nelson, & Simonsohn, 2011). Researchers are allowed to monitor the Bayes factor as the data come in, and stop whenever they feel the evidence is compelling. This flexibility can be of great ethical, practical, and financial importance; in clinical trials for instance, it may limit the number of patients who are needlessly exposed to inferior treatment.

In the field of psychology, interest in hypothesis testing using the Bayes factor has greatly increased over the last years. For instance, Rouder and colleagues have adopted a method for variable selection in regression models (Liang et al., 2008) to yield a Bayesian t test (Rouder et al., 2009; Wetzels et al., 2009); Masson has shown how statistical output from SPSS can be translated to Bayes factors using the BIC approximation (Masson, 2011; see also Raftery, 1995; G. Schwarz, 1978; Wagenmakers, 2007); Hoijtink, Klugkist, and colleagues have promoted Bayes factors for order-restricted inference (e.g., Hoijtink et al., 2008; Klugkist & Hoijtink, 2007; Mulder et al., 2009); finally, we have published two tutorial articles for psychologists to facilitate the computation of Bayes factors (Lodewyckx et al., 2011; Wagenmakers, Lodewyckx, Kuriyal, & Grasman, 2010; see also Morey, Rouder, Pratte, & Speckman, 2011).

Perhaps the greatest impediment to the large-scale adoption of the Bayes factor is the lack of easy-to-use tests for statistical models that psychologists use in practice. For example, the test for the presence of a correlation (and partial correlation) is one of the most popular workhorses in experimental psychology, yet many psychologists will struggle to find a Bayes factor equivalent. In this article we remove this hurdle by providing an easy-to-use Bayes factor alternative to the Pearson correlation test.

In this article we first discuss the standard, frequentist tests for the presence of correlation and partial correlation. Next we explain Bayesian model selection in general, and then focus on a default Bayesian test for correlation and partial correlation. Key concepts and computations are illustrated with three examples of recent psychological experiments.

4.2 Frequentist Test for the Presence of Correlation

We discuss the frequentist correlation test in the context of a study where participants were involved in an intensive meditation training program MacLean et al. (2010). The aim of this program was to investigate if there is an effect of meditation on visual acuity. To assess visual acuity, participants were asked to judge repeatedly whether a vertical line was long or short. Perceptual threshold was defined as the difference in visual angle between the short and the long line that allowed the participant to classify the lines correctly 75% of the time. The main result of the experiment was that the intensive meditation program decreased participants' perceptual threshold.

In addition to this main result, MacLean et al. (2010) wanted to explore if the improved visual acuity is retained five months after termination of the meditation program, and, more specifically, whether at follow-up the participants that had meditated the most also had the lowest threshold. The follow-up involved 54 participants, whose data are replotted in Figure 4.1. Based on these data, MacLean et al. (2010, p. 6) concluded that "this result indicates a correlation between the long-term stability of training-induced discrimination improvement and the maintenance of regular, but less intensive, meditation practice".

To calculate the correlation between threshold and meditation time, we first define the following variables: for person i, mean daily meditation time is denoted x_i, and threshold is denoted y_i. For meditation time and threshold, the sample variances are $s_X^2 = 20,916.68$ and $s_Y^2 = 0.05$, and sample means are $\bar{x} = 121$ and $\bar{y} = 0.56$, respectively. Then, the sample correlation coefficient of X and Y is calculated as follows:

$$r_{XY} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_X s_Y} = \frac{-589}{1629} = -.36,$$

where n is the number of participants ($n = 54$).

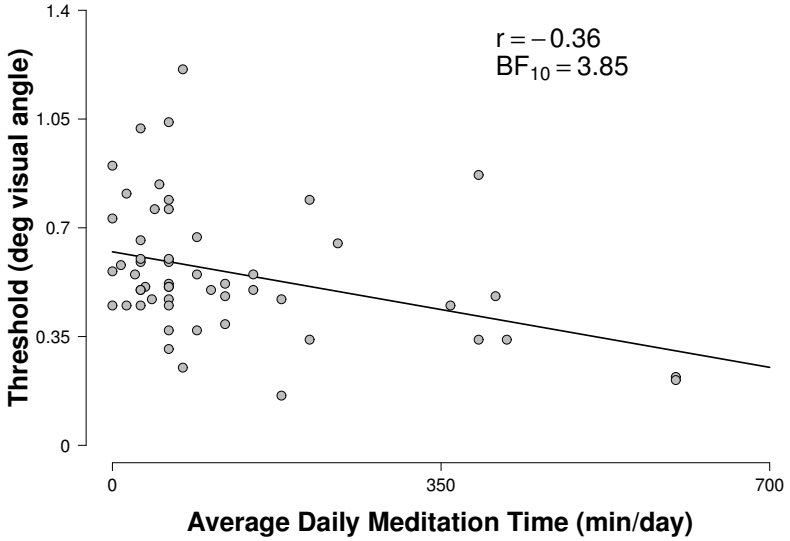


Figure 4.1: The relationship between average daily meditation time and discrimination threshold. A negative correlation suggests that time spent in meditation improves visual perception (i.e., lowers the threshold). Data replotted from MacLean et al., 2010.

In order to test whether we can reject the null hypothesis that the correlation coefficient is zero, $\rho_{XY} = 0$, we calculate the t statistic (using $r_{XY} = -.36$ and $n = 54$):

$$t = r_{XY}\sqrt{\frac{(n-2)}{(1-r_{XY}^2)}} = -2.80,$$

which follows the Student t distribution with $n - 2$ degrees of freedom. This t statistic corresponds to a p value of 0.01. Therefore, with a significance level of $\alpha = 0.05$, researchers may feel that they can confidently reject the null hypothesis of no correlation.

4.3 Frequentist Test for the Presence of Partial Correlation

Partial correlation is the correlation between two variables, say X and Y, after the confounding effect of a third variable Z has been removed. Variable Z is known as the control variable. In psychological research, there are many situations in which one might want to partial out the effects of a control variable.

Consider a recent experiment on the role of implicit prediction in visual search by Lleras, Porporino, Burack, and Enns (2011). Implicit prediction was studied using an interrupted search task featuring three groups of children and one group of adults (i.e., mean age 7, 9, 11, and 19 years). In the search task, participants had to identify a target among a set of distractors (i.e., one "T" among 15 "L" shapes). Crucially, brief looks at the search display (100-500 ms) were interrupted by longer "waits" in which the participant was shown a blank screen (1000-3500 ms). The focus of this study was on *rapid resumption*, the phenomenon that in contrast to the first look at the stimulus

(where only 2% of the correct responses are faster than 500 ms), subsequent looks often show $30 - 50\%$ correct responses faster than 500 ms.

Based on $n = 40$ observations, Lleras et al. (2011) calculated the correlation between mean successful search time (X) and the proportion of rapid resumption responses (Y): $r_{XY} = .51$, a highly significant correlation ($p < 0.01$). However, Lleras et al. (2011) also observed that this correlation does not take the participants' age into account. The correlation between search time (X) and age (Z) is relatively high (i.e., $r_{XZ} = -.78$), and so is the correlation between rapid resumption (Y) and age (i.e., $r_{YZ} = -.66$). Hence, the authors computed a partial correlation to exclude the possibility that age Z caused the correlation between search time X and rapid resumption Y. This is accomplished by the following formula:

$$r_{XY|Z} = \frac{r_{XY} - r_{XZ}r_{YZ}}{\Big[(1 - r_{XZ}^2)(1 - r_{YZ}^2)\Big]^{1/2}} = \frac{.51 - (-.78)(-.66)}{\Big[(1 - (-.78)^2)(1 - (-.66)^2)\Big]^{1/2}} = -.01.$$

This result shows that by controlling for the variable age, the correlation between search time and rapid resumption is virtually eliminated. The *correlation*, r_{XY}, is .51, but the *partial correlation*, $r_{XY|Z}$, is $-.01$. The p value for the partial correlation can be calculated by computing the t statistic (using $r_{XY|Z} = -.01$ and $n = 40$):

$$t = r_{XY|Z}\sqrt{\frac{(n-3)}{(1 - r_{XY|Z}^2)}} = -0.06,$$

which follows the Student t distribution with $n - 3$ degrees of freedom. This t statistic corresponds to a p value of .95. Hence, Lleras et al. (2011) failed to reject the null hypothesis of no correlation between search time and rapid resumption.

Note that this non-significant result leaves the null hypothesis in a state of suspended disbelief. It is not statistically correct to conclude from a non-significant result that the data support the null hypothesis — after all, the same non-significant result could have been due to the fact that the data were relatively noisy. This means that a non-significant p value cannot be used to differentiate between noisy data that yield ambiguous evidence, and clean data that yield compelling evidence – evidence in favor of the null hypothesis. This is one of the prominent p value problems that does not occur in the rival framework of Bayesian inference.

4.4 Bayesian Hypothesis Testing

In order to keep this article self-contained, we now briefly discuss Bayesian hypothesis testing. Readers who are familiar with Bayesian inference can skip to the next section.

In Bayesian model selection or hypothesis testing, the competing statistical hypotheses are assigned prior probabilities. Suppose we have two competing hypotheses, the null hypothesis H_0, and the alternative hypothesis H_1. These hypotheses are assigned prior probabilities $p(H_0)$ and $p(H_1)$. Then, after observing the data $\boldsymbol{Y}$, Bayes' theorem is applied to obtain the posterior probability of both hypotheses. The posterior probability of the alternative hypothesis $p(H_1 \mid \boldsymbol{Y})$, is calculated as follows:

$$p(H_1 \mid \boldsymbol{Y}) = \frac{p(\boldsymbol{Y} \mid H_1)p(H_1)}{p(\boldsymbol{Y} \mid H_1)p(H_1) + p(\boldsymbol{Y} \mid H_0)p(H_0)},$$

where $p(\boldsymbol{Y} \mid H_1)$ denotes the marginal likelihood of the data under the alternative hypothesis (and equivalently for the null hypothesis). The marginal likelihood of the alternative hypothesis is calculated by integrating the likelihood with respect to the prior:

$$p(\boldsymbol{Y} \mid H_1) = \int_{\Theta} p(\boldsymbol{Y} \mid \boldsymbol{\theta}, H_1) p(\boldsymbol{\theta} \mid H_1) d\boldsymbol{\theta}. \tag{4.1}$$

Because the posterior model probabilities are sensitive to the prior probabilities of both hypotheses, $p(H_0)$ and $p(H_1)$, it is common practice to quantify the evidence by the ratio of the marginal likelihoods, also known as the *Bayes factor* (Jeffreys, 1961; Dickey, 1971; J. O. Berger & Sellke, 1987; Kass & Raftery, 1995):

$$BF_{10} = \frac{p(\boldsymbol{Y} \mid H_1)}{p(\boldsymbol{Y} \mid H_0)}. \tag{4.2}$$

The Bayes factor is a weighted average likelihood ratio that indicates the relative plausibility of the data under the two competing hypotheses. Another way to conceptualize the Bayes factor is as the change from prior odds $p(H_1)/p(H_0)$ to posterior odds $p(H_1 \mid \boldsymbol{Y})/p(H_0 \mid \boldsymbol{Y})$ brought about by the data. This change is often interpreted as the *weight of evidence* (Good, 1983, 1985). Hence, the Bayes factor quantifies the evidence that the data provide for one model versus another, and as such it represents "the standard Bayesian solution to the hypothesis testing and model selection problems" (Lewis & Raftery, 1997, p. 648.).

When the Bayes factor has a value greater than 1, this indicates that the data are more likely to have occurred under the alternative hypothesis H_1 than under the null hypothesis H_0, and vice versa when the Bayes factor is below 1. For example, when $BF_{10} = 4$, this indicates that the data are four times as likely to have occurred under the alternative hypothesis H_1 than under the null hypothesis H_0. Moreover, when the two hypotheses were equally likely a priori, one may state that the posterior probability in favor of the alternative hypothesis H_1 equals $4/5 = 0.8$.

Jeffreys (1961) proposed a set of verbal labels to categorize different Bayes factors according to their evidential impact. This set of labels, presented in Table 4.1, facilitates scientific communication but should only be considered an approximate descriptive articulation of different standards of evidence (Kass & Raftery, 1995).

4.5 Default Prior Distributions for the Linear Model

In order to calculate the Bayes factor one needs to specify prior distributions for the parameters in H_0 and H_1 (cf. Equation 4.1). A long line of research in Bayesian statistics has focused on finding appropriate default prior distributions, that is, prior distributions that reflect little subjective information and have desirable characteristics. Much of this statistical development has taken place in the framework of linear regression. In order to capitalize on this work, we later restate the correlation test and the partial correlation test as linear regression:

$$\boldsymbol{Y} = \alpha + \beta \boldsymbol{X} + \boldsymbol{\varepsilon}, \tag{4.3}$$

where $\boldsymbol{X}$ is the vector of predictor variables.

For linear regression, one of the most popular priors is known as Zellner's g-prior (Zellner, 1986). This prior corresponds to a normal distribution on the regression coefficients $\boldsymbol{\beta}$, Jeffreys' prior on the precision ϕ (Jeffreys, 1961), and a uniform prior on the

Bayes factor BF_{10}			Interpretation
	$>$	100	Decisive evidence for H_1
30	–	100	Very Strong evidence for H_1
10	–	30	Strong evidence for H_1
3	–	10	Substantial evidence for H_1
1	–	3	Anecdotal evidence for H_1
	1		No evidence
1/3	–	1	Anecdotal evidence for H_0
1/10	–	1/3	Substantial evidence for H_0
1/30	–	1/10	Strong evidence for H_0
1/100	–	1/30	Very Strong evidence for H_0
	$<$	1/100	Decisive evidence for H_0

Table 4.1: Evidence categories for the Bayes factor BF_{10} (Jeffreys, 1961). We replaced the label "not worth more than a bare mention" with "anecdotal".

intercept α:

$$p(\boldsymbol{\beta} \mid \phi, g, \boldsymbol{X}) \sim N\Big(0, \frac{g}{\phi}(X^T X)^{-1}\Big),$$
$$p(\phi) \sim \frac{1}{\phi},$$
$$p(\alpha) \propto 1.$$

Note that the information in the data about $\boldsymbol{\beta}$ can be conceptualized as $\phi^{-1}(X^T X)^{-1}$. Hence, g is a scaling factor controlling the information that we give the prior on $\boldsymbol{\beta}$, relative to the information in the sample. For example, when $g = 1$, the prior carries the same weight as the observed data; when $g = 10$, the prior carries one tenth as much weight as the observed data.

Obviously, the choice of g is crucial to the analysis and much research has gone into choosing an appropriate g. This is a difficult problem: a default prior should not be very informative, but a prior that is too vague can lead to unwanted behavior. Various choices of g have been proposed—a popular setting is $g = n$, the "unit information prior" (Kass & Wasserman, 1995) but others have argued for $g = k^2$ (Foster & George, 1994) or $g = \max\{n, k^2\}$ (Fernandez, Ley, & Steel, 2001). However, the choice for a single g remains difficult.

The impact of the choice of g can be clarified using an example taken from Kanai, Bahrami, Roylance, and Rees (in press) that concerned the correlation between the number of Facebook friends and the normalized grey matter density at the peak coordinate of the right entorhinal cortex. Figure 4.2 shows the data; people with more Facebook friends have higher grey matter density, $r = .48, p < 0.002$. The effect that a specific choice of g has on the Bayes factor for this data set is shown in Figure 4.3. This figure demonstrates that when g is increased, the support for the null hypothesis can be made arbitrarily large. This phenomenon is known as the Jeffreys-Lindley-Bartlett paradox (Bartlett, 1957; Jeffreys, 1961; D. Lindley, 1980; Shafer, 1982; J. O. Berger & Delampady, 1987; Robert, 1993; but see Vanpaemel, 2010). Clearly, one of the primary desiderata for a default Bayesian hypothesis test is to avoid this paradox.

In a different but related approach, Zellner and Siow (1980) extended the work of Jeffreys (1961) and proposed to assign the regression coefficients a multivariate Cauchy

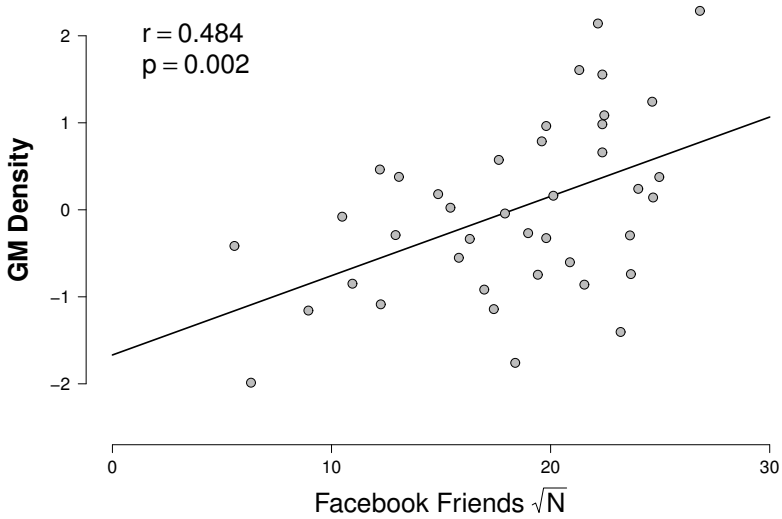


Figure 4.2: The relation between the number of Facebook friends and the normalized grey matter (GM) density at the peak coordinate of the right entorhinal cortex. A positive correlation indicates that people with many Facebook friends have denser grey matter in the right entorhinal cortex. Data replotted from Kanai et al., in press.

prior, with a precision based on the concept of unit information (Liang et al., 2008, p. 414). However, the marginal likelihood for this model specification is not analytically tractable, and therefore this approach did not gain much popularity.

Recently, however, Liang et al. (2008) represented this Jeffreys-Zellner-Siow (JZS) prior as a mixture of g-priors, that is, an Inverse-Gamma$(1/2, n/2)$ prior on g and Jeffreys' prior on the precision ϕ:

$$
\begin{aligned}
p(\boldsymbol{\beta} \mid \phi, g, \boldsymbol{X}) &\propto \int N\Big(0, \frac{g}{\phi}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\Big)p(g)dg \\
p(\phi) &\propto \frac{1}{\phi} \\
p(g) &= \frac{(n/2)^{1/2}}{\Gamma(1/2)} g^{-3/2} e^{-n/(2g)}.
\end{aligned}
$$

From this (mathematically equivalent) perspective, the problem of selecting a single g has been mitigated by assigning g a prior. This above formulation combines the computational advantages of the g-prior with the theoretical advantages of the Cauchy prior (see Liang et al., 2008 for details). Moreover, the mixture representation also facilitates the calculation of the Bayes factor, leaving only one integral that has to be estimated numerically. Note that the same prior set-up underlies the Bayesian JZS t test (Rouder et al., 2009; Wetzels et al., 2009). In the following, we will use this set-up for our correlation and partial correlation test.

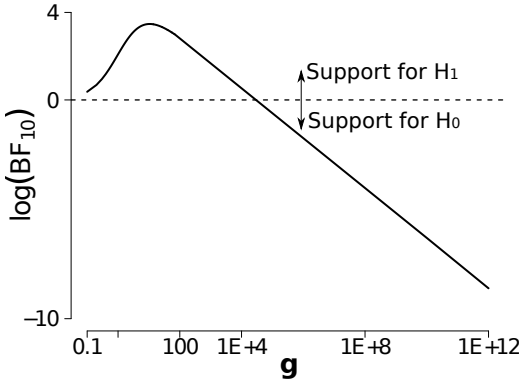


Figure 4.3: An illustration of the Jeffreys-Lindley-Bartlett paradox when the Zellner g prior is applied to the data from Kanai et al. (in press). By increasing g the Bayes factor can be made arbitrarily close to 0, signifying close to infinite support for the null model.

4.6 The JZS Bayes Factor for Correlation and Partial Correlation

In order to calculate the Bayes factor for the JZS (partial) correlation test, we conceptualize these Bayesian tests as a comparison between two regression models, such that the test becomes equivalent to a variable selection test for linear regression (i.e., a test of whether or not the regression coefficient β should be included in the model). This conceptualization allows us to exploit the JZS prior distribution.

Software to compute the two JZS Bayes factors described below are available as an `R` script, a Matlab script, and a description of the steps needed to compute the Bayes factor using the Rouder website.[1] These additional materials are available in the journal archives and on the first author's website.[2]

The JZS Bayes Factor for Correlation

Suppose we have observed data from two variables, $\boldsymbol{X}$ and $\boldsymbol{Y}$, and we are interested in their correlation. Consider the following regression equation:

$$\boldsymbol{Y} = \alpha + \beta \boldsymbol{X} + \boldsymbol{\varepsilon},$$

where α is the intercept, β is the regression coefficient and $\boldsymbol{\varepsilon}$ is the error term, normally distributed with variance σ^2.

Next, we are interested in how well this regression equation fits the data. The standard method to assess this fit is by calculating the coefficient of determination R^2:

$$R^2 = 1 - \frac{SS_{err}}{SS_{tot}},$$

where SS_{err} denotes the residual sum of squares and SS_{tot} denotes the total sum of squares. Note that R^2 is the proportion of variance that is accounted for by the regression

[1] `http://pcl.missouri.edu/bayesfactor`.

[2] `www.ruudwetzels.com`.

model. Specifically, R^2 is an indication of how much better the fit of model $\mathcal{M}_1$ is, when compared to model $\mathcal{M}_0$:

$$\mathcal{M}_0 : \boldsymbol{Y} = \alpha + \epsilon$$
$$\mathcal{M}_1 : \boldsymbol{Y} = \alpha + \beta \boldsymbol{X} + \epsilon.$$

When R^2 is low (i.e., near zero) the addition of the regression coefficient β to $\mathcal{M}_0$ has caused only a small increase in explained variance. As R^2 increases, so does the importance of β. Because R^2 is the square of the sample correlation r, a test for whether or not the correlation equals zero is equivalent to a test for whether or not β equals zero. Hence, the correlation test can be recast as a comparison between two linear regression models, $\mathcal{M}_0$ and $\mathcal{M}_1$ (e.g., Miller & Miller, 2004; Draper & Smith, 1998; Toutenburg & Shalabh, 2009).

The Bayes factor BF_{10} using the JZS prior set-up can then be calculated as follows (see Liang et al., 2008):

$$\begin{aligned} BF_{10} &= \frac{p(\boldsymbol{Y} \mid \mathcal{M}_1)}{p(\boldsymbol{Y} \mid \mathcal{M}_0)} \\ &= \frac{(n/2)^{1/2}}{\Gamma(1/2)} \times \int_0^\infty (1+g)^{(n-2)/2} \times [1 + (1-r^2)g]^{-(n-1)/2} g^{(-3/2)} e^{-n/(2g)} dg. \end{aligned} \tag{4.4}$$

Note that the only input to Equation 4.4 is the usual sample correlation r, and the number of observations n. The resulting Bayes factor BF_{10} quantifies the evidence in favor of the alternative hypothesis. Therefore, Bayes factors greater than 1 indicate evidence for the presence of a correlation and Bayes factors smaller than 1 indicate evidence for the absence of a correlation.

The JZS Bayes Factor for Partial Correlation

Again, we formalize the test as a model selection problem between two regression models. Assume we we have three variables, $\boldsymbol{Y}, \boldsymbol{X}_1$, and $\boldsymbol{X}_2$, and we want to test whether the partial correlation between $\boldsymbol{Y}$ and $\boldsymbol{X}_2$ is zero or not. Analogous to the correlation example, one is interested in whether adding the variable $\boldsymbol{X}_2$ increases R^2 when the variable $\boldsymbol{X}_1$ is already included in the regression model. Hence, we compare the two models (e.g., Draper & Smith, 1998; Abdi, 2003):

$$\mathcal{M}_0 : \boldsymbol{Y} = \alpha + \beta_1 \boldsymbol{X}_1 + \epsilon$$
$$\mathcal{M}_1 : \boldsymbol{Y} = \alpha + \beta_1 \boldsymbol{X}_1 + \beta_2 \boldsymbol{X}_2 + \epsilon.$$

The Bayes factor BF_{10} using the JZS prior set-up can then be calculated as follows:

$$\begin{aligned} BF_{10} &= \frac{p(\boldsymbol{Y} \mid \mathcal{M}_1)}{p(\boldsymbol{Y} \mid \mathcal{M}_0)} \\ &= \frac{\int_0^\infty (1+g)^{(n-1-p_1)/2} \times [1 + (1-r_1^2)g]^{-(n-1)/2} g^{(-3/2)} e^{-n/(2g)} dg}{\int_0^\infty (1+g)^{(n-1-p_0)/2} \times [1 + (1-r_0^2)g]^{-(n-1)/2} g^{(-3/2)} e^{-n/(2g)} dg} \end{aligned} \tag{4.5}$$

Input to Equation 4.5 is the coefficient of determination for H_0 and for H_1 (i.e., r_0^2 and r_1^2), the number of regression coefficients H_0 and H_1 (i.e., p_0 and p_1), and the number of observations n.

The resulting Bayes factor BF_{10} quantifies the evidence in favor of the alternative hypothesis. Therefore, Bayes factors greater than 1 indicate evidence for the presence of a partial correlation and Bayes factors smaller than 1 indicate evidence for the absence of a partial correlation.

Correlation Example: The Meditation Data

In the meditation study, MacLean et al. (2010) tested the hypothesis of a relation between meditation time and visual acuity (see Figure 4.1). The sample correlation between these two variables was found be to $r = -0.36$; the associated p value is .01, significant at the $\alpha = .05$ level.

We can now apply Equation 4.4 to calculate the Bayes factor. We already have the required information: the sample correlation $r = -0.36$, the number of observations $n = 54$. Entering this information in Equation 4.4 yields a Bayes factor $BF_{10} = 3.86$, indicating that the data are 3.86 times more likely to have occurred under H_1 than under H_0, a "substantial" Bayes factor according to the coarse category scheme proposed by Jeffreys. However, note that the factor 3.86 inspires less confidence than the p value, illustrating the well-known point that p values overestimate the evidence against the null (e.g., Edwards et al., 1963; Sellke et al., 2001; Rouder & Morey, 2011; Wetzels et al., 2011).

Correlation Example: The Facebook Data

In the Facebook study, Kanai et al. (in press) investigated the relation between the number of Facebook friends and the normalized grey matter density at the peak coordinate of the right entorhinal cortex (see Figure 4.2). The sample correlation between these two variables was found be to $r = 0.48$; the associated p value is .002, significant at the $\alpha = .05$ level.

We again apply Equation 4.4 to calculate the Bayes factor. We already have the required information: the sample correlation $r = 0.48$, the number of observations $n = 40$. Entering this information in Equation 4.4 yields a Bayes factor $BF_{10} = 17.87$, indicating that the data are 17.87 times more likely to have occurred under H_1 than under H_0, a "strong" Bayes factor according to the coarse category scheme proposed by Jeffreys.

Partial Correlation Example: The Rapid Resumption Data

In the study on rapid resumption, Lleras et al. (2011) tested the partial correlation between search time (X) and rapid resumption (Y) while controlling for age (Z). The partial correlation was found to be $r_{XY|Z} = -0.01$ with a p value of .95.

We can compute the Bayes factor using Equation 4.5, using the coefficient of determination for both models. The null model $\mathcal{M}_0$ regresses search time (X) on age (Y), containing only the regression coefficient for Y. Hence, $p_0 = 1$ and $r^2 = 0.6084$. The alternative model $\mathcal{M}_1$ contains the regression coefficients for Y and Z. Hence, $p_1 = 2$ and $r^2 = 0.6084408$. The sample size n is 40.

The Bayes factor BF_{10} is 0.16, indicating substantial evidence in favor of the null hypothesis: the data are $1/0.16 \approx 6.25$ times as likely to have occurred under the null hypothesis than under the alternative hypothesis (see Table 4.1).

4.7 Concluding Remarks

In this article we outlined a default Bayesian test for correlation and partial correlation. Just as the default Bayesian t test (Rouder et al., 2009; Wetzels et al., 2009), the correlation test follows directly from the regression framework for variable selection proposed by Liang et al. (2008). We did not strive for new statistical development. Instead, our goal

was to show experimental psychologists how they can obtain a default Bayesian hypothesis test for correlation and partial correlation. As we mentioned throughout this article, the Bayesian hypothesis test comes with important practical advantages compared to the standard frequentist test; for instance, the Bayesian hypothesis test can quantify evidence in favor of the null hypothesis, and allows researchers to collect data until a point has been proven or disproven.

It should be noted that Jeffreys (1961, pp. 289-292) also proposed a Bayesian correlation test, one that differs slightly from the one outlined here. We prefer the JZS correlation test because it follows directly from the regression framework by Liang et al., 2008, incorporating modern Bayesian developments into a more general JZS testing framework. This JZS framework now encompasses linear regression, the t test and (partial) correlation. We have also extended the JZS framework to ANOVA, and extensions to other popular statistical models are likely to follow.

By making default Bayes factors easily available to experimental psychologists, we hope and expect that the field will start to turn away from p values and move towards a Bayesian assessment of evidence. This transition is bound to improve statistical inference and accelerate scientific progress.

5 A Default Bayesian Hypothesis Test for ANOVA Designs

Abstract

This article presents a Bayesian hypothesis test for ANOVA designs. The test is an application of standard Bayesian methods for variable selection in regression models. 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.

An excerpt of this chapter has been published as:
Wetzels, R., Grasman, R.P.P.P, & Wagenmakers, E.-J. (in press). A Default Bayesian Hypothesis Test for ANOVA Designs. *The American Statistician*.

5.1 Introduction

Bayesian methods have become increasingly popular in almost all scientific disciplines (e.g.,Poirier, 2006). One important reason for this gain in popularity is the ease with which Bayesian methods can be applied to relatively complex problems involving, for instance, hierarchical modeling or the comparison between non-nested models. However, Bayesian methods can also be applied in simpler statistical scenarios such as those that feature basic testing procedures. Prominent examples of such procedures include ANOVA and the t test; these tests are the cornerstone of data analysis in fields such as biology, economics, sociology, and psychology.

Because Bayesian methods have become more mainstream in recent years, most technically oriented studies now offer at least one course on Bayesian inference in their graduate or undergraduate program. Our own experience in teaching one such course is that students often ask the same questions when Bayesian model selection and hypothesis testing are introduced. First, students are interested to know how they can apply Bayesian methods to testing problems that they face on a regular basis; second, students want to know how prior distributions can be chosen such that a test can be considered default. In this paper we address both questions. We apply the Bayesian method to ANOVA designs and explain the rationale and impact of several default prior distributions.

Thus, the first goal of this article is to show how the Bayesian framework of hypothesis testing with the Bayes factor can be carried out in ANOVA designs. ANOVA is one of the most popular statistical methods to assess whether or not two or more population means are equal—in most experimental settings, ANOVA is used to test for the presence of a treatment effect. Because of its importance and simplicity, ANOVA is taught in virtually every applied statistics course. Nevertheless, the Bayesian hypothesis testing literature on ANOVA is scant; the dominant treatment of ANOVA is still classical or frequentist (e.g., Draper & Smith, 1998; Faraway, 2002), and, although the Bayesian treatment of ANOVA is gaining popularity (e.g., Gelman, Carlin, Stern, & Rubin, 2004; Qian & Shen, 2007; Ntzoufras, 2009; Kaufman & Sain, 2010), the latter has dealt almost exclusively with estimation, not testing (for exceptions, see Westfall & Gönen, 1996; Sen & Churchill, 2001; Ishwaran & Rao, 2003; Ball, 2005; Gönen, Johnson, Lu, & Westfall, 2005; Maruyama, 2009). This is all the more surprising because Bayesian hypothesis testing has been well developed for variable selection in regression models (e.g., Liang et al., 2008), of which ANOVA is a special case.

The second goal of this article is to describe the rationale behind a particular family of default priors—g-priors—and to use these g-priors for default Bayesian tests for ANOVA designs. We hope this work shows students and experimental researchers how Bayesian hypothesis tests can be a valid and practical alternative to classical or frequentist tests.

The outline of this paper is as follows. In the first section we briefly cover Bayesian estimation and Bayesian model selection. In the second section we describe the various g-priors that have been proposed in the literature on variable selection in regression models. Finally, we present two worked examples that show how the regression framework can be applied to one-way and two-way ANOVA designs.

5.2 Bayesian Inference

Bayesian estimation

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

Suppose we have a model $\mathcal{M}$ and we wish to estimate the model parameters $\boldsymbol{\theta}$. Then, we have to define a *prior distribution* over these parameters; $p(\boldsymbol{\theta} \mid \mathcal{M})$. When data $\boldsymbol{Y}$ come in, this prior distribution $p(\boldsymbol{\theta} \mid \mathcal{M})$ is updated to yield the *posterior distribution* $p(\boldsymbol{\theta} \mid \boldsymbol{Y}, \mathcal{M})$ according to Bayes' rule:

$$\begin{aligned} p(\boldsymbol{\theta} \mid \boldsymbol{Y}, \mathcal{M}) &= \frac{p(\boldsymbol{Y} \mid \boldsymbol{\theta}, \mathcal{M}) p(\boldsymbol{\theta} \mid \mathcal{M})}{p(\boldsymbol{Y} \mid \mathcal{M})} \\ &= \frac{p(\boldsymbol{Y} \mid \boldsymbol{\theta}, \mathcal{M}) p(\boldsymbol{\theta} \mid \mathcal{M})}{\int_\Theta p(\boldsymbol{Y} \mid \boldsymbol{\theta}, \mathcal{M}) p(\boldsymbol{\theta} \mid \mathcal{M}) d\boldsymbol{\theta}} \\ &\propto p(\boldsymbol{Y} \mid \boldsymbol{\theta}, \mathcal{M}) p(\boldsymbol{\theta} \mid \mathcal{M}). \end{aligned}$$

Hence, the posterior distribution of $\boldsymbol{\theta}$ is proportional to the likelihood times the prior. In Bayesian parameter estimation, the researcher is interested in the posterior distribution of the model parameters $p(\boldsymbol{\theta} \mid \boldsymbol{Y}, \mathcal{M})$. However, in Bayesian model selection the focus is on $p(\boldsymbol{Y} \mid \mathcal{M})$, the marginal likelihood of the data under model $\mathcal{M}$.

Bayesian model selection

In Bayesian model selection, competing statistical models or hypotheses are assigned prior probabilities. Consider two competing models, $\mathcal{M}_1$ and $\mathcal{M}_2$ with prior probabilities $p(\mathcal{M}_1)$ and $p(\mathcal{M}_2)$.

After observing the data, the relative plausibility of $\mathcal{M}_1$ and $\mathcal{M}_2$ is given by the ratio of posterior model probabilities, that is, the posterior odds:

$$\frac{p(\mathcal{M}_1 \mid \boldsymbol{Y})}{p(\mathcal{M}_2 \mid \boldsymbol{Y})} = \frac{p(\mathcal{M}_1)}{p(\mathcal{M}_2)} \frac{p(\boldsymbol{Y} \mid \mathcal{M}_1)}{p(\boldsymbol{Y} \mid \mathcal{M}_2)}.$$

Hence, the posterior odds are given by the product of the prior odds and the ratio of marginal likelihoods. The latter component is known as the *Bayes factor* (Jeffreys, 1961; Dickey, 1971; J. O. Berger & Sellke, 1987; Kass & Raftery, 1995) and quantifies the change from prior to posterior odds; therefore, the Bayes factor does not depend on the prior model probabilities $p(\mathcal{M}_1)$ and $p(\mathcal{M}_2)$ and quantifies the evidence that the data provide for $\mathcal{M}_1$ versus $\mathcal{M}_2$.

In linear regression and ANOVA, two models of special interest are the null model, $\mathcal{M}_N$, that does not include any of the predictors (but does include the intercept) and the full model, $\mathcal{M}_F$, that includes all relevant predictors. In this scenario, the main difficulty with the Bayes factor is its sensitivity to the prior distribution for the model parameters under test (Press, Chib, Clyde, Woodworth, & Zaslavsky, 2003; J. Berger, 2006; Gelman, 2008).

When there is limited knowledge about the phenomenon under study, the prior distribution for the parameters should be relatively uninformative. However, in order to avoid paradoxical results, the prior distribution cannot be *too* uninformative. In particular, the Jeffreys-Lindley-Bartlett paradox (Bartlett, 1957; Jeffreys, 1961; D. Lindley, 1980; Shafer, 1982; J. O. Berger & Delampady, 1987; Robert, 1993) shows that with vague uninformative priors on the parameters under test the Bayes factor will strongly support the null model. The reason is that the marginal likelihood $p(\boldsymbol{Y} \mid \mathcal{M})$ is obtained by averaging the likelihood over the prior; when the prior is very spread out relative to the data, a large part of the prior distribution is associated with very low likelihoods, decreasing the average. This paradox is illustrated in Figure 5.3 and will be discussed later in the context of a specific model. The next section details how, in the context of

linear regression and ANOVA, one can avoid the Jeffreys-Lindley-Bartlett paradox and nevertheless define prior distributions that are reasonably uninformative.

5.3 Linear Regression, ANOVA, and the Specification of g-Priors

The prior distributions that we will discuss are applicable to model selection in the regression framework. Assume a response vector $\boldsymbol{Y}$ of length n, $\boldsymbol{Y} = (y_1, ..., y_n)^T$, normally distributed with mean vector $\boldsymbol{\mu} = (\mu_1, ..., \mu_n)^T$, precision ϕ, and $\boldsymbol{I}_n$ an $n \times n$ identity matrix,

$$\boldsymbol{Y} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{I}_n/\phi).$$

The mean $\boldsymbol{\mu}$ can be decomposed into an overall common intercept α and the regression coefficients $\boldsymbol{\beta}$. The mean $\boldsymbol{\mu}$ then becomes:

$$\boldsymbol{\mu} = \mathbf{1}_n\alpha + \boldsymbol{X}\boldsymbol{\beta},$$

where $\boldsymbol{X}$ represents the $n \times k$ design matrix and $\boldsymbol{\beta}$ is the k-dimensional vector of regression coefficients.

In the ANOVA setting, the independent variables that are controlled in the experiment are called *factors*, which in turn can have different levels of intensity. Then, the regression coefficients are interpreted as level-specific parameters. The design matrix $\boldsymbol{X}$ is constructed using dummy coding (Draper & Smith, 1998). Because the matrix $[\mathbf{1}_n, \boldsymbol{X}]$ does not necessarily have full column rank, we need to add a constraint. Here we adopt the sum-to-zero constraint. By using this constraint, the intercept is the grand mean, and each regression coefficient describes the deviation from this grand mean—consequently, the regression coefficient of the last level equals minus the sum of the other regression coefficients.

In the one-way ANOVA, we examine the effect of a categorical variable $\boldsymbol{X}$ on the continuous response variable $\boldsymbol{Y}$. The null hypothesis is defined as H_0; all levels have the same mean, and the alternative hypothesis is defined as H_1; at least one of the levels has a different mean.

We can translate this frequentist test to a Bayesian model selection situation by comparing the model with all relevant regression coefficients to the model without these coefficients. In the remainder of this article we focus on the one-way and two-way ANOVA and show how these tests can be carried out in a Bayesian fashion.

The sections below list three default prior distributions. We focus on prior distributions for variable selection in regression as this framework provides the basis for the analysis of ANOVA designs (for more information on Bayesian variable selection see Leamer, 1978; Zellner, 1986, 1987; Mitchell & Beauchamp, 1988; Chipman, 1996; George & McCulloch, 1997). The subsections below detail, in historical order, three versions of the popular g-prior.

Zellner's g-prior

In the case of linear regression, Zellner's g-prior (Zellner, 1986) corresponds to:

$$p(\boldsymbol{\beta} \mid \phi, g, \boldsymbol{X}) \sim N\Big(0, \frac{g}{\phi}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\Big) \quad g > 0,$$

with Jeffreys' prior (Jeffreys, 1961) on the precision:

$$p(\phi) \propto \frac{1}{\phi},$$

and a flat prior on the common intercept α. Note that we assume that the columns of $\boldsymbol{X}$ are centered so that $\mathbf{1}_n^T \boldsymbol{X} = 0$.

This set of prior distributions is of the conjugate Normal-Gamma family, and therefore the marginal likelihood can be calculated analytically. When the design matrix is considered fixed, we are allowed to use it in our prior variance term as $\frac{g}{\phi}(\boldsymbol{X}^T\boldsymbol{X})^{-1}$. Recall that the variance of the maximum likelihood estimator for $\boldsymbol{\beta}, \text{var}(\hat{\boldsymbol{\beta}})$, equals $\phi^{-1}(\boldsymbol{X}^T\boldsymbol{X})^{-1}$. Hence, the term g is a scaling factor for the prior: if we choose g to be 1, we give the prior the same weight as the sample; if we choose g to be 2, the prior is half as important as the sample; if we choose g to be n, the prior is $1/n^{th}$ as important as the sample.

An obvious problem with this prior distribution is how to set parameter g. If g is set low, then the prior distribution for $\boldsymbol{\beta}$ is relatively peaked and informative. If g is set high then this prior is relatively spread out and uninformative. However, as described in the previous section, a prior that is too vague can result in the Jeffreys-Lindley-Bartlett paradox.

Various settings for g have been studied and proposed. A popular setting is $g = n$, corresponding to the so-called "unit information prior". The intuition is that this prior contains as much information as present in a single observation (Kass & Wasserman, 1995); the argument is that the precision of the sample estimate of $\boldsymbol{\beta}$ contains the information of n observations. Then the amount of information in an imaginary single observation is this quantity divided by n, hence $g = n$. Another well-known choice of g is to set it equal to the square of the number of predictors of the regression model: $g = k^2$ (i.e., the Risk Inflation Criterion, Foster & George, 1994). Furthermore, Fernandez et al. (2001) suggested to take $g = \max\{n, k^2\}$ as a "benchmark prior".

A quantity of interest is the so-called shrinkage factor $g/(g+1)$. It can be used to estimate the posterior mean of $\boldsymbol{\beta}$, which is the least squares estimate of $\boldsymbol{\beta}$ multiplied by the shrinkage factor:

$$\mathbb{E}\left[\boldsymbol{\beta} \mid \boldsymbol{Y}, \boldsymbol{X}, \mathcal{M}, g\right] = \frac{g}{g+1}\hat{\boldsymbol{\beta}},$$

where $\hat{\boldsymbol{\beta}}$ is the least squares estimate of $\boldsymbol{\beta}$. A low value of g pulls the posterior mean of $\boldsymbol{\beta}$ to zero, whereas a high value of g yields results similar to the least squares estimate. Note that, somewhat confusingly, a low shrinkage factor means more shrinkage and vice versa.

In order to compute the Bayes factor in the one-way ANOVA design, we compare the full model, $\mathcal{M}_F$ to the null model, $\mathcal{M}_N$. Then, the Bayes factor is given by:

$$\text{BF}[\mathcal{M}_F : \mathcal{M}_N] = (1+g)^{(n-k-1)/2}[1 + g(1 - R^2)]^{-(n-1)/2}, \tag{5.1}$$

where k equals the number of predictors of $\mathcal{M}_F$, n is the sample size, and R^2 the coefficient of determination of $\mathcal{M}_F$ (Note that R^2 for $\mathcal{M}_N$ equals zero as it contains no predictors).

Equation 5.1 shows that, in its general formulation, Zellner's g-prior is potentially vulnerable to the Jeffreys-Lindley-Bartlett paradox: when $g \to \infty$ with n and k fixed, the Bayes factor $\text{BF}[\mathcal{M}_F : \mathcal{M}_N]$ will go to 0, favoring the null model regardless of the observed data (see Figure 5.3 for an example).

Another problem with the Zellner g-prior is that, when the evidence in favor of the full model goes to infinity (i.e., R^2 goes to 1), the Bayes factor converges to the upper bound

$(1+g)^{(n-k-1)/2}$. Liang et al. (2008) term this undesirable property the "information paradox".

Jeffreys-Zellner-Siow prior

To test whether a parameter μ is zero or non-zero (with μ the mean of a normal distribution), (Jeffreys, 1961, pp. 268-270) suggested to apply a Cauchy prior. The Cauchy prior was the simplest distribution to satisfy consistency requirements that Jeffreys considered important for hypothesis testing. One such requirement is that a researcher does not want to favor one model over another on the basis of a single datum.

Extending Jeffreys' suggestion to variable selection in the regression model, (Zellner & Siow, 1980) proposed a multivariate Cauchy prior on the regression coefficients and a flat prior on the common intercept. However, as the marginal likelihood is not analytically tractable, this approach did not gain much popularity.

Recently, however, Liang et al. (2008) represented the Jeffreys-Zellner-Siow (JZS) prior as a mixture of g-priors, that is, an Inverse-Gamma$(1/2, n/2)$ prior on g and Jeffreys' prior on the precision ϕ:

$$
\begin{aligned}
p(\phi) &\propto \frac{1}{\phi} \\
p(\boldsymbol{\beta} \mid \phi, g, \boldsymbol{X}) &\propto \int N\Big(0, \frac{g}{\phi}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\Big)p(g)dg \\
p(g) &= \frac{(n/2)^{1/2}}{\Gamma(1/2)} g^{-3/2} e^{-n/(2g)}.
\end{aligned}
$$

This formulation combines the computational advantages of the g-prior with the statistical advantages of the Cauchy prior. Note that again we assume that the columns of $\boldsymbol{X}$ are centered.

By assigning a prior to g, we avoid having to assign g a specific value; moreover, the prior on g allows us to estimate g from the data and obtain data-dependent shrinkage. Equation 5.2 gives the expected value of the shrinkage factor $g/(g+1)$ with the JZS approach:

$$
\mathbb{E}\left[\frac{g}{g+1} \mid \boldsymbol{Y}, \mathcal{M}\right] = \frac{\int_0^\infty (1+g)^{(n-k-3)/2} \times [1+g(1-R^2)]^{-(n-1)/2} g^{(-1/2)} e^{-n/(2g)} dg}{\int_0^\infty (1+g)^{(n-k-1)/2} [1+g(1-R^2)]^{-(n-1)/2} g^{-3/2} e^{-n/(2g)} dg}. \tag{5.2}
$$

It can be seen from equation (5.2), and later from equation (5.4), that the expected value of $g/(g+1)$ increases with R^2 (Zeugner & Feldkircher, 2009). Hence, there is less shrinkage when more variance is explained by the model.

In the JZS approach, the Bayes factor comparing the full model to the null model is:

$$
\begin{aligned}
&BF[\mathcal{M}_F : \mathcal{M}_N] = \\
&\frac{(n/2)^{1/2}}{\Gamma(1/2)} \times \int_0^\infty (1+g)^{(n-k-1)/2} [1+g(1-R^2)]^{-(n-1)/2} g^{-3/2} e^{-n/(2g)} dg. \qquad (5.3)
\end{aligned}
$$

As pointed out by Liang et al. (2008), the integral is one-dimensional and easily approximated using standard software packages such as R (R Development Core Team, 2004).

A drawback of the JZS prior is that the Bayes factor is not analytically available. However, the JZS prior is not vulnerable to the Jeffreys-Lindley-Bartlett paradox nor to the information paradox (Liang et al., 2008).

Hyper-g priors

As an alternative to the JZS prior, Liang et al. (2008) proposed a family of prior distributions on g and termed this the hyper-g approach:

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

which is a proper distribution if $a > 2$ (Strawderman, 1971; Cui & George, 2008). Because this distribution leads to indeterminate Bayes factors when $a \leq 2$, Liang et al. (2008) study the behavior of this prior for $2 < a \leq 4$. Interestingly, this family of priors on g corresponds to the following prior on the shrinkage factor $g/(1+g)$:

$$\frac{g}{1+g} \sim \text{Beta}\Big(1, \frac{a}{2} - 1\Big).$$

By choosing a, one can tune the prior on the shrinkage factor. When $a = 4$, the prior is uniform between 0 and 1, whereas when a is very close to 2, the prior distribution for the shrinkage factor will have most mass near 1. Figure 5.1 shows the effect of various a on the prior distribution for the shrinkage factor $g/(g+1)$. Furthermore, Dellaportas, Forster, and Ntzoufras (in press) showed that the posterior densities of the parameters are, in terms of posterior shrinkage, insensitive to the choice of a within the recommended range. Only for very high values of a (in their simple linear regression example, $a \approx 20$) was posterior shrinkage considerable.

The expected value of the shrinkage factor $g/(g+1)$ with the hyper-g approach is:

$$\mathbb{E}\left[\frac{g}{g+1} \mid \boldsymbol{Y}, \mathcal{M}\right] = \frac{2}{k+a}\frac{{}_2F_1[(n-1)/2, 2; (k+a)/2+1; R^2]}{{}_2F_1[(n-1)/2, 1; (k+a)/2; R^2]}, \tag{5.4}$$

where ${}_2F_1(a, b; c; z)$ is the Gaussian hypergeometric function (Abramowitz & Stegun, 1972):

$${}_2F_1(a, b; c; z) = \frac{\Gamma(c)}{\Gamma(c-b)\Gamma(b)}\int_0^1 \frac{t^{b-1}(1-t)^{c-b-1}}{(1-tz)^a}dt \quad c > b > 0$$

Just as with the JZS prior, the hyper-g approach estimates g and allows for data-dependent shrinkage.

In order to compare the two models that are important in the one-way ANOVA design, we need to calculate the Bayes factor[1]:

$$BF[\mathcal{M}_F : \mathcal{M}_N] = \frac{a-2}{2}\int_0^\infty (1+g)^{(n-k-1-a)/2} \times [1 + g(1-R^2)]^{-(n-1)/2}dg. \tag{5.5}$$

Just as with the with JZS prior, the hyper-g approach is not vulnerable to the Jeffreys-Lindley-Bartlett paradox, nor to the information paradox (when $a \leq n - k + 1$, Liang et al., 2008).

5.4 A Bayesian One-Way ANOVA

To illustrate the differences between the various priors and the effects they have on the Bayes factor for ANOVA designs, we first discuss the one-way ANOVA. We follow (Box

[1]Note that this Bayes factor is also available in closed form using the Gaussian hypergeometric function.

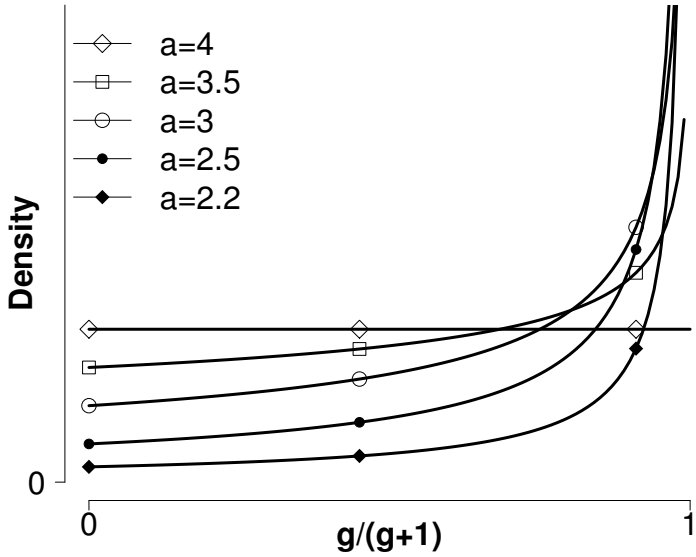


Figure 5.1: Effect of parameter a on the shrinkage factor $g/(g+1)$. When $a = 4$, the prior is uniform between 0 and 1, whereas when a is very close to 2, the prior distribution for the shrinkage factor has most mass near 1. Higher values for $g/(g+1)$ result in less shrinkage.

& Tiao, 1973) and use example data from an experiment that was set up to investigate to what extent yield of dyestuff differs between batches. The experiment featured six batches with five observations each. Figure 5.2 shows the box and whisker plot of yield of dyestuff for the different batches. The left plot shows the original data from Box and Tiao (1973). In order to illustrate the behavior of the Bayes factor when the null hypothesis is true, the right plot shows the same data but with equal means (i.e., the difference between the batch mean and the overall mean was subtracted from the batch data).

First we carried out a classical one-way ANOVA to compute the F statistic and the corresponding p value for both data sets. For the original data set, we compute $F(5, 24) = 4.60, p = 0.004$, suggesting that at least one of the batches has a different yield. In the modified data set with equal means, we compute $F(5, 24) = 0, p = 1$, suggesting that the yield of dyestuff is equal for all batches, although such an inference in favor of the null hypothesis is not warranted in the Fisherian framework of p value significance testing.

Next we designed a Bayesian hypothesis test to contrast two models. The full model, $\mathcal{M}_F$, contains a grand mean α and the predictors for batches 1-5. The predictor for batch 6 is omitted because of the sum-to-zero constraint. The null model, $\mathcal{M}_N$, contains no predictors. Therefore, our test concerns the following two models:

$$\mathcal{M}_F : \boldsymbol{\mu} = \mathbf{1}_n\alpha + \boldsymbol{X}_1\beta_1 + \boldsymbol{X}_2\beta_2 + \boldsymbol{X}_3\beta_3 + \boldsymbol{X}_4\beta_4 + \boldsymbol{X}_5\beta_5$$
$$\mathcal{M}_N : \boldsymbol{\mu} = \mathbf{1}_n\alpha.$$

The results from the Bayesian hypothesis test for the data with unequal group means,

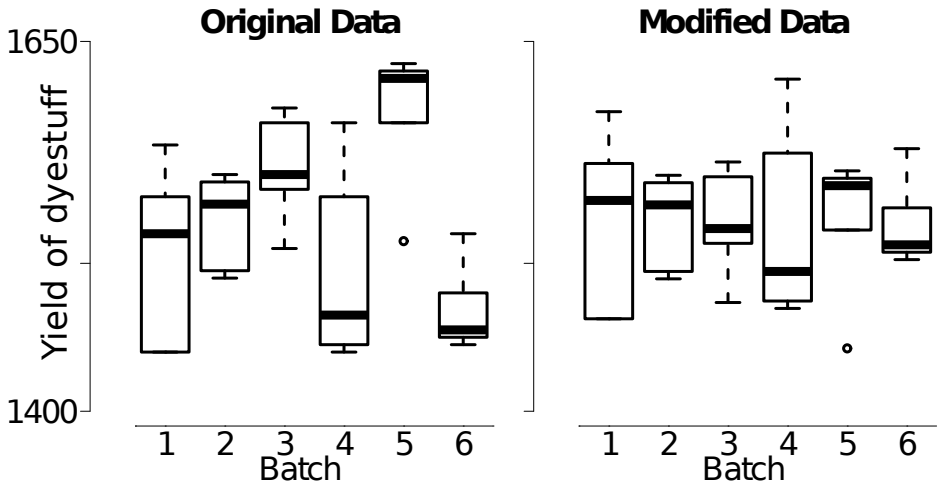


Figure 5.2: Boxplots of yield of dyestuff per batch. The left plot (original data) shows the original data. The right plot (modified data) shows the same data but with the difference between the batch mean and the overall mean subtracted from the batch data.

reported in Table 5.1, show that the two Zellner g-priors and the JZS prior yield only modest Bayes factor support in favor of $\mathcal{M}_F$; the two hyper-g priors yield more convincing support in favor of $\mathcal{M}_F$: overall, the results suggest that the data may be too sparse to allow an unambiguous conclusion. Importantly, a Bayes factor of 3 arguably does not inspire as much confidence as one would glean from a p value as low as .004 (J. O. Berger & Sellke, 1987). This result highlights the general conflict between Bayes factors and p values in terms of their evidential impact (e.g., Sellke et al., 2001; Edwards et al., 1963).

When the models are compared using the modified data, Table 5.1 shows that the two Zellner g-priors and the JZS prior yield considerable Bayes factor support in favor of the null model $\mathcal{M}_N$; the two hyper-g priors also provide evidence in favor of $\mathcal{M}_N$, albeit less extreme. Moreover, the relation between R^2 and the shrinkage factor now becomes clear: for each prior where g is estimated (i.e., JZS, hyper-g with a=3, and hyper-g with a=4), the shrinkage factor is lower when the null model is preferred, as is the case for the modified data.

Finally we use the original dyestuff data with unequal means to illustrate the Jeffreys-

	Unequal means		Equal means	
Prior	$BF_{F:N}$	$\mathbb{E}[g/(g+1) \mid \boldsymbol{Y}]$	$BF_{F:N}$	$\mathbb{E}[g/(g+1) \mid \boldsymbol{Y}]$
Zellner g=n	2.0	0.97	1.87×10^{-4}	0.97
Zellner g=k^2	2.9	0.96	2.90×10^{-4}	0.96
JZS	3.1	0.90	8.51×10^{-4}	0.86
Hyper-g a=3	9.9	0.71	0.17	0.25
Hyper-g a=4	10.1	0.65	0.29	0.22

Table 5.1: Bayes factors and shrinkage factors for the one-way ANOVA example on the dyestuff data, see Figure 5.2. The Bayes factor compares the full model to the null model, testing for a main effect of batch.

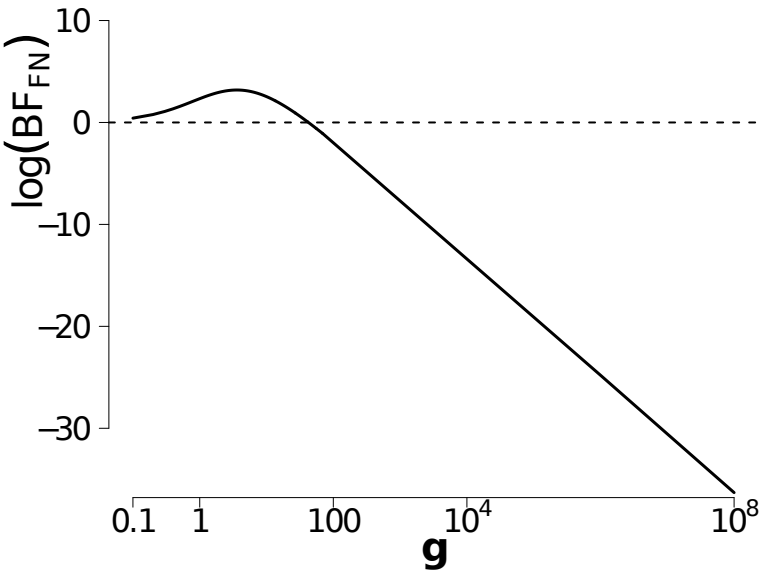


Figure 5.3: An illustration of the Jeffreys-Lindley-Bartlett paradox when the Zellner g prior is applied to the dyestuff data. When g increases from 1 to 4, the Bayes factor in favor of the full model increases as well. By increasing g much further the Bayes factor can be made arbitrarily close to 0, signifying infinite support for the null model.

Lindley-Bartlett paradox for the one-way ANOVA model. Under Zellner's g-prior with $g = n$ or $g = k^2$, the Bayes factor was in favor of the full model. However, Figure 5.3 shows that by increasing g the Bayes factor can be made arbitrarily close to 0, indicating impressive evidence in favor of the null model.

5.5 A Bayesian Two-Way ANOVA

In order to illustrate the Bayesian two-way ANOVA we use a slightly more complex example from Faraway, 2002. As part of an investigation of toxic agents, a total of 48 rats were allocated to three poisons (I, II, III) and four treatments (A, B, C, D). The dependent variable is the reciprocal of the survival time in tens of hours, which can be interpreted as the rate of dying. Figure 5.4 shows the box-and-whisker plot of the survival times in the different experimental conditions.

First we carried out a classical two-way ANOVA to compute the F statistics and the corresponding p values. First we investigate whether the interaction terms should be incorporated in the model. We compute $F(6, 36) = 1.1, p \approx 0.39$, suggesting that poison and treatment do not interact, although, again, such inference in favor of the null hypothesis is not warranted in the Fisherian framework of p value significance testing.

Because the interaction effects were not significant, we remove them from the model. Then, for the main effect of treatment, we compute $F(3, 42) = 27.9, p < 0.001$, suggesting that at least one of the treatments has an effect on rate of dying. For the main effect of poison we compute $F(2, 42) = 71.7, p < 0.001$, suggesting that at least one of the poisons has an effect on rate of dying.

Again we compare the classical results to the Bayesian alternatives. We define the necessary models that are needed to test for each of the main effects and for the interaction

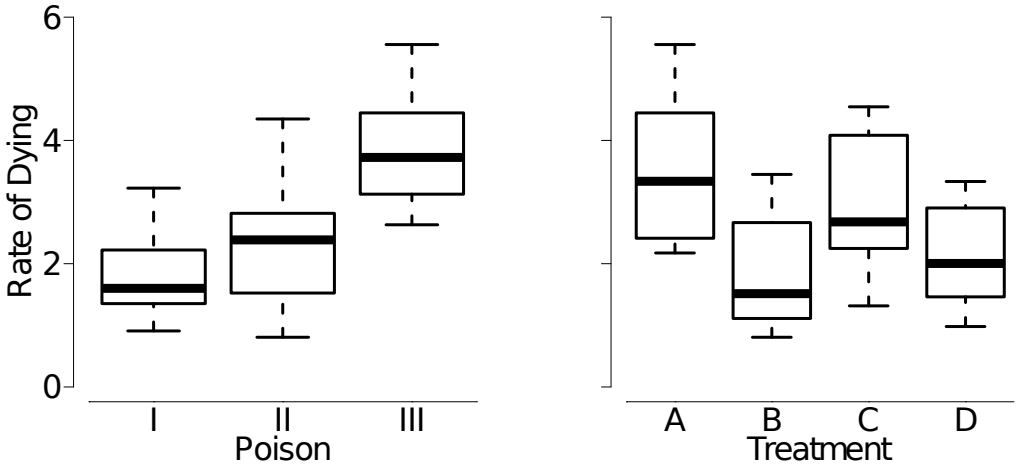


Figure 5.4: Rate of dying per poison and per treatment. Poison group I (the reference level for poison) has a mean of 1.80; the means of groups II and III are 0.47 and 2.00 higher, respectively. Treatment group A (the reference level for treatment) has a mean of 3.52; the means of groups B, C, and D are 1.66, 0.57 and 1.36 lower, respectively.

effects. To test for the effect of the interaction terms we define two models: the full model containing the main and interaction effects $\mathcal{M}_{PT}$, and the same model without the interaction effects $\mathcal{M}_{P+T}$. To test for the main effects we define the no-interaction model with the effects of treatment $\mathcal{M}_T$; the no-interaction model with the effects of poison $\mathcal{M}_P$; and the null model $\mathcal{M}_N$.

$$
\begin{aligned}
\mathcal{M}_{PT} : \boldsymbol{\mu} &= \mathbf{1}_n\alpha + \boldsymbol{X}_I\beta_I + \boldsymbol{X}_{II}\beta_{II} + \boldsymbol{X}_A\beta_A + \boldsymbol{X}_B\beta_B + \boldsymbol{X}_C\beta_C \\
&\quad + \boldsymbol{X}_{I\times A}\beta_{I\times A} + \boldsymbol{X}_{I\times B}\beta_{I\times B} + \boldsymbol{X}_{I\times C}\beta_{I\times C} \\
&\quad + \boldsymbol{X}_{II\times A}\beta_{II\times A} + \boldsymbol{X}_{II\times B}\beta_{II\times B} + \boldsymbol{X}_{II\times C}\beta_{II\times C} \\
\mathcal{M}_{P+T} : \boldsymbol{\mu} &= \mathbf{1}_n\alpha + \boldsymbol{X}_I\beta_I + \boldsymbol{X}_{II}\beta_{II} + \boldsymbol{X}_A\beta_A + \boldsymbol{X}_B\beta_B + \boldsymbol{X}_C\beta_C \\
\mathcal{M}_T : \boldsymbol{\mu} &= \mathbf{1}_n\alpha + \boldsymbol{X}_A\beta_A + \boldsymbol{X}_B\beta_B + \boldsymbol{X}_C\beta_C \\
\mathcal{M}_P : \boldsymbol{\mu} &= \mathbf{1}_n\alpha + \boldsymbol{X}_I\beta_I + \boldsymbol{X}_{II}\beta_{II} \\
\mathcal{M}_N : \boldsymbol{\mu} &= \mathbf{1}_n\alpha
\end{aligned}
$$

Prior	$BF_{PT:P+T}$	$BF_{P+T:P}$	$BF_{P+T:T}$
Zellner g=n	2.61×10^{-04}	6.87×10^{07}	3.09×10^{12}
Zellner g=k^2	1.45×10^{-05}	3.41×10^{08}	4.36×10^{11}
JZS	5.37×10^{-04}	4.52×10^{07}	1.24×10^{12}
Hyper-g a=3	9.41×10^{-04}	2.95×10^{07}	1.81×10^{11}
Hyper-g a=4	1.34×10^{-03}	2.07×10^{07}	6.72×10^{10}

Table 5.2: Bayes factors for the two-way ANOVA for the rats data set from plotted in Figure 5.4. The Bayes factor compares the relevant models to each other in order to test for main effects of poison and treatment, and their interaction.

We compare the reduced models to the larger model in order to test for the effect of the predictors that were left out. If the larger model is preferred over the reduced model then the tested effects matter. However, these models cannot be compared directly using the methods outlined above, as these methods always feature the null model. Instead, we first calculate the Bayes factor comparing the larger model $\mathcal{M}_L$ to the null model: $BF[\mathcal{M}_L : \mathcal{M}_N]$, and the reduced model $\mathcal{M}_R$ to the null model: $BF[\mathcal{M}_R : \mathcal{M}_N]$. The desired Bayes factor, $BF[\mathcal{M}_L : \mathcal{M}_R]$, is then obtained by taking the ratio of Bayes factors:

$$BF[\mathcal{M}_L : \mathcal{M}_R] = \frac{BF[\mathcal{M}_L : \mathcal{M}_N]}{BF[\mathcal{M}_R : \mathcal{M}_N]}.$$

We do not present the shrinkage factors because the model comparison is not between the null model and the full model but between two models with many predictors each.

A test for the interaction involves the comparison between $\mathcal{M}_{PT}$ and $\mathcal{M}_{P+T}$. Table 5.2 shows the results for the Bayes factors that test for the presence of the interaction terms. The different priors do not change the overall conclusion: all priors support the model without the interaction terms. Hence, we drop the interaction terms from the ANOVA model and proceed with the main effects only.

By comparing $\mathcal{M}_{P+T}$ to $\mathcal{M}_P$ we can test for a main effect of treatment. Table 5.2 shows that all Bayesian hypothesis tests favor the model that includes the treatment effect, regardless of the specific choice of prior distribution.

By comparing $\mathcal{M}_{P+T}$ to $\mathcal{M}_T$ we can test for a main effect of poison. The Bayesian hypothesis tests show that all methods favor the full model over the null model, regardless of the specific choice of prior distribution (see Table 5.2). The support for the model with a main effect of poison is considerably higher then the support for the main effect for treatment.

5.6 Conclusion

ANOVA is one of the most often-used statistical methods in the empirical sciences. However, Bayesian hypothesis tests are rarely conducted in ANOVA designs; instead, most theoretical development has concerned the more general problem of selecting variables in regression models (e.g.,Kuo & Mallick, 1998; Mitchell & Beauchamp, 1988; George & McCulloch, 1997; Casella & Moreno, 2006; O'Hara & Sillanpää, 2009). Here we showed how the regression framework can be seamlessly carried over to ANOVA designs, at the same time illustrating various default prior distributions, such as Zellner's g-prior, the JZS approach and the hyper-g approach (for a similar approach see Bayarri & García-Donato, 2007).

Of course, other Bayesian model specifications for ANOVA are possible; ours has the advantage that it follows directly from the regression approach that has been studied in detail. A further didactical advantage is that many students are already familiar with linear regression and the extension to ANOVA is conceptually straightforward. In addition, software programs implemented in R make it easy for students and teachers to apply Bayesian regression and ANOVA to inference problems of practical interest; in addition, this software allows users to compare the substantive Bayesian conclusions to those drawn from the classical p-value approach. In general, the software implementation of the theoretical framework provides students with the opportunity of considerable hands-on experience with Bayesian hypothesis testing, something that is likely to increase not only their understanding, but also their motivation to learn.

We feel it is important for students to realize that there is likely no single correct prior distribution; in fact, it can be informative to use different priors in a sensitivity analysis. If different plausible prior distributions lead to different substantive conclusions it is best to acknowledge that the data are ambiguous.

Although not the focus of this article, post-hoc comparisons can easily be accommodated within the present framework. For instance, one might be interested in testing which group mean is different from the reference category mean. Then it is straightforward to calculate a Bayes factor to compare those means, using a procedure resembling a Bayesian t test (Gönen et al., 2005). Another possibility is to apply model averaging and calculate an inclusion probability for each predictor over all possible models (Clyde, 1999; Hoeting, Madigan, Raftery, & Volinsky, 1999).

Note that although the Bayes factor already has a dimension penalty built in—sometimes called the Bayesian Ockham's razor (J. O. Berger & Jefferys, 1992)—this is not a penalty against multiple comparisons. In order to correct for multiple comparisons, the prior on the model itself must be chosen appropriately (see Scott & Berger, 2010; Stephens & Balding, 2009 and references therein).

In sum, we have outlined a default Bayesian hypothesis test for ANOVA designs by a direct and simple extension of the framework for variable selection in regression models. In the course of doing so we have discussed three of the most popular default priors. We hope that empirical researchers and students can better appreciate and understand Bayesian hypothesis testing when they see how it can be applied to practical research problems for which ANOVA is often the method of choice.

Part II

Bayesian Model Selection: Applied

6 Statistical Evidence in Experimental Psychology: An Empirical Comparison Using 855 t Tests

Abstract

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. 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.

An excerpt of this chapter has been published as:
Wetzels, R., Matzke, D., Lee, M.D., Rouder, J.N., Iverson, G.J., & Wagenmakers, E.-J. (2011). Statistical Evidence in Experimental Psychology: An Empirical Comparison Using 855 t Tests. *Perspectives on Psychological Science, 6*, 291–298.

6.1 Introduction

Experimental psychologists use statistical procedures to convince themselves and their peers that the effect of interest is real, reliable, replicable, and hence worthy of academic attention. A representative example comes from Mussweiler (2006) who studied whether particular actions can activate a corresponding stereotype. To test this hypothesis empirically, Mussweiler unobtrusively induced half the participants, the experimental group, to move in a portly manner that is stereotypic for the overweight. The other half, the control group, made no such movements. Next, all participants were given an ambiguous description of a target person and then used a 9-point scale (1 = not al all, 9 = very) to rate this person on dimensions that correspond to the overweight stereotype (e.g., "unhealthy", "sluggish", "insecure"). To assess whether performing the stereotypic motion affected the rating of the ambiguous target person, Mussweiler computed a t statistic ($t(18) = 2.1$), and found that this value corresponded to a low p value ($p < .05$).[1] Following conventional protocol, Mussweiler concluded that the low p value should be taken to provide "initial support for the hypothesis that engaging in stereotypic movements activates the corresponding stereotype" (Mussweiler, 2006, p. 18).

The use of t tests and corresponding p values in this way constitutes a common and widely accepted practice in the psychological literature. It is, however, not the only possible or reasonable approach to measuring evidence and making statistical and scientific inferences. Indeed, the use of t tests and p values has been widely criticized (e.g., Cohen, 1994; Howard, Maxwell, & Fleming, 2000; Cumming, 2008; Dixon, 2003; Lee & Wagenmakers, 2005; Loftus, 1996; Nickerson, 2000; Wagenmakers, 2007). There are at least two different criticisms, coming from different perspectives, and resulting in different remedies. On the one hand, many have argued that null hypothesis tests should be supplemented with other statistical measures, such as confidence intervals and effect sizes. Within psychology, this approach to remediation has sometimes been institutionalized, being required by journal editors or recommended by the APA (e.g., American Psychological Association, 2010; Cohen, 1988; Erdfelder, 2010; Wilkinson & the Task Force on Statistical Inference, 1999).

A second, more fundamental criticism that comes from Bayesian statistics is that there are basic conceptual and practical problems with p values. Although Bayesian criticism of psychological statistical practice dates back at least to Edwards et al. (1963), it has become especially prominent and increasingly influential in the last decade (e.g., Dienes, 2008; Gallistel, 2009; J. Kruschke, In Press; J. K. Kruschke, 2010a; Lee, 2008; I. J. Myung, Forster, & Browne, 2000; Rouder et al., 2009). One standard Bayesian measure for quantifying the amount of evidence from the data in support of an experimental effect is the *Bayes factor* (Gönen et al., 2005; Rouder et al., 2009; Wetzels et al., 2009). The measure takes the form of an odds ratio: it is the probability of the data under one hypothesis relative to that under another (Dienes, 2011; Kass & Raftery, 1995; Lee & Wagenmakers, 2005).

With this background, it seems that psychological statistical practice currently stands at a three-way fork in the road. Staying on the current path means continuing to rely on p values. A modest change is to place greater focus on the additional inferential information provided by effect sizes and confidence intervals. A radical change is struck by moving to Bayesian approaches such as Bayes factors. The path that psychological science chooses seems likely to matter. It is not just that there are philosophical differences between

[1] The findings suggest that Mussweiler conducted a one-sided t test. In the remainder of this article we conduct two-sided t tests.

the three choices. It is also clear that the three measures of evidence can be mutually inconsistent (e.g., J. O. Berger & Sellke, 1987; Rouder et al., 2009; Wagenmakers, 2007; Wagenmakers & Grünwald, 2006; Wagenmakers et al., 2010).

In this paper, we assess the practical consequences of choosing among inference by p values, by effect sizes, and by Bayes factors. By practical consequences, we mean the extent to which conclusions of extant studies change according to the inference measure that is used. To assess these practical consequences, we re-analyzed 855 t tests reported in articles from the 2007 issues of *Psychonomic Bulletin & Review* (PBR) and *Journal of Experimental Psychology: Learning, Memory and Cognition* (JEP:LMC). For each t test, we compute the p value, the effect size, and the Bayes factor and study the extent to which they provide information that is redundant, complementary, or inconsistent. On the basis of these analyses, we suggest the best direction for measuring statistical evidence from psychological experiments.

6.2 Three Measures of Evidence

In this section, we describe how to calculate and interpret the p value, the effect size, and the Bayes factor. For concreteness, we use Mussweiler's study on the effect of action on stereotypes. The mean score of the control group, M_c, was 5.8 on a weight-stereotype scale ($s_c = 0.69, n_c = 10$), and the mean score of the experimental group, M_e, was 6.4 ($s_e = 0.66, n_e = 10$).

The p Value

The interpretation of p values is not straightforward, and their use in hypothesis testing is heavily debated (Cohen, 1994; Cortina & Dunlap, 1997; Cumming, 2008; Dixon, 2003; Frick, 1996; Gigerenzer, 1993, 1998; Hagen, 1997; Killeen, 2005, 2006; J. Kruschke, In Press; J. K. Kruschke, 2010a; Lee & Wagenmakers, 2005; Loftus, 1996; Nickerson, 2000; Schmidt, 1996; Wagenmakers & Grünwald, 2006; Wainer, 1999). The p value is 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 such as fixed sample size. Fisher (1935) interpreted these p values as evidence against the null hypothesis. The smaller the p value, the more evidence there was against the null hypothesis. Fisher viewed these values as self-explanatory measures of evidence that did not need further guidance. In practice, however, most researchers (and reviewers) adopt a .05 cutoff: p values less than .05 constitute evidence for an effect, and those greater than .05 do not. More fine-grained categories are possible, and Wasserman (2004, p. 157) proposes the gradations in Table 6.1. Note that Table 6.1 lists various categories of evidence *against* the null hypothesis. A basic limitation of null hypothesis significance testing is that it does not allow a researcher to gather evidence *in favor of* the null (Dennis, Lee, & Kinnell, 2008; Gallistel, 2009; Rouder et al., 2009; Wetzels et al., 2009).

For the data from Mussweiler, we compute a p value based on the t test. The t test is designed to test if a difference between two means is significant. First, we calculate the t statistic:

$$t = \frac{M_e - M_c}{\sqrt{s^2_{pooled}\left(\frac{1}{n_e} + \frac{1}{n_c}\right)}} = \frac{6.42 - 5.79}{\sqrt{0.46\left(\frac{1}{10} + \frac{1}{10}\right)}} = 2.09,$$

p Value			Interpretation
	$<$	0.001	Decisive Evidence Against H_0
0.001	–	0.01	Substantive Evidence Against H_0
0.01	–	0.05	Positive Evidence Against H_0
	$>$	0.05	No Evidence Against H_0

Table 6.1: Evidence categories for p values, adapted from Wasserman (2004, p. 157).

where M_c and M_e are the means of both groups, n_c and n_e are the sample sizes, and s^2_{pooled} estimates the common population variance:

$$s^2_{pooled} = \frac{(n_e - 1)s_e^2 + (n_c - 1)s_c^2}{n_e + n_c - 2}.$$

Next, the t statistic with $n_e + n_c - 2 = 18$ degrees of freedom results in a p value slightly larger than 0.05 (≈ 0.051). For our concrete example, Table 6.1 leads to the conclusion that the p value is on the cusp between "no evidence against H_0" and "positive evidence against H_0".

The Effect Size

Effect sizes quantify the magnitude of an effect and serves as a measure of how much the results deviate from the null hypothesis (Cohen, 1988; Thompson, 2002; Richard et al., 2003; Rosenthal, 1990; Rosenthal & Rubin, 1982). For the data from Mussweiler the effect size d is calculated as follows:

$$d = \frac{M_e - M_c}{s_{pooled}} = \frac{6.42 - 5.79}{0.68} = 0.93.$$

Note that in contrast to the p value, the effect size is independent of sample size; increasing the sample size does not increase effect size but instead allows it to be estimated more accurately.

Effect sizes are often interpreted in terms of the categories introduced by Cohen (1988), as listed in Table 6.2, ranging from "small" to "very large". For our concrete example, $d = 0.93$, and we conclude that this effect is large to very large. Interestingly, the p value was on the cusp between the categories "no evidence against H_0" and "positive evidence against H_0" whereas the effect size indicates the effect to be strong.

Effect Size			Interpretation
	$<$	0.2	Small Effect Size
0.2	–	0.5	Small to Medium Effect Size
0.5	–	0.8	Medium to Large Effect Size
	$>$	0.8	Large to Very Large Effect Size

Table 6.2: Evidence categories for effect sizes as proposed by Cohen (1988).

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).

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). The Bayes factor is the probability of the data under one hypothesis relative to the other. When a hypothesis is a simple point, such as the null, then the probability of the data under this hypothesis is simply the likelihood evaluated at that point. When a hypothesis consists of a range of points, such as all positive effect sizes, then the probability of the data under this hypothesis is the weighted average of the likelihood across that range. This averaging automatically controls for the complexity of different models, as has been emphasized in Bayesian literature in psychology (e.g., Pitt, Myung, & Zhang, 2002; Rouder et al., 2009).

We take as the null that a parameter α is restricted to 0 (i.e., $H_0 : \alpha = 0$), and take as the alternative that α is not zero (i.e., $H_A : \alpha \neq 0$). In this case, the Bayes factor given data D is simply the ratio

$$BF_{A0} = \frac{p\left(D \mid H_A\right)}{p\left(D \mid H_0\right)} = \frac{\int p\left(D \mid H_A, \alpha\right) p\left(\alpha \mid H_A\right) \, \mathrm{d}\alpha}{p\left(D \mid H_0\right)},$$

where the integral in the denominator takes the average evidence over all values of α, weighted by the prior probability of those values $p\left(\alpha \mid H_A\right)$ under the alternative hypothesis.

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; Good, 1985). Before seeing the data D, the two hypotheses H_0 and H_A are assigned prior probabilities $p(H_0)$ and $p(H_A)$. The ratio of the two prior probabilities defines the *prior odds*. When the data D are observed, the prior odds are updated to *posterior odds*, which is defined as the ratio of the posterior probabilities $p(H_0 \mid D)$ and $p(H_A \mid D)$:

$$\frac{p(H_A \mid D)}{p(H_0 \mid D)} = \frac{p(D \mid H_A)}{p(D \mid H_0)} \times \frac{p(H_A)}{p(H_0)}. \tag{6.1}$$

Equation 6.1 shows that the change from prior odds to posterior odds is quantified by $p(D \mid H_A)/p(D \mid H_0)$, the Bayes factor BF_{A0}.

Under either conceptualization, the Bayes factor has an appealing and direct interpretation as an odds ratio. For example, $BF_{A0} = 2$ implies that the data are twice as likely to have occurred under H_A than under H_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 6.3, facilitates scientific communication but should only be considered an approximate descriptive articulation of different standards of evidence (Kass & Raftery, 1995).

In general, calculating Bayes factors is more difficult than calculating p values and effect sizes. However, psychologists can now turn to easy-to-use webpages to calculate the Bayes factor for many common experimental situations or use software such as WinBUGS

Bayes factor			Interpretation
	$>$	100	Decisive evidence for H_A
30	–	100	Very Strong evidence for H_A
10	–	30	Strong evidence for H_A
3	–	10	Substantial evidence for H_A
1	–	3	Anecdotal evidence for H_A
	1		No evidence
1/3	–	1	Anecdotal evidence for H_0
1/10	–	1/3	Substantial evidence for H_0
1/30	–	1/10	Strong evidence for H_0
1/100	–	1/30	Very Strong evidence for H_0
	$<$	1/100	Decisive evidence for H_0

Table 6.3: Evidence categories for the Bayes factor BF_{A0} (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.

(D. J. Lunn et al., 2000; Wetzels et al., 2009; Wetzels, Lee, & Wagenmakers, in press).[2] In this paper, we use the Bayes factor calculation described in Rouder et al. (2009). Rouder et al.'s development is suitable for one-sample and two-sample designs, and the only necessary input is the t value and sample size.

The Bayes factor that we report in this article is the result of a *default* Bayesian t test (for details see Rouder et al., 2009). The test is default because it applies regardless of the phenomenon under study: for every experiment one uses the same prior on effect size for the alternative hypothesis, the Cauchy(0,1) distribution. This prior has statistical advantages that make it an appropriate default choice (for example, it has excellent theoretical properties in the limit, when $N \to \infty$ and $t \to \infty$; for details see Liang et al., 2008).

The default test is easy to use and avoids informed specification of prior distributions that other researchers may contest. On the other hand, one may argue that the informed specification of priors is the appropriate way to take problem-specific prior knowledge into account. Bayesian statisticians are divided over the relative merits of default versus informed specifications of prior distributions (Press et al., 2003). In our opinion, the default test provides an excellent starting point of analysis, one that may later be supplemented with a detailed problem-specific analysis (see Dienes, 2011, 2008; J. K. Kruschke, 2011, 2010a, 2010b for additional discussion of informed priors).

In our concrete example, the resulting Bayes factor for $t = 2.09$ and a sample size of 20 observations is $BF_{A0} = 1.56$. Accordingly, the data are 1.56 times more likely to have occurred under the alternative hypothesis than under the null hypothesis. This Bayes factor falls into the category "anecdotal". In other words, this Bayes factor indicates that although the alternative hypothesis is slightly favored, we do not have sufficiently strong evidence from the data to reject or accept either hypothesis.

[2] A webpage for computing a Bayes factor online is `http://pcl.missouri.edu/bayesfactor` and a webpage to download a tutorial and a flexible R/WinBUGS function to calculate the Bayes factor can be found at `www.ruudwetzels.com`.

6.3 Comparing p Values, Effect Sizes and Bayes Factors

For our concrete example, the three measures of evidence are not in agreement. The p value was on the cusp between the categories "no evidence against H_0" and "positive evidence against H_0", the effect size indicates a large to very large effect size, and the Bayes factor indicates that the data support the null hypothesis almost as much as they support the alternative hypothesis. If this example is not an isolated one, and the measures differ in many psychological applications, then it is important to understand the nature of those differences.

To address this question, we studied all of the empirical results evaluated by a t test in the Year 2007 volumes of *Psychonomic Bulletin & Review* (PBR) and *Journal of Experimental Psychology: Learning, Memory and Cognition* (JEP:LMC). This sample was comprised of 855 t tests from 252 articles. These articles covered 2394 journal pages, and addressed many topics that are important in modern experimental psychology. Our sample suggests, on average, that an article published in PBR and JEP:LMC contains about 3.4 t tests, which amounts to one t test for every 2.8 pages. For simplicity we did not include t tests that result from multiple comparisons in ANOVA designs (for a Bayesian perspective on multiple comparisons see Scott and Berger (2006)). Even though our t tests are sampled from the field of experimental/cognitive psychology, we expect our findings to generalize to many other subfields of psychology, as long as the studies in these subfields use the same level of statistical significance, approximately the same number of participants, and approximately the same number of trials per participant (Howard et al., 2000).

In the next sections we describe the empirical relation between the three measures of evidence, starting with the relation between effect sizes and p values.

Comparing Effect Sizes and p Values

The relationship between the obtained p values and effect sizes is shown as a scatter plot in Figure 6.1. Each point corresponds to one of the 855 comparisons. Different panels are introduced to distinguish the different evidence categories, as given in Tables 6.1 and Table 6.2.

Figure 6.1 suggests that p values and effect sizes capture roughly the same information in the data. Large effect sizes tend to correspond to low p values, and small effect sizes tend to correspond to large p values. The two measures, however, are far from identical. For instance, a p value of 0.01 can correspond to effect sizes ranging from about 0.2 to 1, and an effect size near 0.5 can correspond to p values ranging from about 0.001 to 0.05. The triangular points in the top-right panel of Figure 6.1 highlight gross inconsistencies. These 8 studies have a large effect size, above 0.8, but their p values do not indicate evidence against the null hypothesis. A closer examination revealed that these studies had p values very close to 0.05, and were comprised of small sample sizes.

Comparing Effect Sizes and Bayes Factors

The relationship between the obtained Bayes factors and effect sizes is shown in Figure 6.2. Much as with the comparison of p values with effect sizes, it seems clear that the default Bayes factor and effect size generally agree, though not exactly. No striking inconsistencies are apparant: No study with an effect size greater than 0.8 coincides with a Bayes factor below 1/3, nor does a study with very low effect size below 0.2 coincide with a Bayes factor above 3. The two measures, however, are not identical. They differ in the assessment of

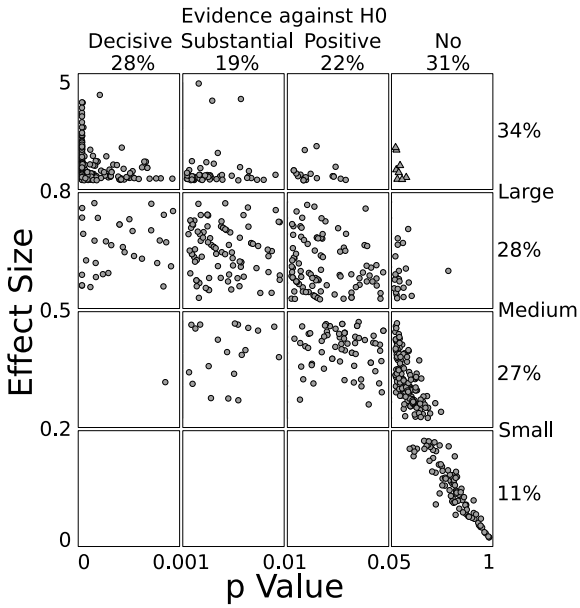


Figure 6.1: The relationship between effect size and p values. Points denote comparisons (855 in total. Points denoted by circle indicate relative consistency between the effect size and p value, while those denoted by triangles indicate gross inconsistency. The scale of the axes is based on the decision categories, as given in Table 6.1 and Table 6.2.

strength of evidence. Effect sizes above 0.8 range all the way from anecdotal to decisive evidence in terms of the Bayes factor. Also note that small to medium effect sizes (i.e., those between 0.2 and 0.5) can correspond to Bayes factor evidence in favor of either the alternative or the null hypothesis.

This last observation highlights that Bayes factors may quantify support for the null hypothesis. Figure 6.2 shows that about one-third of all studies produced evidence in favor of the null hypothesis. In about half of these studies favoring the null, the evidence is substantial. Because of the file-drawer problem (i.e., only significant effects tend to get published) this is an underestimate of the true amount of null findings and their Bayes factor support.

Comparing p Values and Bayes Factors

The relationship between the obtained Bayes factors and p values is shown in Figure 6.3, again using interpretative panels. It is clear that default Bayes factors and p values largely covary with each other. Low Bayes factors correspond to high p values and high Bayes factors correspond to low p values, a relationship that is much more exact than for

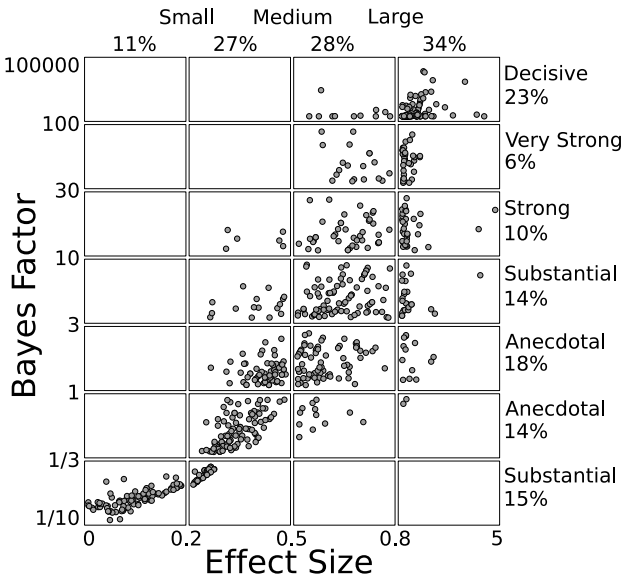


Figure 6.2: The relationship between Bayes factor and effect size. Points denote comparisons (855 in total). The scale of the axes is based on the decision categories, as given in Table 6.2 and Table 6.3.

our previous two comparisons. The main difference between default Bayes factors and p values is one of calibration; p values accord more evidence against the null than do Bayes factors. Consider the p values between .01 and .05, values that correspond to "positive evidence" and that usually pass the bar for publishing in academia. According to the default Bayes factor, 70% of these experimental effects convey evidence in favor of the alternative hypothesis that is only "anecdotal". This difference in the assessment of the strength of evidence is dramatic and consequential.

6.4 Conclusions

We compared p values, effect sizes and default Bayes factors as measures of statistical evidence in empirical psychological research. Our comparison was based on a total of 855 different t statistics from all published articles in two major empirical journals in 2007. In virtually all studies, the three different measures of evidence are broadly consistent: small p values correspond to large effect sizes and large Bayes factors in favor of the alternative hypothesis. Despite the fact that the measures of evidence reach the same conclusion about what hypothesis is best supported by the data, however, the measures differ with respect to the strength of that support. In particular, we noted that p values between .01 and .05 often correspond to what, in Bayesian terms, is only anecdotal evidence favor of the alternative hypothesis. The practical ramifications of this are considerable.

Practical Ramifications

Our results showed that when the p value falls in the interval from .01 to .05, there is a 70% chance that the default Bayes factor indicates the evidence for the alternative

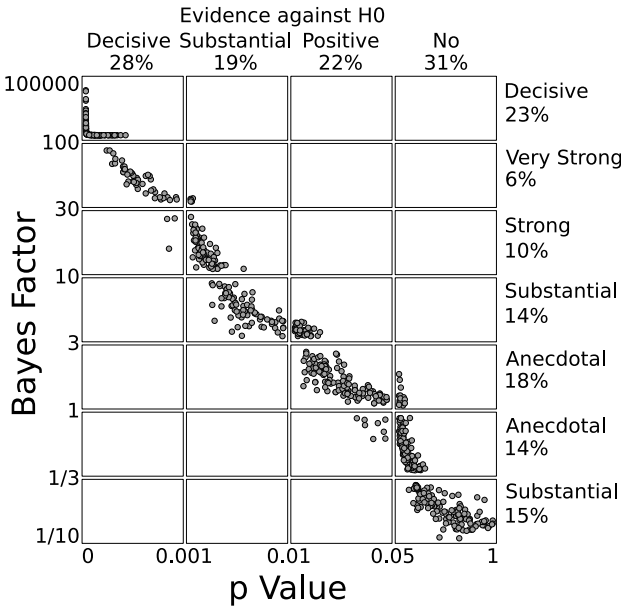


Figure 6.3: The relationship between Bayes factor and p value. Points denote comparisons (855 in total). The scale of the axes is based on the decision categories, as given in Table 6.1 and Table 6.3.

hypothesis to be only anecdotal or "worth no more than a bare mention"; this means that the data are no more than three times more likely under the alternative hypothesis than they are under the null hypothesis. Hence, for the studies under consideration here, it seems that a p value criterion more conservative than .05 is appropriate. Alternatively, researchers could avoid computing a p value altogether and instead compute the Bayes factor. Both methods help prevent researchers from overestimating the strength of their findings, and help the field from incorporating ambiguous findings as if they were real and reliable (Ioannidis, 2005).

As a practical illustration, consider a series of recent experiments on precognition (Bem, 2011).[3] In nine experiments with over 1000 participants, Dr. Bem intended to show that precognition exists, that is, that people can foresee the future. And indeed, eight out of nine experiments yielded a significant result. However, most p values fell in the ambiguous range of .01 to .05, and, across all nine experiments, a Bayes factor analysis indicates about as much evidence for the alternative hypothesis as against it (J. K. Kruschke, 2011; Wagenmakers, Wetzels, Borsboom, & van der Maas, in press). We believe that this situation typifies part of what could be improved in psychological

[3]A preprint of Bem's article is available at `http://dbem.ws/FeelingFuture.pdf`.

research today. It is simply too easy to obtain a p value below .05 and subsequently publish the result.

When researchers publish ambiguous results as if they were real and reliable, this damages the field as a whole – time, effort, and money will be invested to replicate the phenomenon, and, when replication fails, the burden of proof is almost always on the part of the researcher who, after all, failed to replicate a phenomenon that was demonstrated to be present (with a p value in between .01 and .05).

Thus, our empirical comparison shows that the academic criterion of .05 is too liberal. Note this problem would not be solved by opting for a stricter significance level, such as .01. It is well known that the p value decreases as the sample size n increases. Hence, if psychologists switch to a significance level of .01 but inevitably increase their sample sizes to compensate for the stricter statistical threshold, then the phenomenon of anecdotal evidence will start to plague p values even when these p values are lower than .01. Therefore, we make a case for Bayesian statistics in the next section.

A Case for Bayesian Statistics

We have compared the conclusions from the different measures of evidence. It is easy to make a case for Bayesian statistical inference in general, based on arguments already well documented in statistics and psychology (e.g., Dienes, 2008; Jaynes, 2003; J. Kruschke, In Press; J. K. Kruschke, 2010a; Lee & Wagenmakers, 2005; D. V. Lindley, 1972; Wagenmakers, 2007). We briefly mention three arguments here.

Firstly, unlike null hypothesis testing, Bayesian inference does not violate basic principles of rational statistical decision-making such as the stopping rule principle or the likelihood principle (J. O. Berger & Wolpert, 1988; J. O. Berger & Delampady, 1987). This means that the results of Bayesian inference do not depend on the intention with which the data were collected. As stated by Edwards et al. (1963, p. 193), "the rules governing when data collection stops are irrelevant to data interpretation. It is entirely appropriate to collect data until a point has been proven or disproven, or until the data collector runs out of time, money, or patience".

Secondly, Bayesian inference takes model complexity into account in a rational way. Specifically, the Bayes factor has the attraction of not assigning a special status to the null hypothesis, and so makes it theoretically possible to measure evidence in favor of the null (e.g., Dennis et al., 2008; Gallistel, 2009; Kass & Raftery, 1995; Rouder et al., 2009).

Thirdly, we believe that Bayesian inference provides the kind of answers that researchers care about. In our experience, researchers are usually not that interested in the probability of encountering data at least as extreme as those that were observed, given that the null hypothesis is true and the sample was generated according to a specific intended procedure. Instead, most researchers want to know what they have learned from the data about the relative plausibility of the hypotheses under consideration. This is exactly what is quantified by the Bayes factor.

These advantages notwithstanding, the Bayes factor is not a measure of the mere size of an effect. Hence the measure of effect size confers additional information, particularly when small numbers of participants or trials are involved. So, especially for these sorts of studies, there is an argument for reporting both a Bayes factor and an effect size. We note that, from a Bayesian perspective, the effect size can naturally be conceived as a (summary statistic of) the posterior distribution of a parameter representing the effect, under an uninformative prior distribution. In this sense, a standard Bayesian combination of parameter estimation and model selection could encompass all of the useful measures of

evidence we observed (for an example of how Bayes factor estimation can be incorporated in a Bayesian estimation framework, see for example J. K. Kruschke, 2011).

Our final thought is that reasons for adopting a Bayesian approach now are amplified by the promise of using an extended Bayesian approach in the future. In particular, we think the hierarchical Bayesian approach, which is standard in statistics (e.g. Gelman & Hill, 2007), and is becoming more common in psychology (e.g. J. Kruschke, In Press; J. K. Kruschke, 2010b; Lee, 2011; Rouder & Lu, 2005) could fundamentally change how psychologists identify effects. Hierarchical Bayesian analysis can be a valuable tool both for meta-analyses and for the analysis of a single study. In the meta-analytical context, multiple studies can be integrated, so that what is inferred about the existence of effects and their magnitude is informed, in a coherent and quantitative way, by a domain of experiments. In the context of a single experiment, a hierarchical analysis can be used to take variability across participants or items into account.

In sum, our empirical comparison of 855 t tests shows that three often-used measures of evidence – p values, effect sizes, and 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 those data sets with p values in between .01 and .05, about 70% are associated with a Bayes factor that indicates the evidence to be only anecdotal or "worth no more than a bare mention" (Jeffreys, 1961). This analysis suggests that many results that have been published in the literature are not established as strongly as one would like.

7 Why Psychologists Must Change the Way They Analyze Their Data: The Case of Psi

Abstract

Does psi exist? In a recent 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.

An excerpt of this chapter has been published as:
Wagenmakers, E.–J., Wetzels. R., Borsboom. D., & van der Maas. H. L. J. (2011). Why Psychologists Must Change the Way They Analyze Their Data: The Case of Psi. *Journal of Personality and Social Psychology, 100*, 426–432.

7.1 Introduction

In a recent article for *Journal of Personality and Social Psychology*, Bem (2011) presented nine experiments that test for the presence of psi.[1] Specifically, the experiments were designed to assess the hypothesis that future events affect people's thinking and people's behavior in the past (henceforth precognition). As indicated by Bem, precognition—if it exists—is an anomalous phenomenon, because it conflicts with what we know to be true about the word (e.g., weather forecasting agencies do not employ clairvoyants, casino's make profit, etc.). In addition, psi has no clear grounding in known biological or physical mechanisms.[2]

Despite the lack of a plausible mechanistic account of precognition, Bem was able to reject the null hypothesis of no precognition in eight out of nine experiments. For instance, in Bem's first experiment 100 participants had to guess the future position of pictures on a computer screen, left or right. And indeed, for erotic pictures, the 53.1% mean hit rate was significantly higher than chance ($t(99) = 2.51$, $p = .01$).

Bem takes these findings to support the hypothesis that people "use psi information implicitly and nonconsciously to enhance their performance in a wide variety of everyday tasks". In further support of psi, Utts (1991, p. 363) concluded in a *Statistical Science* review article that "(...) the overall evidence indicates that there is an anomalous effect in need of an explanation" (but see Diaconis, 1978; Hyman, 2007). Do these results mean that psi can now be considered real, replicable, and reliable?

We think that the answer to this question is negative, and that the take home message of Bem's research is in fact of a completely different nature. One of the discussants of the Utts review paper made the insightful remark that "Parapsychology is worth serious study. (...) if it is wrong [i.e., psi does not exist], it offers a truly alarming massive case study of how statistics can mislead and be misused." (Diaconis, 1991, p. 386). And this, we suggest, is precisely what Bem's research really shows. Instead of revising our beliefs regarding psi, Bem's research should instead cause us to revise our beliefs on methodology: the field of psychology currently uses methodological and statistical strategies that are too weak, too malleable, and offer far too many opportunities for researchers to befuddle themselves and their peers.

The most important flaws in the Bem experiments, discussed below in detail, are the following: (1) confusion between exploratory and confirmatory studies; (2) insufficient attention to the fact that the probability of the data given the hypothesis does not equal the probability of the hypothesis given the data (i.e., the fallacy of the transposed conditional); (3) application of a test that overstates the evidence against the null hypothesis, an unfortunate tendency that is exacerbated as the number of participants grows large. Indeed, when we apply a Bayesian t test (Gönen et al., 2005; Rouder et al., 2009) to quantify the evidence that Bem presents in favor of psi, the evidence is sometimes slightly in favor of the null hypothesis, and sometimes slightly in favor of the alternative hypothesis. In almost all cases, the evidence falls in the category "anecdotal", also known as "worth no more than a bare mention" (Jeffreys, 1961).

[1] The preprint that this article is based on was downloaded September 25th, 2010, from `http://dbem.ws/FeelingFuture.pdf`.

[2] Some argue that modern theories of physics are consistent with precognition. We cannot independently verify this claim, but note that work on precognition is seldom published in reputable physics journals (in fact, we failed to find a single such publication). But even if the claim were correct, the fact that an assertion is consistent with modern physics does not make it true. The assertion that the CIA bombed the twin towers is consistent with modern physics, but this fact alone does not make the assertion true. What is needed in the case of precognition is a plausible account of the process that leads future events to have perceptual effects in the past.

We realize that the above flaws are not unique to the experiments reported by Bem. Indeed, many studies in experimental psychology suffer from the same mistakes. However, this state of affairs does not exonerate the Bem experiments. Instead, these experiments highlight the relative ease with which an inventive researcher can produce significant results even when the null hypothesis is true. This evidently poses a significant problem for the field, and impedes progress on phenomena that are replicable and important.

7.2 Problem 1: Exploration Instead of Confirmation

In his well-known book chapters on writing an empirical journal article, Bem (2000, 2003) rightly calls attention to the fact that psychologists do not often engage in purely confirmatory studies. That is,

> "The conventional view of the research process is that we first derive a set of hypotheses from a theory, design and conduct a study to test these hypotheses, analyze the data to see if they were confirmed or disconfirmed, and then chronicle this sequence of events in the journal article. (...) But this is not how our enterprise actually proceeds. Psychology is more exciting than that (...)" (Bem, 2000, p. 4).

How is it then that psychologists analyze their data? Bem notes that senior psychologists often leave the data collection to their students, and makes the following recommendation:

> "To compensate for this remoteness from our participants, let us at least become intimately familiar with the record of their behavior: the data. Examine them from every angle. Analyze the sexes separately. Make up new composite indexes. If a datum suggests a new hypothesis, try to find further evidence for it elsewhere in the data. If you see dim traces of interesting patterns, try to reorganize the data to bring them into bolder relief. If there are participants you don't like, or trials, observers, or interviewers who gave you anomalous results, place them aside temporarily and see if any coherent patterns emerge. Go on a fishing expedition for something–anything–interesting." (Bem, 2000, pp. 4-5)

We agree with Bem in the sense that empirical research can benefit greatly from a careful exploration of the data; dry adherence to confirmatory studies stymies creativity and the development of new ideas. As such, there is nothing wrong with fishing expeditions. But it is vital to indicate clearly and unambiguously which results are obtained by fishing expeditions and which results are obtained by conventional confirmatory procedures. In particular, when results from fishing expeditions are analyzed and presented as if they had been obtained in a confirmatory fashion, the researcher is hiding the fact that the same data were used *twice*: first to discover a new hypothesis, and then to test that hypothesis. If the researcher fails to state that the data have been so used, this practice is at odds with the basic ideas that underlie scientific methodology (see Kerr, 1998, for a detailed discussion).

Instead of presenting exploratory findings as confirmatory, one should ideally use a two-step procedure: first, in the absence of strong theory, one can explore the data until one discovers an interesting new hypothesis. But this phase of exploration and discovery needs to be followed by a second phase, one in which the new hypothesis is tested against new data in a confirmatory fashion. This is particularly important if one wants to convince a skeptical audience of a controversial claim: after all, confirmatory

studies are much more compelling than exploratory studies. Hence, explorative elements in the research program should be explicitly mentioned, and statistical results should be adjusted accordingly. In practice, this means that statistical tests should be corrected to be more conservative.

The Bem experiments were at least partly exploratory. For instance, Bem's Experiment 1 tested not just erotic pictures, but also neutral pictures, negative pictures, positive pictures, and pictures that were romantic but non-erotic. Only the erotic pictures showed any evidence for precognition. But now suppose that the data would have turned out differently and instead of the erotic pictures, the positive pictures would have been the only ones to result in performance higher than chance. Or suppose the negative pictures would have resulted in performance lower than chance. It is possible that a new and different story would then have been constructed around these other results (Bem, 2003; Kerr, 1998). This means that Bem's Experiment 1 was to some extent a fishing expedition, an expedition that should have been explicitly reported and should have resulted in a correction of the reported p value.

Another example of exploration comes from Bem's Experiment 3, in which response time (RT) data were transformed using either an inverse transformation (i.e., 1/RT) or a logarithmic transformation. These transformations are probably not necessary, because the statistical analysis were conducted on the level of participant mean RT; one then wonders what the results were for the untransformed RTs—results that were not reported.

Furthermore, in Bem's Experiment 5 the analysis shows that "Women achieved a significant hit rate on the negative pictures, 53.6%, $t(62) = 2.25$, $p = .014$, $d = .28$; but men did not, 52.4%, $t(36) = 0.89$, $p = .19$, $d = .15$." But why test for gender in the first place? There appears to be no good reason. Indeed, Bem himself states that "the psi literature does not reveal any systematic sex differences in psi ability".

Bem's Experiment 6 offers more evidence for exploration, as this experiment again tested for gender differences, but also for the number of exposures: "The hit rate on control trials was at chance for exposure frequencies of 4, 6, and 8. On sessions with 10 exposures, however, it fell to 46.8%, $t(39) = -2.12$, two-tailed p = .04." Again, conducting multiple tests requires a correction.

These explorative elements are clear from Bem's discussion of the empirical data. The problem runs deeper, however, because we simply do not know how many other factors were taken into consideration only to come up short. We can never know how many other hypotheses were in fact tested and discarded; some indication is given above and in Bem's section "The File Drawer". At any rate, the foregoing suggests that strict confirmatory experiments were not conducted. This means that the reported p values are incorrect and need to be adjusted upwards.

7.3 Problem 2: Fallacy of the Transposed Conditional

The interpretation of statistical significance tests is liable to a misconception known as the fallacy of the transposed conditional. In this fallacy, the probability of the data given a hypothesis (e.g., $p(D|H)$, such as the probability of someone being dead given that they were lynched, a probability that is close to 1) is confused with the probability of the hypothesis given the data (e.g., $P(H|D)$, such as the probability that someone was lynched given that they are dead, a probability that is close to zero).

This distinction provides the mathematical basis for Laplace's Principle that extraordinary claims require extraordinary evidence. This principle holds that even compelling data may not make a rational agent believe that psi exists (see also Price, 1955). Thus, the

prior probability attached to a given hypothesis affects the strength of evidence required to make a rational agent change his or her mind.

Suppose, for instance, that in the case of psi we have the following hypotheses:

$$H_0 = \text{Precognition does not exist;}$$
$$H_1 = \text{Precognition does exist.}$$

Our personal prior belief in precognition is very low; two reasons for this are outlined below. We accept that each of these reasons can be disputed by those who believe in psi, but this is not the point—we do not mean to disprove psi on logical grounds. Instead, our goal is to indicate why most researchers currently believe psi phenomena are unlikely to exist.[3]

As a first reason, consider that Bem (2011) acknowledges that there is no mechanistic theory of precognition (see Price, 1955 for a discussion). This means, for instance, that we have no clue about how precognition could arise in the brain—neither animals nor humans appear to have organs or neurons dedicated to precognition, and it is unclear what electrical or biochemical processes would make precognition possible. Note that precognition conveys a considerable evolutionary advantage (Bem, 2011), and one might therefore assume that natural selection would have lead to a world filled with powerful psychics (i.e., people or animals with precognition, clairvoyance, psychokineses, etc.). This is not the case, however (see also Kennedy, 2001). The believer in precognition may object that psychic abilities, unlike all other abilities, are not influenced by natural selection. But the onus is then squarely on the believer in psi to explain why this should be so.

Second, there is no real-life evidence that people can feel the future (e.g., nobody has ever collected the $1,000,000 available for anybody who can demonstrate paranormal performance under controlled conditions[4], etc.). To appreciate how unlikely the existence of psi really is, consider the facts that (a) casinos make profit, and (b) casinos feature the game of French roulette. French roulette features 37 numbers, 18 colored black, 18 colored red, and the special number 0. The situation we consider here is where gamblers bet on the color indicated by the roulette ball. Betting on the wrong color results in a loss of your stake, and betting on the right color will double your stake. Because of the special number 0, the house holds a small advantage over the gambler; the probability of the house winning is 19/37.

Consider now the possibility that the gambler could use psi to bet on the color that will shortly come up, that is, the color that will bring great wealth in the immediate future. In this context, even small effects of psi result in substantial payoffs. For instance, suppose a player with psi can anticipate the correct color in 53.1% of cases—the mean percentage correct across participants for the erotic pictures in Bem's Experiment 1. Assume that this psi-player starts with only 100 euros, and bets 10 euro every time. The gambling stops whenever the psi-player is out of money (in which case the casino wins) or the psi-player has accumulated one million euros. After accounting for the house advantage, what is the probability that the psi-player will win one million euros? This probability, easily calculated from random walk theory (e.g., Feller, 1970, 1971) equals 48.6%. This means that, in this case, the expected profit for a psychic's night out at the casino equals $485,900. If Bem's psychic plays the game all year round, never raises the stakes, and always quits at a profit of a million dollars, the expected return is $177,353,500.[5]

[3] This is evident from the fact that psi research is almost never published in the mainstream literature.

[4] See `http://www.skepdic.com/randi.html` for details.

[5] The break-even point for the house lies at a success probability of 0.514. However, even if the

Clearly, Bem's psychic could bankrupt all casinos on the planet before anybody realized what was going on. This analysis leaves us with two possibilities. The first possibility is that, for whatever reason, the psi effects are not operative in casinos, but they are operative in psychological experiments on erotic pictures. The second possibility is that the psi effects are either nonexistent, or else so small that they cannot overcome the house advantage. Note that in the latter case, all of Bem's experiments overestimate the effect.

Returning to Laplace's Principle, we feel that the above reasons motivate us to assign our prior belief in precognition a number very close to zero. For illustrative purposes, let us set $P(H_1) = 10^{-20}$, that is, .00000000000000000001. This means that $P(H_0) = 1 - P(H_1) = .99999999999999999999$. Our aim here is not to quantify precisely our personal prior belief in psi. Instead, our aim is to explain Laplace's Principle by using a concrete example and specific numbers. It is also important to note that the Bayesian t test outlined in the next section does not depend in any way on the prior probabilities $P(H_0)$ and $P(H_1)$.

Now assume we find a flawless, well-designed, 100% confirmatory experiment for which the observed data are unlikely under H_0 but likely under H_1, say by a factor of 19 (as indicated below, this is considered "strong evidence"). In order to update our prior belief, we apply Bayes' rule:

$$\begin{aligned} p(H_1|D) &= \frac{p(D|H_1)p(H_1)}{p(D|H_0)p(H_0) + p(D|H_1)p(H_1)} \\ &= \frac{.95 \times 10^{-20}}{.05(1 - 10^{-20}) + .95 \times 10^{-20}} \\ &= .00000000000000000019. \end{aligned}$$

True, our posterior belief in precognition is now higher than our prior belief. Nevertheless, we are still relatively certain that precognition does not exist. In order to overcome our skeptical prior opinion, the evidence needs to be much stronger. In other words, extraordinary claims require extraordinary evidence. This is neither irrational nor unfair; if the proponents of precognition succeed in establishing its presence, their reward is eternal fame, (and, if Bem were to take his participants to the casino, infinite wealth).

Thus, in order to convince scientific critics of an extravagant or controversial claim, one is required to pull out all the stops. Even when Bem's experiments had been confirmatory (which they were not, see above), and even if they would have conveyed strong statistical evidence for precognition (which they did not, see below), eight experiments are not enough to convince a skeptic that the known laws of nature have been bent. Or, more precisely, that these laws were bent only for erotic pictures, and only for participants who are extraverts.

7.4 Problem 3: p values Overstate the Evidence Against the Null

Consider a data set for which $p = .001$, indicating a low probability of encountering a test statistic that is at least as extreme as the one that was actually observed, given that the

success rate is smaller, say, 0.510, one can boost one's success probability by utilizing a team of psychics and using their majority vote. This is so because Condorcet's jury theorem ensures that, whenever the success probability for an individual voter lies above 0.5, the probability of a correct majority vote approaches 1 as the number of voters grows large. If the individual success probability is 0.510, for instance, using the majority vote of a team of 1000 psychics gives a probability of .73 for the majority vote being correct.

null hypothesis H_0 is true. Should we proceed to reject H_0? Well, this depends at least in part on how likely the data are under H_1. Suppose, for instance, that H_1 represents a very small effect—then it may be that the observed value of the test statistic is almost as unlikely under H_0 as under H_1. What is going on here?

The underlying problem is that evidence is a relative concept, and it is of limited interest to consider the probability of the data under just a single hypothesis. For instance, if you win the state lottery you might be accused of cheating; after all, the probability of winning the state lottery is rather small. This may be true, but this low probability in itself does not constitute evidence—the evidence is assessed only when this low probability is pitted against the much lower probability that you could somehow have obtained the winning number by acquiring advance knowledge on how to buy the winning ticket.

Therefore, in order to evaluate the strength of evidence that the data provide for or against precognition, we need to pit the null hypothesis against a specific alternative hypothesis, and not consider the null hypothesis in isolation. Several methods are available to achieve this goal. Classical statisticians can achieve this goal with the Neyman-Pearson procedure, statisticians who focus on likelihood can achieve this goal using likelihood ratios (Royall, 1997), and Bayesian statisticians can achieve this goal using a hypothesis test that computes a weighted likelihood ratio (e.g., Rouder et al., 2009; Wagenmakers et al., 2010; Wetzels et al., 2009). As an illustration, we focus here on the Bayesian hypothesis test.

In a Bayesian hypothesis test, the goal is to quantify the change in prior to posterior odds that is brought about by the data. For a choice between H_0 and H_1, we have

$$\frac{p(H_0|D)}{p(H_1|D)} = \frac{p(H_0)}{p(H_1)} \times \frac{p(D|H_0)}{p(D|H_1)}, \tag{7.1}$$

which is often verbalized as

$$\text{Posterior model odds} = \text{Prior model odds} \times \text{Bayes factor}. \tag{7.2}$$

Thus, the change from prior odds $p(H_0)/p(H_1)$ to posterior odds $p(H_0|D)/p(H_1|D)$ brought about by the data is given by the ratio of $p(D|H_0)/p(D|H_1)$, a quantity known as the *Bayes factor* (Jeffreys, 1961). The Bayes factor (or its logarithm) is often interpreted as the weight of evidence provided by the data (Good, 1985; for details see J. O. Berger & Pericchi, 1996, Bernardo & Smith, 1994, Chapter 6, Gill, 2002, Chapter 7, Kass & Raftery, 1995, and O'Hagan, 1995).

When the Bayes factor for H_0 over H_1 equals 2 (i.e., $BF_{01} = 2$) this indicates that the data are twice as likely to have occurred under H_0 then under H_1. Even though the Bayes factor has an unambiguous and continuous scale, it is sometimes useful to summarize the Bayes factor in terms of discrete categories of evidential strength. Jeffreys (1961, Appendix B) proposed the classification scheme shown in Table 7.1.

Several researchers have recommended Bayesian hypothesis tests (e.g., J. O. Berger & Delampady, 1987; J. O. Berger & Sellke, 1987; Edwards et al., 1963; see also Wagenmakers & Grünwald, 2006), particularly in the context of psi (e.g., Bayarri & Berger, 1991; Jaynes, 2003, Chap. 5; Jefferys, 1990).

To illustrate the extent to which Bem's conclusions depend on the statistical test that was used, we have reanalyzed the Bem experiments with a default Bayesian t test (Gönen et al., 2005; Rouder et al., 2009). This test computes the Bayes factor for H_0 versus H_1, and it is important to note that the prior model odds plays no role whatsoever in its calculation (see also Equations 7.1 and 7.2). One of the advantages of this Bayesian test is that it also allows researchers to quantify the evidence in favor of the null hypothesis,

Table 7.1: Classification scheme for the Bayes factor, as proposed by Jeffreys (1961). We replaced the labels "worth no more than a bare mention" with "anecdotal", and "decisive" with "extreme".

Bayes factor, BF_{01}			Interpretation
	$>$	100	Extreme evidence for H_0
30	–	100	Very Strong evidence for H_0
10	–	30	Strong evidence for H_0
3	–	10	Substantial evidence for H_0
1	–	3	Anecdotal evidence for H_0
	1		No evidence
1/3	–	1	Anecdotal evidence for H_1
1/10	–	1/3	Substantial evidence for H_1
1/30	–	1/10	Strong evidence for H_1
1/100	–	1/30	Very strong evidence for H_1
	$<$	1/100	Extreme evidence for H_1

something that is impossible with traditional p values. Another advantage of the Bayesian test that it is *consistent*: as the number of participants grows large, the probability of discovering the true hypothesis approaches 1.

The Bayesian t test

Ignoring for the moment our concerns about the exploratory nature of the Bem studies, and the prior odds in favor of the null hypothesis, we can wonder how convincing the statistical results from the Bem studies really are. After all, each of the Bem studies featured at least 100 participants, but nonetheless in several experiments Bem had to report one-sided (not two-sided) p values in order to claim significance at the .05 level. One might intuit that such data do not constitute compelling evidence for precognition.

In order to assess the strength of evidence for H_0 (i.e., no precognition) versus H_1 (i.e., precognition) we computed a default Bayesian t test for the critical tests reported in Bem (2011). This default test is based on general considerations that represent a lack of knowledge about the effect size under study (Gönen et al., 2005; Rouder et al., 2009; for a generalization to regression see Liang et al., 2008). More specific assumptions about the effect size of psi would result in a different test. We decided to first apply the default test because we did not feel qualified to make these more specific assumptions, especially not in an area as contentious as psi.

Using the Bayesian t test web applet provided by Dr. Rouder[6] it is straightforward to compute the Bayes factor for the Bem experiments: all that is needed is the t-value and the degrees of freedom (Rouder et al., 2009). Table 7.2 shows the results. Out of the 10 critical tests, only one yields "substantial" evidence for H_1, whereas three yield "substantial" evidence in favor of H_0. The results of the remaining six tests provide evidence that is only "anecdotal" or "worth no more than a bare mention" (Jeffreys, 1961).

In sum, a default Bayesian test confirms the intuition that, for large sample sizes, one-sided p values higher than .01 are not compelling (see also Wetzels et al., 2011[7]).

[6] See `http://pcl.missouri.edu/bayesfactor`.

[7] A preprint is available at `http://www.ruudwetzels.com/`.

Table 7.2: The results of 10 crucial tests for the experiments reported in Bem (in press), reanalyzed using the default Bayesian t test.

Exp	df	$\|t\|$	p	BF_{01}	Evidence category (in favor of $H_\cdot$)
1	99	2.51	0.01	0.61	Anecdotal (H_1)
2	149	2.39	0.009	0.95	Anecdotal (H_1)
3	96	2.55	0.006	0.55	Anecdotal (H_1)
4	98	2.03	0.023	1.71	Anecdotal (H_0)
5	99	2.23	0.014	1.14	Anecdotal (H_0)
6	149	1.80	0.037	3.14	Substantial (H_0)
6	149	1.74	0.041	3.49	Substantial (H_0)
7	199	1.31	0.096	7.61	Substantial (H_0)
8	99	1.92	0.029	2.11	Anecdotal (H_0)
9	49	2.96	0.002	0.17	Substantial (H_1)

Overall, the Bayesian t test indicates that the data of Bem do not support the hypothesis of precognition. This is despite the fact that multiple hypotheses were tested, something that warrants a correction (for a Bayesian correction see Scott & Berger, 2010; Stephens & Balding, 2009).

Note that, even though our analysis is Bayesian, we did not select priors to obtain a desired result: the Bayes factors that were calculated are independent of the prior model odds, and depend only on the prior distribution for effect size—for this distribution, we used the default option. We also examined other options, however, and found that our conclusions are robust: for a wide range of different, non-default prior distributions on effect size the evidence for precognition is either non-existent or negligible.[8]

At this point, one may wonder whether it is feasible to use the Bayesian t test and eventually obtain enough evidence against the null hypothesis to overcome the prior skepticism outlined in the previous section. Indeed, this is feasible: based on the mean and sample standard deviations reported in Bem's Experiment 1, it is straightforward to calculate that around 2000 participants are sufficient to generate an extremely high Bayes factor BF_{01} of about 10^{-24}; when this extreme evidence is combined with the skeptical prior, the end result is firm belief that psi is indeed possible. On the one hand, 2000 participants seems excessive; on the other hand, this is but a small subset of participants that have been tested in the field of parapsychology during the last decade. Of course, this presupposes that the experiment under consideration was 100% confirmatory, and that it has been conducted with the utmost care.

7.5 Guidelines for Confirmatory Research

As discussed earlier, exploratory research is useful but insufficiently compelling to change the mind of a skeptic. In order to provide hard evidence for or against an empirical proposition, one has to resort to strictly confirmatory studies. The degree to which the scientific community will accept semi-confirmatory studies as evidence depends partly on the plausibility of the claim under scrutiny: again, extraordinary claims require extraor-

[8]This robustness analysis is reported in an online appendix available on the first author's website, `http://www.ejwagenmakers.com/papers.html`.

dinary evidence. The basic characteristic of confirmatory studies is that all choices that could influence the result have been made before the data are observed. We suggest that confirmatory research in psychology observes the following guidelines:

1. Fishing expeditions should be prevented by selecting participants and items *before* the confirmatory study takes place. Of course, previous tests, experiments, and questionnaires may be used to identify those participants and items who show the largest effects—this method increases power in case the phenomenon of interest really does exist; however, no further selection or subset testing should take place once the confirmatory experiment has started.

2. Data should only be transformed if this has been decided beforehand. In confirmatory studies, one does not "torture the data until they confess". It also means that—upon failure—confirmatory experiments are not demoted to exploratory pilot experiments, and that—upon success—exploratory pilot experiments are not promoted to confirmatory experiments.

3. In simple examples, such as when the dependent variable is success rate or mean response time, an appropriate analysis should be decided upon *before* the data have been collected.

4. It is prudent to report more than a single statistical analysis. If the conclusions from p values conflict with those of, say, Bayes factors, then this should be clearly stated. Compelling results yield similar conclusions, irrespective of the statistical paradigm that is used to analyze the data.

In our opinion, the above guidelines are sufficient for most research topics. However, the researcher who wants to convince a skeptical community of academics that psi exists may want to go much further. In the context of psi, Price (1955, p. 365) argued that "(...) what is needed is something that can be demonstrated to the most hostile, pig-headed, and skeptical of critics." This is also consistent with Hume's maxim that "(...) no testimony is sufficient to establish a miracle, unless the testimony be of such a kind, that its falsehood would be more miraculous, than the fact, which it endeavours to establish (...)" (Hume, 1748, Chapter 10). What this means is that in order to overcome the skeptical bias against psi, the psi researcher might want to consider more drastic measures to ensure that the experiment was completely confirmatory:

5. The psi researcher may make stimulus materials, computer code, and raw data files publicly available online. The psi-researcher may also make the decisions made with respect to guidelines 1-4 publicly available online, and do so *before* the confirmatory experiment is carried out.

6. The psi researcher may engage in an adversarial collaboration, that is, a collaboration with a true skeptic, and preferably more than one (Price, 1955; Wiseman & Schlitz, 1997). This echoes the advice of Diaconis (1991, p. 386), who stated that the studies on psi reviewed by (Utts, 1991) were "crucially flawed (...) Since the field has so far failed to produce a replicable phenomena, it seems to me that any trial that asks us to take its findings seriously should include full participation by qualified skeptics."

The psi researcher who also follows the last two guidelines makes an effort that is slightly higher than usual; we believe this is a small price to pay for a large increase in credibility. It should after all be straightforward to document the intended analyses, and in most universities a qualified skeptic is sitting in the office next door.

7.6 Concluding Comment

In eight out of nine studies, Bem reported evidence in favor of precognition. As we have argued above, this evidence may well be illusory; in several experiments it is evident that exploration should have resulted in a correction of the statistical results. Also, we have provided an alternative, Bayesian reanalysis of Bem's experiments; this alternative analysis demonstrated that the statistical evidence was, if anything, slightly in favor of the null hypothesis. One can argue about the relative merits of classical t tests versus Bayesian t tests, but this is not our goal; instead, we want to point out that the two tests yield very different conclusions, something that casts doubt on the conclusiveness of the statistical findings.

In this article, we have assessed the evidential impact of Bem's experiments in isolation. It is certainly possible to combine the information across experiments, for instance by means of a meta-analysis (Storm, Tressoldi, & Di Risio, 2010; Utts, 1991). We are ambivalent about the merits of meta-analyses in the context of psi: one may obtain a significant result by combining the data from many experiments, but this may simply reflect the fact that some proportion of these experiments suffer from experimenter bias and excess exploration. When examining different answers to criticism against research on psi, Price (1955, p. 367) concluded "But the only answer that will impress me is an adequate experiment. Not 1000 experiments with 10 million trials and by 100 separate investigators giving total odds against change of 10^{1000} to 1—but just one good experiment."

Although the Bem experiments themselves do not provide evidence for precognition, they do suggest that our academic standards of evidence may currently be set at a level that is too low (see also Wetzels et al., 2011). It is easy to blame Bem for presenting results that were obtained in part by exploration; it is also easy to blame Bem for possibly overestimating the evidence in favor of H_1 because he used p values instead of a test that considers H_0 vis-a-vis H_1. However, Bem played by the implicit rules that guide academic publishing—in fact, Bem presented many more studies than would usually be required. It would therefore be mistaken to interpret our assessment of the Bem experiments as an attack on research of unlikely phenomena; instead, our assessment suggests that something is deeply wrong with the way experimental psychologists design their studies and report their statistical results. It is a disturbing thought that many experimental findings, proudly and confidently reported in the literature as real, might in fact be based on statistical tests that are explorative and biased (see also Ioannidis, 2005). We hope the Bem article will become a signpost for change, a writing on the wall: psychologists must change the way they analyze their data.

8 An Agenda for Purely Confirmatory Research

Abstract

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. We illustrate our proposals with a confirmatory replication attempt of a study on ESP.

An excerpt of this chapter has been submitted as:
Wagenmakers, E.–J., Wetzels. R., Borsboom. D., van der Maas. H. L. J. & Kievit, R.A. (2012). An Agenda for Purely Confirmatory Research. *Perspectives on Psychological Science*.

You cannot find your starting hypothesis in your final results. It makes the stats go all wonky. – Ben Goldacre, 2009, p. 221, Bad Science.

Psychology is a challenging discipline. Empirical data are noisy, formal theory is scarce, and the processes of interest (e.g., attention, jealousy, loss aversion) cannot be observed directly. Nevertheless, psychologists have managed to generate many key insights about human cognition and behavior. For instance, research has shown that people tend to seek confirmation rather than disconfirmation of their beliefs – a phenomenon known as confirmation bias (Nickerson, 1998). Confirmation bias operates in at least three ways. First, ambiguous information is readily interpreted to be consistent with one's prior beliefs; second, people tend to search for information that confirms rather than disconfirms their preferred hypothesis; third, people more easily remember information that supports their position. We also know that people experience cognitive dissonance when the facts do not correspond to the desired state of the world, an unwelcome tension that people will attempt to reduce; moreover, we know that people fall prey to hindsight bias, the tendency to judge an event as more predictable after it has occurred (Christensen–Szalanski & Willham, 1991).

In light of these and other human biases[1] it would be naive to believe that, without special protective measures, the scientific research process is somehow exempt from these systematic imperfections of the mind. When bias influences the research process this means that researchers seek to confirm, not falsify, their main hypothesis (Sterling, 1959; Sterling, Rosenbaum, & Weinkam, 1995). This is especially relevant in an environment that puts a premium on output quantity: when academic survival depends on how many papers one publishes, researchers are attracted to methods and procedures that maximize the probability of publication (Bakker, van Dijk, & Wicherts, in press; John, Loewenstein, & Prelec, 2012; Nosek, Spies, & Motyl, in press; Neuroskeptic, in press). It should be noted that such behavior is ecologically rational in the sense that it maximizes the proximal goals of the researcher. However, when each researcher acts this way in an entirely understandable attempt at academic self-preservation, the cumulative effect on the field as a whole can be catastrophic. The primary concern is that many published results may simply be false, as they have been obtained partly by dubious or inappropriate methods of observation, analysis, and reporting (Jasny, Chin, Chong, & Vignieri, 2011; Sarewitz, 2012).

Several years ago, Ioannidis (2005) famously argued that "most published research findings are false". And indeed, recent results from biomedical and cancer research suggest that replication rates are lower than 50%, with some as low as 11% (Begley & Ellis, 2012; Osherovich, 2011; Prinz, Schlange, & Asadullah, 2011). If the above results carry over to psychology, this suggests that our discipline is in serious trouble (S. Carpenter, 2012; Roediger, 2012; Yong, 2012). Research findings that do not replicate are worse than fairy tales; with fairy tales the reader is at least aware that the work is fictional.

In this article we first discuss four popular practices that result in bad science[2]; we call these "fairy tale factors", because each factor increases the probability that a presented finding is fictional and hence non-replicable. Next we propose two radical remedies to ensure scientific integrity and inoculate the research process against the inalienable biases of human reasoning. We end by illustrating the remedies to a replication attempt of an ESP experiment reported by Bem (2011).

[1] For an overview see for instance `http://en.wikipedia.org/wiki/List_of_cognitive_biases`.
[2] This list is not meant to be exhaustive.

8.1 Bad Science

Science can be bad in many ways. Flawed design, faulty logic, and limited scholarship engender no enthusiasm whatsoever.[3] Here we list four factors that bias the research process and make experimental results appear to be more compelling than they really are (see also Simmons et al., 2011).

Fairy Tale Factor 1: Exploratory Analyses, Confirmatory Conclusions

Our main concern is that almost no psychological research is conducted in a purely confirmatory fashion (e.g., Kerr, 1998; Wagenmakers et al., in press). Only rarely do psychologists indicate, in advance of data collection, the specific analyses they intend to carry out. In the face of human biases and the vested interest of the experimenter, such freedom of analysis provides access to a Pandora's box of tricks that can be used to achieve any desired result (e.g., John et al., 2012; Simmons et al., 2011; for what may happen to psychologists in the afterlife see Neuroskeptic, in press). For instance, researchers can engage in cherry-picking: they can measure many variables (gender, personality characteristics, age, etc.) and only report those that yield the desired result; they can include in their papers only those experiments that produced the desired outcome, even though these experiments were designed as pilot experiments, ready to be discarded had the results turned out less favorably. Researchers can also explore various transformations of the data, rely on one-sided p values, and construct post-hoc hypotheses that have been tailored to fit the observed data. In the past decades, the development of statistical software has resulted in a situation where the number of opportunities for massaging the data is virtually infinite.

True, researchers may not use these tricks with the explicit purpose to deceive—for instance, hindsight bias often makes exploratory findings appear perfectly sensible. Even researchers who advise their students to "torture the data until they confess"[4] are hardly evil geniuses out to deceive the public or their peers. Instead, these researchers may genuinely believe that they are giving valuable advice that leads the student to analyze the data more thoroughly, increasing the odds of a publication along the way. How could such advice be wrong?

In fact, the advice to torture the data until they confess is not wrong – just as long as this torture is clearly acknowledged in the research report. Academic deceit sets in when this does not happen and partly exploratory research is analyzed as if it had been completely confirmatory. At the heart of the problem lies the statistical law that, for the purpose of *hypothesis testing*, the data may be used only once. So when you turn your data set inside and out, looking for interesting patterns, you have used the data to help you formulate a specific hypothesis. Although the data may still serve many purposes after such fishing expeditions, there is one purpose for which the data are no longer appropriate; namely, for testing the hypothesis that they helped to suggest. Just like conspiracy theories are never disproved by the facts that they were designed to explain, a hypothesis that is developed on the basis of exploration of a data set is unlikely to be refuted by that same data. Thus, for testing one's hypothesis, one always needs a fresh data set. This also means that the interpretation of common statistical tests in terms of

[3] We are indebted to an anonymous reviewer of a different paper for bringing this sentence to our attention.

[4] The expression is attributed to Ronald Coase. Earlier, Mackay (1852/1932, p. 552) made a similar statement, one that is perhaps even more apt: "When men wish to construct or support a theory, how they torture facts into their service!".

type I and type II error rates is valid only if (a) the data were used only once, and (b) the statistical test was not chosen on the basis of relevant characteristics of the data. If you carry out a hypothesis test on the very data that inspired that test in the first place then the statistics are invalid (or "wonky", as Ben Goldacre put it). In neuroimaging, this has been referred to as double dipping (Kriegeskorte, Simmons, Bellgowan, & Baker, 2009; Vul, Harris, Winkielman, & Pashler, 2009). If a researcher uses double dipping strategies, type I error rates will be inflated considerably, and as a result p values are no longer trustworthy.

As illustrated in Figure 8.1, psychological studies can be placed on a continuum from purely exploratory, where the hypothesis is found in the data, to purely confirmatory, where the entire analysis plan has been explicated before the first participant is tested. Every study in psychology falls somewhere along this continuum; the exact location may differ depending on the initial outcome (i.e., poor initial results may encourage exploration), the clarity of the research question (i.e., vague questions allow more exploration), the amount of data collected (i.e., more dependent variables encourage more exploration), the *a priori* beliefs of the researcher (i.e., strong belief in the presence of an effect encourages exploration when the initial result is ambiguous), and so on. Hence, the amount of exploration, data dredging, or data torture may differ widely from one study to the next; consequently, so does the reliability of the statistical results. It is important to stress again that we do not disapprove of exploratory research as long as its exploratory character is openly acknowledged. If fishing expeditions are sold as hypothesis tests, however, it becomes impossible to judge the strength of the evidence reported.

Fairy Tale Factor 2: Publication Bias or Aversion to the Null

Few researchers like null results. Compared to statistically significant results (i.e., $p < .05$), null results (i.e., $p > .1$) are inherently ambiguous: perhaps the research was carried out poorly, or perhaps the experiment did not have enough power. Also, null results are sometimes uninteresting (e.g., "people are just as creative when they are inside or outside a large box"). When only positive studies are published and the null results are rejected or disappear into the file drawer, the literature does not fairly represent the true state of affairs. There is ample evidence that the file drawer effect in psychology is rather large; for instance, Sterling et al. (1995) found that more than 95% of articles in psychology journals confirm their main hypothesis (see Bones, 2012, for an alternative account).

Fairy Tale Factor 3: Optional Stopping

Optional stopping or "sampling to a foregone conclusion" is a popular method of data collection that, within the framework of p value hypothesis testing, is nevertheless tantamount to cheating (e.g., Jennison & Turnbull, 1990; Strube, 2006; Wagenmakers, 2007). The method consists of taking multiple looks at the data as they come in, and stopping data collection whenever the desired result is obtained. The problem is that the standard tests work as advertised only when the number of participants has been determined in advance. Without a correction for multiple looks, the probability of falsely rejecting the null hypothesis is larger than .05. It is again important to note that optional stopping in itself is not a problem, as long as it is openly reported and as long as the relevant statistics are corrected for this strategy.

The problem is perhaps clearest when a researcher tests 20 participants, finds $p = .11$, proceeds to test 10 more participants, and then reports $n = 30$ and $p = .04$, the latter value computed as if 30 participants were scheduled from the start. It is less clear that the

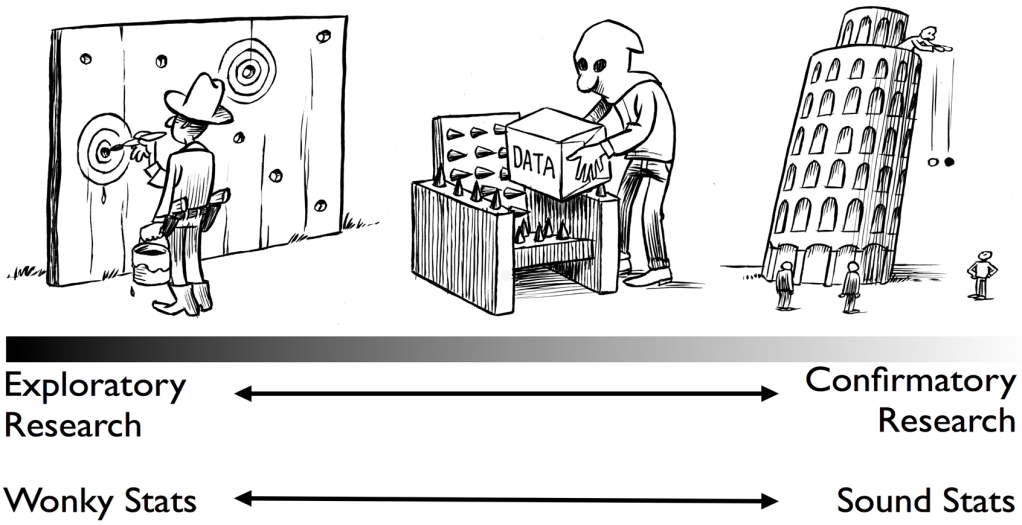


Figure 8.1: A continuum of experimental exploration and the corresponding continuum of statistical wonkiness. On the far left of the continuum, researchers find their hypothesis in the data by post-hoc theorizing, and the corresponding statistics are "wonky", dramatically overestimating the evidence for the hypothesis. On the far right of the continuum, researchers pre-register their studies such that data collection and data analyses leave no room whatsoever for exploration; the corresponding statistics are "sound" in the sense that they are used for their intended purpose. Much empirical research operates somewhere in between these two extremes, although for any specific study the exact location may be impossible to determine. In the grey area of exploration, data are tortured to some extent, and the corresponding statistics are somewhat wonky. Figure downloaded from Flickr, courtesy of Dirk-Jan Hoek.

problem of optional stopping may also exist when a researcher never takes a sneak peak at the data at all. For instance, a researcher may test 30 participants in one sitting, find $p = .04$, and report the result – and this could nevertheless be tantamount to cheating. This is because one may ask the researcher "what would you have done in case the results had not been significant after the initial 30 participants"? One possible answer is "When the p value is higher than .05 but lower than .15 I would have tested 10 participants more". This answer reveals that the sampling plan did not fix the number of participants in advance, and hence the problem of optional stopping persists.

Fairy Tale Factor 4: p Values Overestimate the Evidence Against the Null

The p value is the probability of obtaining a value for a test statistic that is at least as extreme as the one that was actually observed, given that the null hypothesis is true and relevant statistical assumptions are met (these may involve characteristics of the data, like normality and linearity, or characteristics of the research process, such as a the prior specification of a sampling plan). If the p value is low, researchers reject the null hypothesis and assume, explicitly or implicitly, that the alternative hypothesis is much better supported by the data. But this reasoning is fallacious. The observed data (summarized

by a test statistic) could be just as unlikely under the alternative hypothesis. Thus, what matters is the relative likelihood of the data under various competing explanations, not the probability of the data under a single explanation (e.g., D. V. Lindley, 1993).

Because the p value is not comparative, it can be shown to overestimate the evidence against the null. In particular, Sellke et al. (2001) considered the diagnosticity of the p value: how much more likely is a particular observed p value under H_1 than it is under H_0? The results are shocking. A p value of $p = .037$, for instance, is at best 3 times more likely under H_1 than under H_0; a p value of $p = .01$ is at best 8 times more likely under H_1 than under H_0. Since these values are upper bounds, derived by cherry-picking the single most competitive H_1, the true diagnosticity is likely to be even lower.

When researchers use a low-diagnostic criterion to detect experimental effects, it should come as no surprise if many of these effects do not replicate.

Together, these and other fairy tale factors create a perfect storm that threatens to unravel the very fabric of our field. This special issue features several papers that propose remedies to right what is wrong, for instance through changes in incentive structures (Nosek et al., in press) and an increased focus on replicability (Bakker et al., in press; Frank & Saxe, in press; Grahe et al., in press). In the next section we stress two radical remedies that hold great promise, not just for the state of the entire field but also for researchers individually.

8.2 Good Science

Science can be good in many ways, but a key characteristic is that the researcher is honest. Unfortunately, a call for more honesty is unlikely to change anything. Blinded by confirmation bias and hindsight bias, researchers may be convinced that they are honest even when they are not. We therefore focus on a more tangible characteristic of good science, namely that it should minimize the impact of the fairy tale factors discussed above.

Solution 1: Separate Exploratory from Confirmatory Experiments

The articles by Simmons et al. (2011) and John et al. (2012) suggest to us that considerable care needs to be taken before researchers are allowed near their own data: they may well torture them until a confession is obtained, even if the data are perfectly innocent. More importantly, researchers may then proceed to analyze and report their data as if these had undergone a spa treatment rather than torture. Psychology is not the only discipline in which exploratory methods masquerade as confirmatory, thereby polluting the field and eroding public trust (Sarewitz, 2012). In his fascinating book "Bad Science", Ben Goldacre discusses several fairy tale factors in public health science and medicine, and concludes:

> "What's truly extraordinary is that almost all of these problems – the suppression of negative results, data dredging, hiding unhelpful data, and more – could largely be solved with one very simple intervention that would cost almost nothing: a clinical trial register, public, open, and properly enforced (...) Before you even start your study, you publish the 'protocol' for it, the methods section of the paper, somewhere public. This means that everyone can see what you're going to do in your trial, what you're going to measure, how, in how many people, and so on, *before you start*. The problems of publication bias, duplicate publication and hidden data on side-effects – which all

> cause unnecessary death and suffering – would be eradicated overnight, in one fell swoop. If you registered a trial, and conducted it, but it didn't appear in the literature, it would stick out like a sore thumb." (Goldacre, 2009, pp. 220-221)

We believe this idea has great potential for psychological science as well (see also Bakker et al., in press; Nosek et al., in press, and the NeuroSkeptic blog[5]) By pre-registering the study design and the analysis plan, the first fairy tale factor (i.e., presenting and analyzing exploratory results as if they were confirmatory) is eliminated entirely. The second factor (i.e., aversion to the null) is also avoided, as non-published pre-registered experiments "stick out like a sore thumb".

To some, pre-registering an experiment may seem a Draconian measure. To us, this response only highlights how exceptional it is for psychologists to commit to a specific method of analysis in advance of data collection. Also, we wish to emphasize that we have nothing against exploratory work per se. Exploration is an essential component of science, key to new discoveries and scientific progress; without exploratory studies the scientific landscape is sterile and uninspiring. However, we do believe that it is important to separate exploratory from confirmatory work, and we do *not* believe that researchers can be trusted to observe this distinction if they are not forced to.[6]

Hence, in the first stage of a research program, researchers should feel free to conduct exploratory studies and do whatever they please: turn the data inside out, discard participants and trials at will, and enjoy the fishing expedition. However, exploratory studies cannot be presented as strong evidence in favor of a particular claim; instead, the focus of exploratory work should be on describing interesting aspects of the data, on determining which tentative findings are of particular interest, and on proposing efficient ways in which future studies may confirm or disconfirm the initial exploratory results.

In the second stage of a research program, a purely confirmatory approach is desired. This requires that the psychological science community set up an online repository comparable to the usual article submission websites such as *Manuscript Central*.[7] Before a single participant is tested, the researcher submits to the online repository a document that details what dependent variables will be collected and how the data will be analyzed (i.e., which hypotheses are of interest, which statistical tests will be used, and which outlier criteria or data transformations will be applied). When p-values are used, the researcher also needs to indicate exactly how many participants will be tested. When researchers wish to claim that their studies are confirmatory, the online document then becomes part of the review process.

An attractive implementation of this two-step procedure is to collect the data all at once and then split the data in an exploratory and a confirmatory subset. For example, researchers can decide to freely analyze only the even-numbered participants, exploring the data however they like. In the next stage, however, the favored hypothesis can be tested on the odd-numbered participants in a purely confirmatory fashion. To enforce academic self-discipline, the second stage still requires pre-registration. Although it is always possible for researchers to cheat, the main advantage of pre-registration is that it removes the effects of confirmation bias and hindsight bias. In addition, researchers

[5] See in particular `http://neuroskeptic.blogspot.co.uk/2008/11/registration-not-just-for-clinical.html`, `http://neuroskeptic.blogspot.co.uk/2011/05/how-to-fix-science.html`, and `http://neuroskeptic.blogspot.co.uk/2012/04/fixing-science-systems-and-politics.html`.

[6] This should not be taken personally: we distrust ourselves as well.

[7] The Open Science Framework may provide such a service, see `http://openscienceframework.org/`.

who cheat with respect to pre-registration of experiments are well aware that they have committed a serious academic offense.

What we propose is a method to ensure academic honesty: there is nothing wrong with exploration as long as it is explicitly acknowledged as such. The only way to safeguard academics against fooling themselves, their readers, reviewers, and the general public, is to demand that confirmatory results are clearly separated from work that is exploratory. In a way, our proposal is merely a matter of common sense, and we have not met many colleagues who wish to argue against it; nevertheless, we know of almost no research in experimental psychology that follows this procedure.

Solution 2: Bayesian Hypothesis Tests

Fairy tale factor four (i.e., p values overestimate the evidence against the null) can be eliminated by calculating a comparative measure of evidence. One such measure, the one we focus on here, is the Bayes factor. The Bayes factor BF_{01} quantifies the evidence that the data provide for H_0 vis-a-vis H_1 (Hoijtink et al., 2008; Jeffreys, 1961; Kass & Raftery, 1995; Masson, 2011; Rouder et al., 2009; Wagenmakers et al., 2010; Wetzels et al., 2009). For instance, when $BF_{01} = 10$ the observed data are 10 times as likely to have occurred under H_0 than under H_1. When $BF_{01} = 1/5 = .20$ the observed data are 5 times as likely to have occurred under H_1 than under H_0. It is important to realize that the Bayes factor quantifies the evidence brought about by the data and does not in any way depend on the prior probabilities that are assigned to H_0 and H_1.

Thus, the complication with the Bayes factor does not reside in the prior probabilities that are assigned to the hypotheses. Instead, the main complication lies in the requirement to fully specify H_1. In particular, we need to state what effect sizes can be expected should H_1 be true; in Bayesian terminology, we need to assign effect size a prior distribution. This is a choice that should be made judiciously, and much work in Bayesian statistics has concerned this crucial problem (Jeffreys, 1961; Liang et al., 2008; Rouder et al., 2009; Zellner & Siow, 1980). One method, the one we prefer, is to use a default specification process based on general principles (i.e., an "objective Bayesian hypothesis test").[8] Another method is to use substantive knowledge and specify prior distributions for effect size on the inference problem at hand. Of course, such subjective specification should be carried out before the data are analyzed in order to prevent the data analyst from falling prey to hindsight bias and assigning effect size a prior distribution that matches the observed data too closely.

An additional bonus of using the Bayes factor is that it eliminates fairy tale factor three (i.e., optional stopping). As noted in the classic article by Edwards et al. (1963, p. 193), "the rules governing when data collection stops are irrelevant to data interpretation. It is entirely appropriate to collect data until a point has been proven or disproven, or until the data collector runs out of time, money, or patience." (see also Kerridge, 1963). This means that researchers who use the Bayes factor should feel entirely uninhibited to continue data collection in case the initial results are not sufficiently compelling. Likewise, when early results are compelling, researchers who use the Bayes factor can just stop data collection and report the result without feeling any pressure to continue collecting more data. Space constraints prevent us from discussing the other advantages of using Bayes factors (e.g., the ability to quantify evidence in favor of H_0, the fact that discovery of the truth is guaranteed as the number of data points increases).[9]

[8] One method is to use unit-information priors, that is, priors that contain as much information as a single observation.

[9] Denny Borsboom wishes to state that, unlike the first author, he is not an evangelical Bayesian

8.3 Example: Precognitive Detection of Erotic Stimuli?

In 2011, Dr. Bem published an article in the *Journal of Personality and Social Psychology*, the flagship journal of social psychology, in which he claimed that people can look into the future (Bem, 2011). In his first experiment, "precognitive detection of erotic stimuli", participants were instructed as follows: "(...) on each trial of the experiment, pictures of two curtains will appear on the screen side by side. One of them has a picture behind it; the other has a blank wall behind it. Your task is to click on the curtain that you feel has the picture behind it. The curtain will then open, permitting you to see if you selected the correct curtain." In the experiment, the location of the pictures was random and chance performance is therefore 50%. Nevertheless, Bem's participants scored 53.1%, significantly higher than chance; however, the effect was present only for erotic pictures, and not for neutral pictures, positive pictures, negative pictures, and romantic-but-not-erotic pictures. Bem also claimed that the psi effects are more pronounced for extraverts, and that for certain erotic pictures women show psi but men do not.

In order to illustrate our proposal we set out to replicate Bem's experiment in a purely confirmatory fashion. First we detailed our method, design, and planned analyses in a document that we posted online before a single participant was tested.[10] As outlined in the online document, our replication focused on Bem's key findings; therefore, we tested only women, used only neutral and erotic pictures, and included a standard extraversion questionnaire. We also tested each participant in two contiguous sessions. Each session featured the same pictures, but presented them in a different random order. The idea is that individual differences in psi –if these exist– lead to a positive correlation between performance on session 1 and session 2. Performance is quantified by the proportion of times that the participant chooses the curtain that hides the picture. Each session featured 60 trials, with 45 neutral pictures and 15 erotic pictures.

A vital part of the online document concerns the *a priori* specification of our analyses. First we outlined our main analysis tool, the Bayes factor t-test:

> "Data analysis proceeds by a series of Bayesian tests. For the Bayesian t-tests, the null hypothesis H_0 is always specified as the absence of a difference. Alternative hypothesis 1, H_1, assumes that effect size is distributed as Cauchy(0,1); this is the default prior proposed by Rouder et al. (2009). Alternative hypothesis 2, H_2, assumes that effect size is distributed as a half-normal distribution with positive mass only and the 90^{th} percentile at an effect size of 0.5; this is the "knowledge-based prior" proposed by Bem et al. (submitted).[11] We will compute the Bayes factor for H_0 vs. H_1 (BF_{01}) and for H_0 vs. H_2 (BF_{02})."

Next we outlined a series of six hypotheses to test. For instance, the second analysis was specified as follows:

> "(2) Based on the data of session 1 only: Does performance for erotic pictures differ from chance (in this study 50%)? To address this question we

fundamentalist. Han van der Maas does not know what he is, Rogier Kievit is an agnostic pragmatist, and, as a graduate student of the first author, Ruud Wetzels has no choice in the matter whatsoever. All authors agree, however, that it is important to utilize methods that give the null hypothesis a fair chance in data analysis.

[10] See `http://confrep.blogspot.nl/` and `http://dl.dropbox.com/u/1018886/Advance_Information_on_Experiment_and_Analysis.pdf`.

[11] Current addendum: this paper has since been published (i.e., Bem, Utts, & Johnson, 2011).

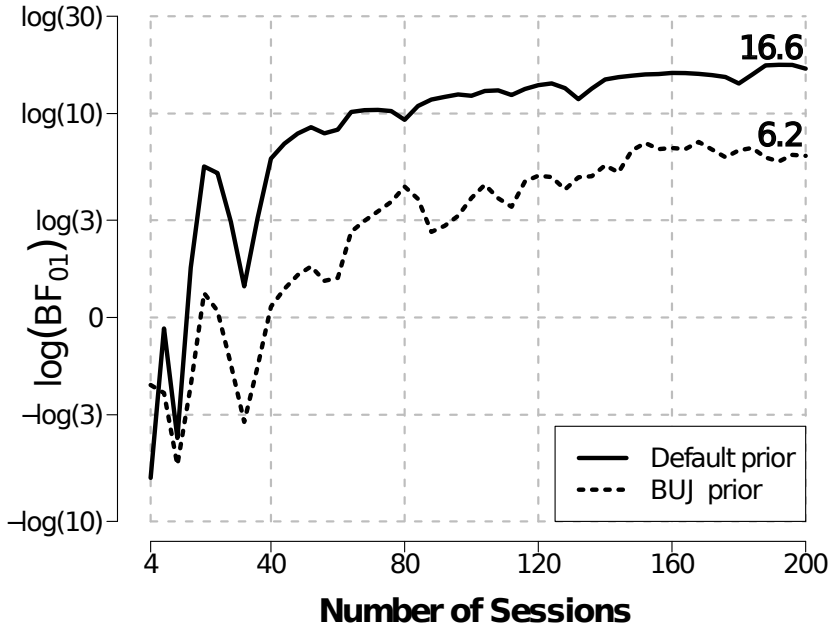


Figure 8.2: Results from a purely confirmatory replication test for the presence of precognition. The intended analysis was specified online in advance of data collection. The evidence (i.e., the logarithm of the Bayes factor) supports "H_0 : performance for erotic stimuli does not differ from chance". Note that the evidence may be monitored as the data accumulate. See text for details.

> compute a one-sample t-test and monitor BF_{01} and BF_{02} as the data come in."

And the sixth analysis was specified as follows:

> "(6) Same as (2), but now for the combined data from sessions 1 and 2."

Readers curious to know whether people can look into the future are invited to examine the results for all six hypotheses in an online appendix.[1]2 Here we only present the results from our sixth hypothesis. Figure 8.2 shows the development of the Bayes factor as the data accumulate. It is clear that the evidence in favor of H_0 increases as more participants are tested and the number of sessions increases. With the default prior the data are 16.6 times more likely under H_0 than under H_1; with the "knowledge-based prior" from Bem et al. (2011) the data are 6.2 times more likely under H_0 than under H_1. Note that we did not have to indicate in advance that we were going to test 100 participants. We calculated the Bayes factor two or three times as the experiment was running, and after 100 participants we inspected Figure 8.2 and decided that for the present purposes the results were sufficiently compelling. Note how the Bayes factor can be used to quantify evidence in favor of the null hypothesis.

[1]2Available from the first author's webpage or directly from `https://dl.dropbox.com/u/1018886/Appendix_PoPS_WagenmakersEtAl.pdf`.

The results reported here are *purely confirmatory*: absolutely everything that we have done here was decided before we saw the data. In this respect, these results are exceptional in experimental psychology, a state of affairs that we hope will change in the future.

Naturally, it is possible that our data had shown something unexpected and interesting, or that we could have forgotten to include an important analysis in our pre-registration document. It is also possible that reviewers of this paper will ask for additional information (e.g., a credible interval for effect size). How should we deal with such alterations of the original data-analysis scheme? We suggest that, rather than walking the fine line of trying to decide which alterations are appropriate and which are not, *all* such findings and analyses should be mentioned in a separate section entitled "exploratory results". When such exploratory results are analyzed it is important to realize that the data have been used more than once, and the inferential statistics may therefore to some extent be wonky.

Pre-registration of our study was sub-optimal. The key document was posted on the first author's website and a purpose-made blog, and therefore the file would have been easy to alter, remove, or ignore. With the online resources of the current day, however, the field should find it easy to construct a professional repository to push academic honesty to greater heights. We believe that researchers who use pre-registration will quickly realize how different this procedure is from what is now standard practice. Top journals could facilitate the transition to more confirmatory research by implementing a policy to reward empirical manuscripts that feature at least one confirmatory experiment; for instance, these manuscripts could be published in a separate section explicitly containing "confirmatory research". We hope that our proposal will increase the transparency of the scientific process, diminish the proportion of false findings, and improve the status of psychology as a rigorous scientific discipline.

9 Discussion

9.1 Discussion

In this thesis we have proposed Bayesian alternatives to frequentist null hypothesis tests. More specifically, we have outlined a Bayesian t test, a Bayesian correlation test, a Bayesian test for partial correlations, and a Bayesian one-way and two-way ANOVA. All these tests are essential tools for empirical research in psychology.

In this thesis we also compared the proposed Bayesian null hypothesis tests to their frequentist counterparts, discussed their behavior, and explained how and when these tests can be applied. In the second part of this thesis we discussed the practical benefits of the Bayesian null hypothesis tests, that is, we explain how social science can benefit from applying Bayesian methods. This dissertation points out several Bayesian solutions to problems that concern 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.

In the remainder of this discussion, we will first recap four main problems with p value hypothesis testing that social science research is confronted with. Then we will discuss how the application of Bayesian can be of help, or even solve the problem.

Bayesian Methods Allow Evidence in Favor of the Null Hypothesis

Bayesian methods allow researchers to gather support in favor of the null hypothesis. This is an important feature because the current social science literature has serious problems with ad-hoc theories and models that are being discussed in the literature as being "true" while they might very well be false. This is a concern to psychological scientists, because when a certain theory is established in the literature, it is relatively difficult to overthrow. The reason is that, within the frequentist framework of null hypothesis testing, it is impossible to gather evidence in favor of the null. This makes it difficult to eliminate false results.

Fortunately, researchers can gather evidence in favor of the null hypothesis when they compute a Bayes factor that contrasts the null hypothesis with a specific alternative hypothesis. Hence, the Bayes factor allows researchers to disprove false theories more easily. This will greatly benefit the social sciences, as it enables researchers to publish the results of replication attempts for well-known experiments, even when the original finding is not replicated and evidence in favor of the null hypothesis is found instead. By encouraging replication studies, the evaluation of psychological theories and models becomes easier, something scientists can only benefit from.

Bayesian Methods Treat the Alternative and Null Hypothesis Alike

When a frequentist null hypothesis test is conducted, the alternative hypothesis is not evaluated. More specifically, the question that is not evaluated is whether the data are likely under the alternative hypothesis. In psychology, the alternative hypothesis is usually (implicitly) considered to be the theory or model that is being tested. It only

seems sensible to evaluate the plausibility of the data under both the null hypothesis and the alternative hypothesis.

The Bayes factor does compare the two models directly. The alternative theory is being evaluated just as the null hypothesis, resulting in a balanced, comparative measure of evidence. Consider cases where the data are highly unlikely under the null hypothesis but also highly unlikely under the alternative hypothesis. In such cases the p value rejects the null hypothesis whereas the Bayes factor indicates that the data are inconclusive. We believe that the latter approach is more sensible and it is more closely linked to the research question at hand.

Bayes Factors are More Easily Interpreted Than the p Value

Many people misinterpret the p value as the probability of the null hypothesis being true. This is not correct as the p value is a conditional probability: It is the probability of the data, or data more extreme, given that the null hypothesis is true and given a specific design. This is a complicated probability to interpret, as shown by a famous quote of Jeffreys: "What the use of p implies, therefore, is that a hypothesis that may be true may be rejected because it has not predicted observable results that have not occurred." (1961, p. 385).

Because of this confusing definition, many researchers seem to confuse this probability with its transposed probability, the probability of the null hypothesis given the data. These two probabilities are not the same, as the p value is calculated given that the null hypothesis is true, therefore it can not also be the probability that the null hypothesis is true. A clear example that $p(D \mid L) \neq p(L \mid D)$ is that when D is the event of someone dying, and L is the event of someone being hit by lightning. It is clear that the $p(D \mid L)$, the probability of someone dying when she was hit by lightning is much larger than $p(L \mid D)$, the probability that someone who is dead was hit by lightning (there are of course many other ways to die, and not many people die by being hit by lightning).

More specifically, in the specific case of null hypothesis testing we can show that these two conditional probabilities are not the same by Lindley's paradox. In his article, D. V. Lindley (1957) gives an example that shows that if H is a simple hypothesis, and y the result of an experiment, the following two phenomena can occur at the same time:

1 a significance test for H reveals that y is significant at the 5% level

2 the posterior probability of H, given y, is, for quite small prior probabilities of H, as high as 95%.

It might be that the p value is so often misinterpreted as the probability of the null hypothesis given the observed data, because this is often the probability that researchers want to calculate. Therefore, computing the Bayes factor might be a convenient solution to this misinterpretation. The Bayes factor has a clear interpretation as the change from prior to posterior odds, brought about by the data. Assuming that this is what researchers are interested in, applying Bayesian methods would not only yield a measure that is easier to interpret, but it would also yield a measure that gives an answer to the question that is being asked.

Bayes Factors are Not Vulnerable to Optional Stopping

The optional stopping problem is usually defined as the fact that the p value has an exact interpretation as the probability of the observed data or data more extreme, given that

the null hypothesis is true and given a pre-defined design and sample size. Hence, after data collection starts one is not allowed to stop before the experiment was supposed to be finished and interpret the p value. In the same vain, one is not allowed to continue testing after the predefined sample size is reached.

In sum, researchers are not allowed to monitor the p value as the data come in and stop when it falls below .05 (the usual critical value below which the data is considered significant). At the same time, researchers are not allowed to continue testing when the planned sample size is reached. When one computes a p value, both early stopping and continued testing are considered to be statistical cheating.

However, in some situations it can even be unethical *not* to practice optional stopping. Let's assume that a researcher conducts an experiment to investigate whether a new medicine to treat a disease has a positive effect. She constructs a control group with participants receiving placebo treatment and an experimental group with participants receiving the new medicine. Each participant receives one pill each week for a total of 20 weeks. Now, what if the new medicine is so successful that after 10 weeks it is obvious that it cures the disease much more effectively than the old treatment? Then, according to the statistical rules, the researcher is not allowed to stop experimenting and make the new medicine available to the patients in the control group. However, it is arguably not ethical to prevent the control group from using the medicine, as the patients in the control group may experience needless discomfort (or even death) before the experiment is finished.

Bayesian model selection is not vulnerable to the optional stopping problem. Edwards, Lindman, and Savage (1963, p. 193) note that "the rules governing when data collection stops are irrelevant to data interpretation. It is entirely appropriate to collect data until a point has been proven or disproven, or until the data collector runs out of time, money, or patience". Hence, the researcher can monitor the Bayes factor when the data come in. If the data are convincing enough, she can stop data collection and vice versa, if the data are not convincing enough, she can keep on adding data until she finds that her point has been proven.

9.2 Future Directions

In this dissertation, we have shown that researchers in social science can benefit greatly from adding Bayesian methods to their statistical toolbox. We have pointed out various pitfalls of frequentist p value null hypothesis testing and we have shown how computing Bayes factors can be used to circumvent these pitfalls.

However, Bayesian methods have only recently become popular. Hence, there is still a lot of room for development of Bayesian methods for mainstream scientific purposes. In the remaining part of the discussion, we list a few remaining issues for future research.

How to Make a Choice Between the Various Default Prior Distributions?

This thesis deals with the application of default priors for Bayesian model selection. Default priors are the preferred choice for standard testing situations, because we feel that a statistical test should be as objective as possible. Moreover, a statistical test can yield a reference point for the behavior of the Bayes factor when other prior distributions are used. Hence, a personal (i.e., subjective) prior distribution would be difficult to use

for standard testing situations, and the same holds for a prior distribution that is based on the data.

However, in recent years, different default priors have been proposed, each having slightly different properties and differing asymptotic behavior. This induces a new question that is slightly paradoxical, is the choice between various objective priors a subjective choice? We see at least two ways to study this problem, one is to investigate the behavior of the various default prior distribution which answers the more pragmatic question whether the choice for a particular prior makes a practical difference (see next subsection). Another way to study this problem is to investigate whether it would be possible to create one overarching prior distribution that encompasses all the default options.

What is the Behavior of the Various Default Priors for Different Linear Models?

Much statistical research is focused on linear models. As discussed earlier, there are different default prior distributions that can be used. We already indicated that it may be hard to choose one of these various "default" choices (e.g., the original g prior with $g = n$, $g = \#parameters$, $g = ...$; the Jeffreys-Zellner-Siow prior; one of the Liang et al. scale mixture priors). It is interesting to investigate whether in practice, these priors result in substantially differing conclusions. We have conducted an extensive simulation study comparing the most common default prior distributions, for linear models, generalized linear models, and generalized linear mixed models (results not reported in this thesis). The interim conclusion is that when a reasonable sample size is used, there is not much difference between the default priors. If this result holds more generally, maybe an Occam's razor for the specification of priors should be proposed: If various priors yield the same results, the least complex prior should be chosen.

How to Interpret the Bayes Factor Scale?

How to interpret the Bayes factor scale? Jeffreys proposed a scale for the interpretation of the Bayes factor, a scale that is used throughout this thesis. However, many prior distributions can be considered a valid choice and this choice influences the Bayes factor. Hence, if the choice for a prior is somewhat ad-hoc, then the resulting Bayes factor scale is also somewhat ad-hoc.

One potential solution to this problem is to interpret the Bayes factor scale in terms of statistical power. For example, Cohen's d is interpreted as follows; an effect size d below 0.3 is considered to be a "small" effect size, a d of 0.5 is a "medium" effect size and a d of 0.8 or higher is a "large" effect size. A researcher could combine the scale of Cohen and Jeffreys. One could calculate the sample sizes needed to obtain a certain Bayes factor, assuming a certain effect size. For example, if one expects a small effect size, $d = 0.3$ and one wishes to obtain a Bayes factor of 3, the expected sample size could be $n = 40$. However, if the researchers wish to obtain a higher Bayes factor of, say 30, the expected sample size could be $n = 200$. This information could be used to calibrate the Bayes factor and give it a scale that is interpretable across different prior distributions.

How to Choose a Prior Distribution that is Based on an Experimental Question or Design?

There are many different choices to make when it comes to choosing a prior distribution. In some situations, a default prior distribution is the preferred choice, but there might be

situations in which one would be better off choosing a subjective or an empirical prior. This depends on the scientific question that is being asked. We can imagine situations where implementing as much prior information in the model as possible makes sense. For example, when one is comparing two models that have a clear psychological interpretation. Then, when comparing these two models (or theories), the prior distributions are a substantive part of the psychological theory and hence should be chosen in line with this theory.

Vanpaemel (2010) discusses an example where he formalized three hypotheses that concern a decision maker who has to choose between two alternatives. The first hypothesis states total indifference between the alternatives. Hence, the probability of choosing alternative 1 over alternative 2 is equal to 0.5, $\theta = 0.5$. Hypothesis two is that the decision maker is biased towards one alternative over the other. The probability of choosing alternative 1 over alternative 2 could then be anything between zero and one, $\theta \sim \text{dbeta}(1, 1)$. If one assumes that there are correct and incorrect alternatives, one could formulate a the third hypothesis; the decision maker performs better than chance $\theta > 0.5$. Notice that the model equations of these three models are assumed to be the same for all hypotheses. The only difference, the difference that is psychologically relevant, is implemented in the model through the prior distributions on the parameter θ.

It would be convenient if the psychological community would decide when it is appropriate to use one of the various prior choices that are available, defining a choice protocol and distinguishing various experimental settings and combine these settings with a heuristic for choosing a prior distribution.

What are the Pros and Cons of Calculating the Bayes Factor Versus Parameter Estimation?

This thesis focuses on the calculation of the Bayes factor, comparing different models or hypotheses. However, the calculation of the Bayes factor is often not easy, and the influence of the prior on the Bayes factor is considerable. Hence, Bayes factors should be used and interpreted with care. In comparison, the influence of the prior distribution on parameter estimation is far less. To avoid complex discussions on the merits of Bayes factors, one could also revert to Bayesian parameter estimation. Note however, these two approaches are by no means mutually exclusive. A study giving a roadmap on how to combine these two approaches in psychological research would be valuable to researchers interested in applying Bayesian methods.

How to Handle the Prior Probability of a Model or Hypothesis?

It is difficult to interpret or specify the probability of a model or a hypothesis. For example, what does it mean that a model has a probability p of being true? Moreover, it is difficult to define how multiple comparisons should be taken into account. If one is comparing many different models at the same time, it might be important to take the prior probability of a specific model into account, based on the number of models, or maybe even based on the number of parameters in the model (Scott & Berger, 2010). There are ideas on how to do this for well-defined models but for psychological process models this is much more complicated.

Furthermore, what is the prior probability of the null model? For example, considering the t test, there is no definite agreement about whether the null hypothesis is ever exactly true. If it is true, is the null equally probable as the alternative hypothesis that there is

a difference between the means? Again, for psychologically relevant models the situation becomes even more complex. For the calculation of the Bayes factor these questions are not directly relevant, as the prior model probability is not taken into account. However, the questions themselves remain interesting and relevant to psychological researchers that are using Bayes for their inference.

Part III

Appendices

A Bayesian Parameter Estimation in the Expectancy Valence Model of the Iowa Gambling Task

Abstract

The purpose of the popular Iowa gambling task is to study decision making deficits in clinical populations by mimicking real-life decision making in an experimental context. Busemeyer and Stout (2002) proposed an "Expectancy Valence" reinforcement learning model that estimates three latent components which are assumed to jointly determine choice behavior in the Iowa gambling task: weighing of wins versus losses, memory for past payoffs, and response consistency. In this article 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.

An excerpt of this chapter has been published as:
Wetzels, R., Vandekerckhove, J., Tuerlinckx, F., & Wagenmakers, E.-J. (2010). Bayesian parameter estimation in the Expectancy Valence model of the Iowa gambling task. *Journal of Mathematical Psychology, 54*, 14-27.

Every neuroscientist knows the tale of Phineas Gage, the railroad worker who suffered an unfortunate accident: in 1848, an explosion drove an iron rod straight through Gage's frontal cortex. Although Gage miraculously survived the accident, the resultant brain trauma did cause a distinct change in his personality. Prior to the accident, Gage was capable and reliable, but after the accident he was described as impatient, stubborn, and impulsive. Gage was no longer able to plan ahead in order to achieve long-term goals.[1]

The symptoms of Phineas Gage are characteristic for patients with damage to the ventromedial prefrontal cortex (vmPFC). These patients often take irresponsible decisions and do not seem to learn from their mistakes. The observed real-life decision making deficits are not caused by low intelligence, as vmPFC patients generally perform adequately on standard IQ tests.

In order to study the decision making behavior of clinical populations such as vmPFC patients under controlled conditions, Bechara and Damasio developed the now-famous "Iowa gambling task" (IGT; Bechara, Damasio, Damasio, & Anderson, 1994; Bechara, Damasio, Tranel, & Damasio, 1997), described in more detail below. Successful performance on the IGT requires that participants learn to prefer cautious (i.e., low rewards, low losses) alternatives over risky (i.e., high rewards, high losses) alternatives. The IGT is one of the most often used clinical tools to study deficits in decision making, and it has been applied to older adults, chronic cocaine users, cannabis users, children, criminals, patients with Huntington disease, patients with Asperger's syndrome, patients with obsessive-compulsive disorder, patients with Parkinson's disease, etc. (see Caroselli, Hiscock, Scheibel, & Ingram, 2006; Crone & van der Molen, 2004; Yechiam, Busemeyer, Stout, & Bechara, 2005; Yechiam et al., 2008 and references therein).

Although most clinical populations perform relatively poorly on the IGT, in the sense that their learning rate is lower than that of normal controls, it is as yet unclear whether or not the poor performance of these different clinical groups has the same origin. The IGT is a relatively complex task that requires the participant to correctly integrate information, remember this information, and converge upon a decision. Poor performance on the IGT could be due to any of these subcomponents that together determine choice behavior. In order to address this issue formally one needs a reinforcement learning model for task performance in the IGT. Such a model was developed and popularized by Jerry Busemeyer, Julie Stout, Eldad Yechiam, and co-workers (Busemeyer & Stout, 2002; Stout, Busemeyer, Lin, Grant, & Bonson, 2004; Wood, Busemeyer, Koling, Cox, & Davis, 2005; Yechiam, Stout, Busemeyer, Rock, & Finn, 2005; Yechiam, Busemeyer, et al., 2005; Yechiam & Busemeyer, 2005; Yechiam et al., 2008), whose Expectancy Valence (EV) model can presently be considered the default model of performance in the IGT.

When researchers use the EV model to draw conclusions about underlying processes, it is of course important that they can rely on estimation routines to accurately recover parameter values. Despite its importance, much is still unknown about the statistical characteristics of parameter estimation in the EV model. The primary goal of the present article is to analyze and improve on the estimation routines that are currently standard in the field.

The outline of this article is as follows. Part I provides a detailed explanation of the IGT and the EV model. Part II discusses the statistical properties of the EV model when parameters are estimated using maximum likelihood. Part III outlines a Bayesian graphical model for the EV model, both for single participant analysis and for a hierarchical analysis. Part IV applies the standard maximum likelihood estimation and the novel

[1] For more information about Phineas Gage see for instance `http://www.deakin.edu.au/hmnbs/psychology/gagepage/`.

	Bad Decks		Good Decks	
	A	B	C	D
reward per trial	100	100	50	50
number of losses per 10 cards	5	1	5	1
loss per 10 cards	1250	1250	250	250
net profit per 10 cards	−250	−250	250	250

Table A.1: Rewards and Losses in the IGT. Cards from decks A and B yield higher rewards than cards from decks C and D, but they also yield higher losses. The net profit is highest for cards from decks C and D.

Bayesian estimation to data from an experiment that was designed to provide a test of specific influence.

A.1 Part I: Explanation of the Iowa Gambling Task and the Expectancy Valence Model

The Iowa Gambling Task

In the IGT, participants have to discover, through trial and error, the difference between risky and safe decisions. In the computerized version of the IGT, the participant starts with $2000 in play money. Next, the computer screen shows four decks of cards (A, B, C, and D), and the participant has to select a card from one of the decks. Each card is associated with a reward, but potentially also with a loss. The default payoff scheme is presented in Table A.1.

As can be seen from Table A.1, decks A and B yield a reward of $100 everytime a card from those decks is selected, compared to only $50 for decks C and D. However, the relatively large rewards associated with decks A and B are more than undone by large occasional losses; in five out of every ten selections from deck A, the reward is overshadowed by a loss that ranges from $150 to $350 for a total of $1250 for every ten selections. For deck B, only one out of every ten selections is accompanied by a loss, but this loss is a whopping $1250.

The rewards associated with decks C and D may be relatively meagre, but so are the losses; for deck C, five out of every ten selections yields a loss, ranging from $25 to $75 for a total of $250. For deck D, only one out every ten selections yields a loss, and that loss is $250. This means that it is in the participants' financial interest to avoid decks A and B (i.e., the bad decks with large rewards, but even larger losses) and prefer cards from decks C and D (i.e., the good decks with modest rewards, but relatively small losses). The fact that the A and B decks are bad, and the C and D decks are good is something that the participant has to discover through experience.

At the start of the IGT, the participant is told to maximize net profit. During the task, the participant is presented with a running tally of the net profit. The task terminates after the participant has made a certain number of card selections. Depending on the experiment, this number varies from 100 or 150 to as much as 250.

The Expectancy Valence Model

From a statistical perspective, the IGT is a so-called four-armed bandit problem (Berry & Fristedt, 1985). Bandit problems are a special case of the more general reinforcement learning problems, in which an agent has to learn an environment by choosing actions and experiencing the consequences of those actions (e.g., Estes, 1950; Steyvers, Lee, & Wagenmakers, 2008; Sutton & Barto, 1998). It is easy to formulate a reinforcement learning problem, but it is difficult to solve such a problem in an optimal fashion. Optimal performance depends on a delicate tradeoff between "exploration" and "exploitation"; in order to discover the best option, the agent first has to try out or explore the various opportunities. However, if the agent only has a limited number of trials left, it is optimal to gradually stop exploring and instead exploit the option that has turned out to produce the highest profit in the past.

Many reinforcement problems such as the IGT are practically impossible to solve optimally. However, the reinforcement literature contains several solutions that are sensible and produce relatively good results. Interestingly, the parameters of a reinforcement learning method can often be given a clear psychological interpretation (e.g., Daw, O'Doherty, Dayan, Seymour, & Dolan, 2006). The EV model developed by Busemeyer and Stout (2002) is a case in point.

The EV model proposes that choice behavior in the IGT comes about through the interaction of three latent psychological processes. Each of these three processes is vital to producing successful performance typified by an increase in preference for the good decks over the bad decks with increasing experience. First, the model assumes that the participant, after selecting a card from deck k, $k \in \{1, 2, 3, 4\}$ on trial t, calculates the resulting net profit or valence. This valence v_k is a combination of the experienced reward $W(t)$ and the experienced loss $L(t)$:

$$v_k(t) = (1 - w) \cdot W(t) + w \cdot L(t). \tag{A.1}$$

Thus, the first parameter of the EV model is w, the *attention weight* of losses relative to rewards, $w \in [0, 1]$. A rational decision maker would assign equal weight to losses and rewards and hence use $w = .5$. Stout et al. (2004) found that the mean value of w was .25 for chronic cocaine users, in contrast to .63 for control participants. This result supports the idea that, compared to normal controls, cocaine users focus on rewards and deemphasize the possible negative consequences of their behavior.

On the basis of the sequence of valences v_k experienced in the past, the participant forms an expectation Ev_k of the valence for deck k. In order to learn, new valences need to modify continually the expected valence Ev_k. If the experienced valence v_k is higher or lower than expected, Ev_k needs to be adjusted upward or downward, respectively. This intuition is captured by the equation

$$Ev_k(t + 1) = Ev_k(t) + a \cdot (v_k(t) - Ev_k(t)), \tag{A.2}$$

in which the *updating rate* $a \in [0, 1]$ determines the impact of recently experienced valences. A high value of a means that the participant quickly adjusts the expected valence as a result of recent experiences. As a consequence, such a participant pays little heed to past events and has limited memory. Wood et al. (2005) found that older adults have higher values of the updating rate parameter than younger adults. This means that older adults show relatively large recency effects and exhibit more rapid forgetting.

Upon first consideration, it may seem rational to always prefer the deck with the highest expected valence. This "greedy" strategy, however, leaves very little room for

exploration, and the danger is that the decision maker quickly gets stuck choosing an inferior option. What is needed is some procedure to ensure that participants initially explore the decks, and only after a certain number of trials decide to always prefer the deck with the highest expected valence. One of the standard reinforcement learning methods to achieve this is to use what is called softmax selection or Boltzmann exploration (Kaelbling, Littman, & Moore, 1996; Luce, 1959):

$$\Pr[S_k(t+1)] = \frac{\exp\left(\theta(t) Ev_k\right)}{\sum_{j=1}^{4} \exp\left(\theta(t) Ev_j\right)}. \tag{A.3}$$

In this equation, $1/\theta(t)$ is the "temperature" at trial t and $\Pr(S_k)$ is the probability of selecting a card from deck k. When the temperature is very high, deck preference is almost completely random, allowing for a lot of exploration. As the temperature decreases, deck preference is guided more and more by the expected valences. When the temperature is zero, participants always prefer the deck with the highest expected valence.

In the EV model, the temperature is assumed to vary with the number of observations according to

$$\theta(t) = (t/10)^c, \tag{A.4}$$

where c is the *response consistency* or sensitivity parameter. In fits to data, this parameter is usually constrained to the interval $[-5, 5]$. When c is positive, response consistency θ increases with the number of observations (i.e., the temperature $1/\theta$ decreases). This means that choices will be more and more guided by the expected valences. When c is negative, choices will become more and more random as the number of card selections increases. Busemeyer and Stout (2002) found that patients with Huntington's disease had negative values for the response consistency parameter, which indicates that these patients became tired or bored as the task progressed, and consequently started to select cards at random.

In sum, the Expectancy Valence model decomposes choice behavior in the Iowa gambling task in three components or parameters:

1. An attention weight parameter w that quantifies the weighting of losses versus rewards.

2. An updating rate parameter a that quantifies the memory for rewards and losses.

3. A response consistency parameter c that quantifies the amount of exploration.

Although several suggestions have been made to change minor aspects of the EV model, the version of the model that is currently preferred is the version that was originally proposed by Busemeyer and Stout (2002). Current practice is to estimate the parameters of the EV model separately for each participant through the method of maximum likelihood.

A.2 Part II: Maximum Likelihood Estimation

Researchers who work with the EV model generally estimate parameters by minimizing the sum of one-step-ahead prediction errors. That is, based on the feedback from the previous t card selections, the EV model uses Equation A.3 to assign probabilities to each of the four decks. These probabilities can be thought of probabilistic forecasts for card selection $t + 1$. The parameter values that yield the best forecasts are the point estimates that are used for further statistical analysis.

Specifically, let a sequence of T observations (e.g., all card selections and the associated feedback) be denoted by $y^T = (y_1, ..., y_T)$; for example, y_{t-1} denotes the $(t-1)$th individual observation, whereas y^{t-1} denotes the entire sequence of observations ranging from y_1 up to and including y_{t-1}. Here we quantify predictive performance for a single observation by the logarithmic loss function $-\ln \hat{p}_t(y_t)$, that is, the larger the probability that $\hat{p}_t$ (determined based on the previous observations y^{t-1}) assigns to the observed outcome y_t, the smaller the loss. Thus, in the current EV parameter estimation routines, participant-specific parameters w_i, a_i, and c_i are adjusted in order to find the point estimates that minimize the sum of the one-step-ahead prediction errors: $\sum_{t=1}^{T} -\ln p(y_t|y^{t-1}, w_i, a_i, c_i)$. The method of parameter estimation is applied to each individual participant i separately.

The above procedure of finding parameter point estimates is in fact equivalent to that of maximum likelihood estimation (MLE; for a tutorial see I. J. Myung, 2003). To see this, recall that MLE seeks to determine those parameters under which the occurrence of the observed data is most likely, that is, $\{\hat{w}_i, \hat{a}_i, \hat{c}_i\} = \text{argmax}_{\{w_i, a_i, c_i\}} p(y^T|w_i, a_i, c_i)$. From the definition of conditional probability, i.e., $p(y_t|y^{t-1}) = p(y^t)/p(y^{t-1})$, it follows that $p(y^T)$ may be decomposed as a series of sequential, "one-step-ahead" probabilistic predictions (Dawid, 1984; Wagenmakers, Grünwald, & Steyvers, 2006):

$$\begin{aligned} p(y^T|w_i, a_i, c_i) &= p(y_1, ..., y_T|w_i, a_i, c_i) \\ &= p(y_T|y^{T-1}, w_i, a_i, c_i)p(y_{T-1}|y^{T-2}, w_i, a_i, c_i)...p(y_2|y_1, w_i, a_i, c_i)p(y_1|w_i, a_i, c_i). \end{aligned} \tag{A.5}$$

Thus, Equation A.5 shows that the MLE point estimates that maximize $p(y^T)$ are the same as those that minimize the sum of one-step-ahead prediction errors under log loss, as $-\ln p(y^T|w_i, a_i, c_i) = \sum_{t=1}^{T} -\ln p(y_t|y^{t-1}, w_i, a_i, c_i)$.

In the next three sections, we use simulations to examine performance of maximum likelihood parameter estimation for the EV model.[2] In particular, we address the following three interrelated questions:

1. How well can the EV parameters be recovered for single simulated participants?

2. What are the correlations between the EV parameters across many simulated participants?

3. To what extent are the EV parameters identifiable?

Parameter Recovery for Single Synthetic Participants

The clinical contribution of the EV model is to allow researchers to decompose choice performance into three latent psychological processes. These psychological processes are represented by model parameters, and hence it is vital to know the extent to which these parameters are estimated accurately and reliably.

We addressed this issue by simulating 1,000 synthetic participants in a 150-trial IGT, all with exactly the same EV model parameters: $w = 0.5$, $a = 0.35$, and $c = 0.35$. The values of these parameters were informed by previous research that suggests these values to be fairly typical of choice performance in the IGT. We then used the standard MLE procedure to determine parameter point estimates separately for each of the 1,000 synthetic participants. Consistent with current practice, we constrained the c parameter such that $c \in [-5, 5]$. Parameters w and a are probabilities and hence $\{w, a\} \in [0, 1]$.

[2] MLE routines were programmed in R, a free software environment for statistical computing and graphics (R Development Core Team, 2004).

Figure A.1 shows the density estimates (i.e., smoothed normalized histograms consisting of 1,000 estimates) for the each parameter separately. It is clear that parameter estimation is relatively unbiased, that is, the true parameter value with which the data were generated is about equal to the mean of the 1,000 estimated parameter values. Specifically, the mean estimated values for w, a, and c are 0.54, 0.36, and 0.36, respectively.

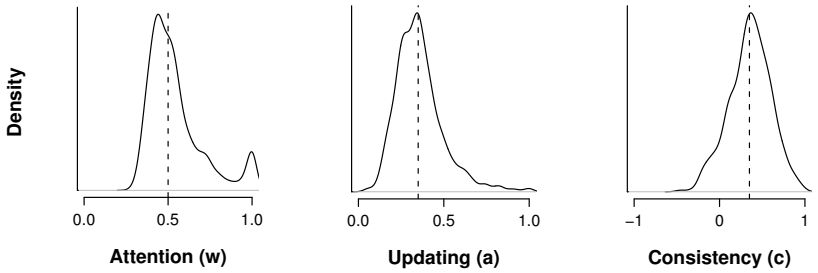


Figure A.1: EV parameter recovery for single participants. Dotted lines indicate true parameter values: Attention weight $w = 0.5$, updating rate $a = 0.35$, and response consistency $c = 0.35$. Data come from 1,000 synthetic participants, each completing a 150-trial IGT.

It is also clear that, for single participants, the variability in the estimates is considerable. In fact, this variability is so large that we believe it is hazardous to draw any kind of clinical conclusion based on the performance of an *individual* participant. For instance, an individual participant could have a perfectly normal updating rate of $a = .35$, but still stand a considerable chance of being assigned a point estimate that is either much lower or much higher.

Figure A.1 also reveals that the density of the parameter estimates for attention weight w is bimodal with a peak on the boundary of the parameter space. This is worrisome, as it indicates that, even when the true value of w is 0.5, a substantial proportion of participants will have a MLE of $\hat{w} = 1$; in the present simulation, this was the case for 50 out of 1,000 participants. We will revisit this issue later.

In sum, for single participants EV parameter recovery is virtually unbiased, but has relatively high variance. Of course, when the EV model is used in an experimental setting, high-variance individual parameter estimates are combined into a group average, and this group average has a much lower variability than the individual point estimates. However, the group averaging procedure ignores the commonalities that are shared by the participants within a particular group, a disadvantage that is remedied by the Bayesian hierarchical model proposed later.

Parameter Correlations Across Single Synthetic Participants

Ideally, parameter point estimates show little correlation across synthetic participants. The presence of such correlations could indicate that the effects of overestimating a certain parameter, say w, can be compensated by overestimating another parameter, say a. Such interactions between parameters lower the efficiency of parameter estimation and urge caution with respect to the ensuing statistical analysis (Ratcliff & Tuerlinckx, 2002, pp. 452–455).

To investigate this issue, we studied the correlational patterns between the parameters for the synthetic data described in the previous section. Figure A.2 plots the parameters

against each other. The dotted lines indicate the true parameter values. Figure A.2 shows that the correlation between attention weight w and updating rate a is positive but not very strong (i.e., $r = .20$). However, there is a substantial negative correlation between attention weight w and response consistency c (i.e., $r = -.53$); in other words, synthetic participants who appear to pay relatively much attention to losses will also appear to have a relatively low choice consistency. The relationship between updating rate a and response consistency c is also negative (i.e., $r = -.33$), such that synthetic participants who appear to have a relatively high updating rate will also appear to have a relatively low choice consistency.

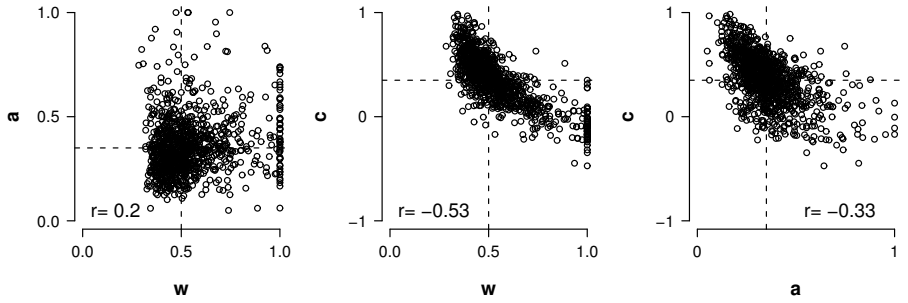


Figure A.2: EV parameter correlations based on MLEs from 1,000 synthetic participants, each completing a 150-trial IGT. Dotted lines indicate true parameter values: Attention weight $w = 0.5$, updating rate $a = 0.35$, and response consistency $c = 0.35$.

Finally, Figure A.2 also highlights the substantial variability in the parameter recovery for individual participants, and shows again the fact that several of the MLEs for w are on the boundary of the parameter space (i.e., $w = 1$).

Identifiability Within Single Human and Synthetic Participants

The previous two sections have revealed high variability of parameter recovery, and substantial correlations between parameter values across synthetic participants. These results suggest that, at least on the level of an individual participant, maximum likelihood parameter estimation in the EV model may suffer from a problem of identifiability. That is, it may be difficult in the particular probabilistic environment of the IGT to determine uniquely the most likely values for the parameters.

To examine the issue of identifiability more closely, we plotted *log likelihood contours* or log likelihood landscapes, that is, graphs that show how the logarithm of the likelihood changes across different parameter values for w, a, and c. Ideally, a log likelihood landscape has a single, pronounced peak that falls off equally quickly in all directions.

For the first log likelihood contour plot, we consider empirical data from a single human participant. This participant completed a 150-trial IGT for which the experimental details are described in Part IV of this article.[3] The EV maximum likelihood of this participant was the highest among a total of 165 participants, and therefore this participant can be considered a relatively ideal specimen.

Figure A.3 shows the log likelihood contours for our ideal participant. Each panel shows the log likelihood values as a function of two EV parameters – the third parameter

[3]The participant under consideration here completed the "reward condition" of the experiment described later.

is fixed at its maximum likelihood estimate. The three right-hand panels are a zoomed-in version of the three left-hand panels. The three left-hand panels show that the log likelihood landscape is somewhat irregular, particularly for the bottom panel w vs. a landscape. Nevertheless, the right-hand panels suggest that this irregularity is less of a problem in the neighborhood of the maximum. For our ideal participant, the top right and bottom right landscapes indicate that small changes in the attention weight parameter w are accompanied by relatively large changes in the response consistency parameter c and the update parameter a, respectively. This makes c and a relatively difficult to identify. Note that the log likelihood contours depend on the parameters used to generate the data. Our parameter values (e.g., $w = 0.5$, $a = 0.35$, and $c = 0.35$) were informed by previous research and are fairly typical; nevertheless, it should be kept in mind that different parameter values may lead to different log likelihood contours.

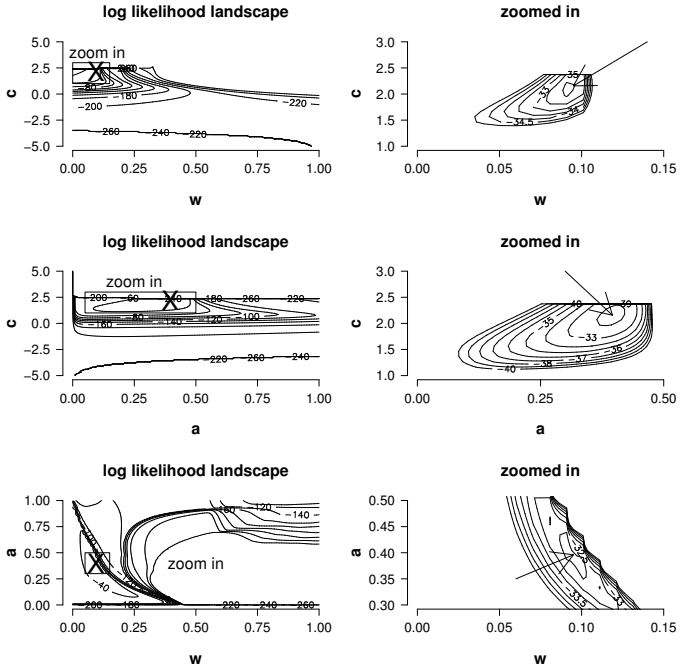


Figure A.3: Log likelihood contours for two EV parameters, with the third one fixed at its most likely value. The three right panels are a zoomed-in version of the left three panels. The arrows in the right panels point to the MLEs. Data come from an "ideal" human participant completing a 150-trial IGT (see text for details).

For the second log likelihood contour plot, we conducted a simulation with a synthetic participant who completed a 10,000 trial IGT. The parameter values in this simulation were the same as those used previously, that is, $w = 0.5$, $a = 0.35$, and $c = 0.35$. One would expect that with 10,000 trials, the log likelihood contours would be much better behaved.

Contrary to intuition, Figure A.4 shows that the shape of the log likelihood landscape again gives cause for concern, even when estimation is based on 10,000 trials from a simulated participant. Specifically, the elongated landscapes for w and a when plotted against c suggest that small changes in c can compensate for large changes in w and a. When c is fixed at its true value, the log likelihood landscape looks much better. Despite

these concerns about the log likelihood contours, it should be acknowledged that in the case of 10,000 trials, the parameters are recovered relatively accurately.

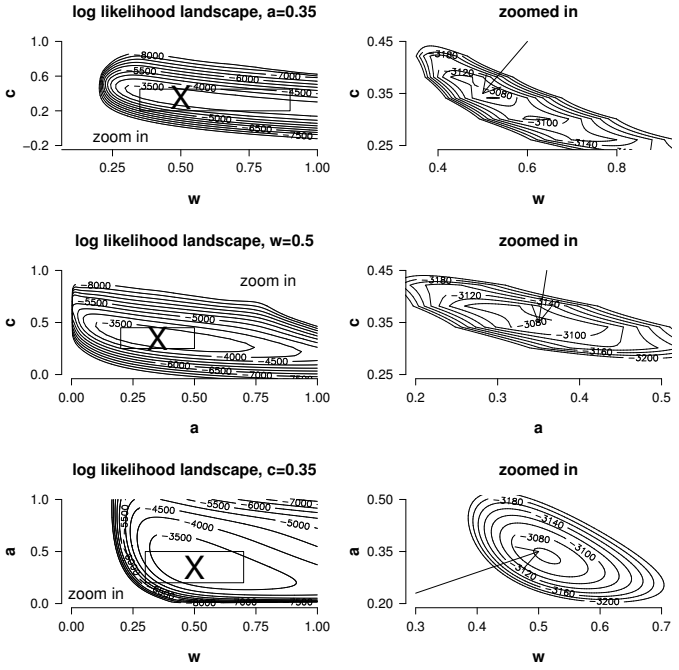


Figure A.4: Log likelihood contours for two EV parameters, with the third one fixed at its true value (i.e., $w = 0.5$, $a = 0.35$, and $c = 0.35$). The three right panels are a zoomed-in version of the left three panels. The arrows in the right panels point to the MLEs. Data come from a synthetic participant completing a 10,000-trial IGT.

The foregoing analyses have revealed that the EV parameter estimation routine is not immune to problems. In particular, the large variability that characterizes the parameter estimation for individual participants means that (1) it is valuable to have access to and use the uncertainty that accompanies parameter estimation for individual participants; and (2) it is necessary to combine information across different participants. One of the most principled ways to accomplish these goals is to turn to Bayesian inference.

A.3 Part III Bayesian Estimation

In Bayesian estimation (e.g., Bernardo & Smith, 1994; D. V. Lindley, 2000), all uncertainty about parameters is quantified by probability distributions. Prior parameter distributions are updated by incoming data to yield posterior distributions. These posterior distributions quantify our uncertainty about the parameters after having seen the data (for introductions to Bayesian inference for psychologists see for instance Edwards et al., 1963; Lee & Wagenmakers, 2005, and Rouder & Lu, 2005).

The Bayesian approach holds many advantages over the orthodox maximum likelihood approach (for a review see Wagenmakers, Lee, et al., 2008). One of the more general advantages is that the axiomatic foundations of the Bayesian approach guarantee that it is *coherent*; in the statistical sense of the word, this means that information from different

sources is combined in a principled manner such that inferential statements cannot be internally inconsistent.

Other prime advantages of the Bayesian approach include flexibility, generality, and practicality. For instance, Bayesian nonlinear models are easily equipped with hierarchical extensions. Indeed, some researchers profess to adopt the Bayesian approach for its practical advantages alone (e.g., Rouder & Lu, 2005, p. 599).

In the context of the EV model, a concrete advantage of the Bayesian procedure is that it yields posterior distributions for w, a, and c. These posterior distributions directly convey the uncertainty associated with individual parameter estimates. Below, we first introduce the Bayesian EV model for inference on the level of a single participant, and then add a hierarchical structure that allows information from different participants to be combined in coherent fashion.

The Bayesian Graphical EV Model for a Single Participant Analysis

It is often insightful to represent Bayesian models graphically, as this notation highlights the model structure, the dependence between the models parameters, and the way in which the likelihood can be factorized to reduce computational effort (for introductions to graphical models, see for instance Gilks et al., 1994; Griffiths, Kemp, & Tenenbaum, 2008; Lauritzen, 1996; Lee, 2008; Spiegelhalter, 1998).

The Bayesian graphical EV model for a single participant analysis is shown in Figure A.5. In this notation, nodes represent variables of interest, and the graph structure is used to indicate dependencies between the variables, with children depending on their parents. The double borders indicate that the variables under consideration are deterministic (i.e., they are calculated without noise from other variables) rather than stochastic. Continuous variables are represented with circular nodes and discrete variables are represented with square nodes; observed variables are shaded and unobserved variables are not shaded. In Figure A.5, for instance, the observed variable W_{t-1} indicates the rewards obtained by the participant on trial $t-1$. We also use plate notation, enclosing with square boundaries subsets of the graph that have independent replications in the model. The plate of Figure A.5 reads $t = 1, ...150$ and this corresponds to the 150 choices in the IGT.

Figure A.5 shows that the psychological processes associated with parameters w, a, and c are unobserved (i.e., the nodes are unshaded) and continuous (i.e., the nodes are circular). The quantities v_{t-1}, Ev_t, θ_{t-1}, and $\Pr[S_t]$ are deterministic (i.e., the nodes have double borders), as these quantities are calculated without noise from Equations A.1, A.2, A.4, and A.3, respectively. To avoid crowding the figure, we have suppressed the notation that indexes the deck number k.

In order to get off the ground, the Bayesian inference machine needs prior distributions for its parameters. For the EV model, we choose noninformative priors, that is, priors that are uniform across their range. For ease for application, we initially programmed this model in the WinBUGS environment (Spiegelhalter, Thomas, Best, & Lunn, 2003) that has been developed to approximate distributions by sampling values from them using Markov chain Monte Carlo techniques. The acronym BUGS stands for Bayesian inference Using Gibbs Sampling (Casella & George, 1992), and it greatly facilitates Bayesian modeling and communication (for a review see Cowles, 2004).[4]

[4]At the time of writing, WinBUGS is freely available at `http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml`.

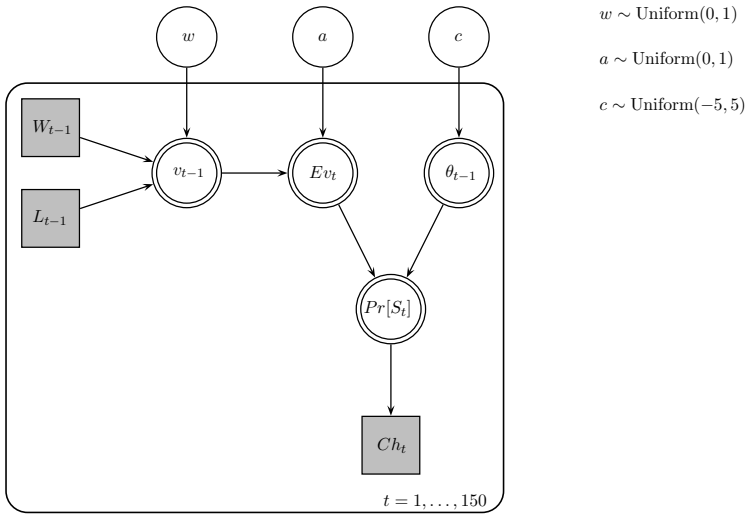


Figure A.5: Bayesian graphical EV model for a single participant analysis.

The direct implementation of the EV model in WinBUGS is relatively straightforward, but the program takes about five minutes to obtain a reliable estimation of the parameters for a single participant, and occasionally crashes. When the EV model is hand-coded as a WinBUGS function with the help of the WinBUGS Development Interface (WBDev, D. Lunn, 2003), the program no longer crashes and runtime decreases to about 8 seconds for a single participant.

Illustrative Results for a Single Synthetic Participant

We illustrate the Bayesian Markov chain Monte Carlo (MCMC) parameter estimation routine for the EV model by applying the method to data from a synthetic participant in a 150-trial IGT. As in our previous simulations, the true parameter values were $w = 0.5$, $a = 0.35$, and $c = 0.35$. Figure A.6 shows the result.

The top panels of Figure A.6 show that the medians of the posterior distributions are relatively close to the true generating values for the parameters. More importantly, the posterior distributions directly indicate the uncertainty about the parameters. For instance, one only needs to glance at the top panels to learn that the attention weight parameter w is likely to lie somewhere in between 0.25 and 0.75, that the updating rate parameter a lies somewhere in between 0.20 and 0.75, and that the response consistency parameter c lies somewhere in between -0.5 and 0.5.

The bottom panels of Figure A.6 show the MCMC chains that form the basis for the posterior distributions in the top panels. Visual inspection suggests that these chains are relatively well-behaved, in the sense that appear to be draws from the stationary distribution.

In addition to plotting the posterior distributions for the three parameters separately, the MCMC samples can also be used to plot joint posterior distributions. The joint distributions provide useful information with respect to how the parameters for a single participant relate to each other. Figure A.7 plots the MCMC values from joint distributions for three parameter pairs. The results show that there is a substantial negative

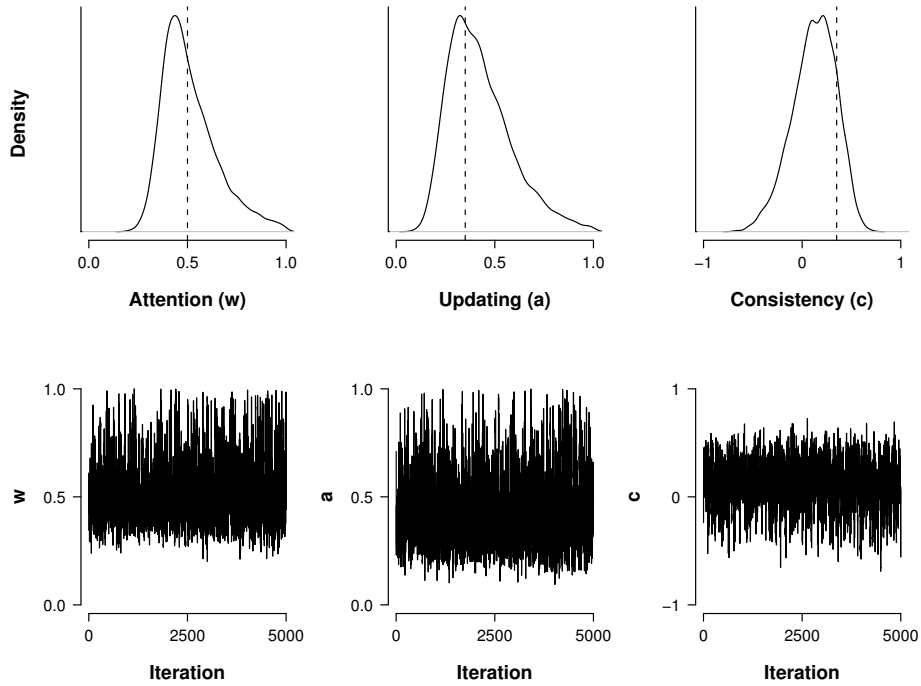


Figure A.6: Density estimates for posterior distributions (top row) and MCMC chains (bottom row) for the three EV parameters based on data from a single synthetic participant completing a 150-trial IGT. The dotted lines in the top panels indicate the true parameter values (i.e., $w = 0.5$, $a = 0.35$, and $c = 0.35$).

correlation between the c parameter and the w and a parameters. This correlational pattern echoes the earlier result based on the MLEs for 1,000 synthetic participants (see Figure A.2).

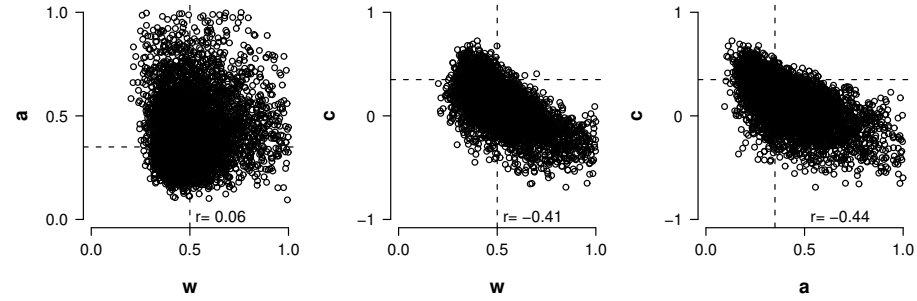


Figure A.7: Joint posterior distributions for EV parameter pairs, based on MCMC samples from a Bayesian analysis of a single synthetic participant completing a 150-trial IGT. The dotted lines indicate the true parameter values (i.e., $w = 0.5$, $a = 0.35$, and $c = 0.35$).

Illustrative Results for a Single Human Participant

Here we illustrate the Bayesian parameter estimation routine by application to the data from the same human participant whose data were also analyzed by maximum likelihood (cf. Figure A.3). The top panels of Figure A.8 show that the medians of the poste-

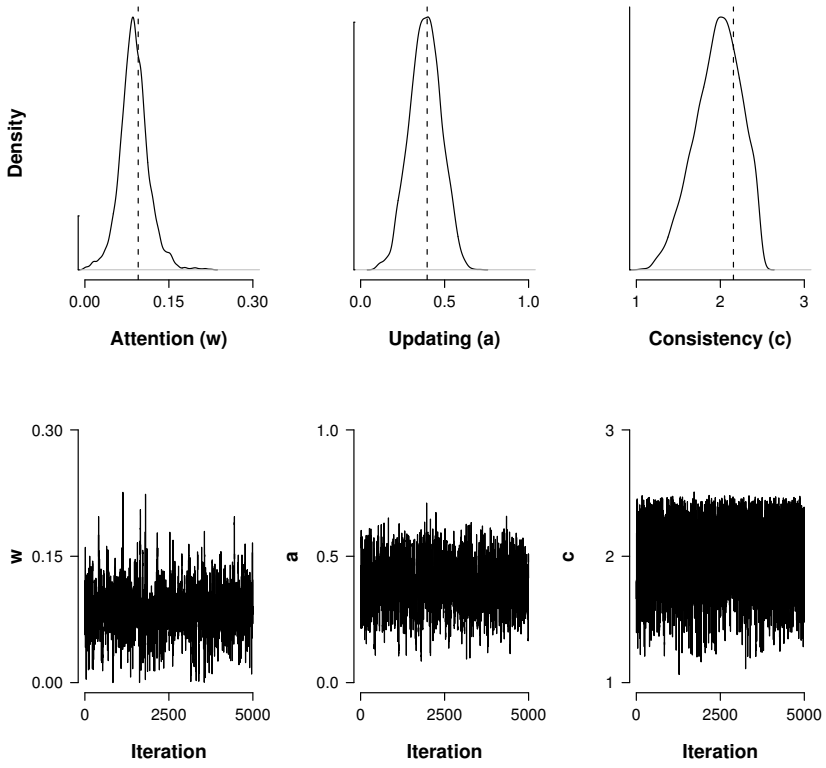


Figure A.8: Density estimates for posterior distributions (top row) and MCMC chains (bottom row) for the three EV parameters based on data from an "ideal" human participant completing a 150-trial IGT. The dotted lines in the top panels indicate the MLE parameter values (i.e., $\hat{w} = 0.10$, $\hat{a} = 0.40$, and $\hat{c} = 2.17$).

rior distributions are very close to the MLE estimates. These panels also show that uncertainty about w is relatively small, whereas uncertainty about a and c remains substantial. Visual inspection of the chains, plotted in the bottom three panels, strongly suggests convergence to the stationary distribution.

Figure A.9 shows MCMC samples from the joint posterior distributions for our ideal human participant. The left-hand and middle panels show that small changes in the attention weight parameter w are associated with relatively large changes in the update parameter a and the response consistency parameter c, respectively. This echoes the results from the earlier analysis of the log likelihood landscapes in Figure A.3.

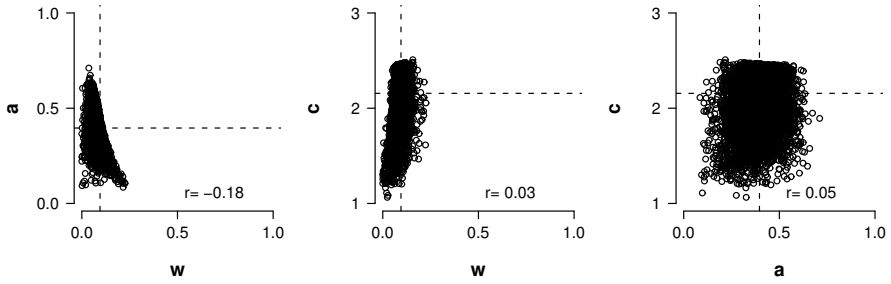


Figure A.9: Joint posterior distributions for EV parameter pairs, based on MCMC samples from a Bayesian analysis of an "ideal" human participant completing a 150-trial IGT. The dotted lines indicate the MLE parameter values (i.e., $\hat{w} = 0.10$, $\hat{a} = 0.40$, and $\hat{c} = 2.17$).

The Bayesian Graphical EV Model for a Hierarchical Analysis

Historically, the field of experimental psychology has mostly ignored individual differences, pretending instead that each new participant is a replicate of the previous one (Batchelder, 2007). As Bill Estes and others have shown, however, individual differences that are ignored can lead to averaging artifacts in which the data that are averaged over participants are no longer representative for any of the participants (Estes, 1956, 2002; Heathcote, Brown, & Mewhort, 2000). One way to address this issue, popular in psychophysics, is to measure each individual participant extensively, and deal with the data on a participant-by-participant basis.

In between the two extremes of assuming that participants are completely the same and that they are completely different lies the compromise of hierarchical modeling (see also Lee & Webb, 2005). The theoretical advantages and practical relevance of a Bayesian hierarchical analysis for common experimental designs have been repeatedly demonstrated by Jeff Rouder and others (Morey, Pratte, & Rouder, 2008; Morey, Rouder, & Speckman, 2008; Navarro, Griffiths, Steyvers, & Lee, 2006; Rouder, Lu, Speckman, Sun, & Jiang, 2005; Rouder & Lu, 2005; Rouder et al., 2007, 2008). Although hierarchical analyses can be carried out using orthodox methodology (i.e., Hoffman & Rovine, 2007), there are convincing philosophical and practical reasons to prefer the Bayesian methodology (e.g., D. V. Lindley, 2000 and Gelman & Hill, 2007, respectively).

In Bayesian hierarchical models, parameters for individual people are assumed to be drawn from a group-level distribution. Such multi-level structures naturally incorporate both the differences and the commonalities between people, and therefore provide experimental psychology with the means to settle the age-old problem of how to deal with *individual differences*.

The flexibility of the Bayesian paradigm makes it straightforward to extend the single participant model from Figure A.5 in a hierarchical fashion. As Figure A.10 shows, the hierarchical model differs from the individual model in that it adds a plate to indicate independent replications for $i = 1, ..., N$ participants. In addition, the hierarchical model transforms c to lie between 0 and 1 (instead of between -5 and $+5$), so that all EV parameters are now on a rate scale (this transformation is not shown in the figure).

In the graphical model notation of Figure A.10, all three parameters w_i, a_i, and c_i are deterministic; this is because instead of modeling w_i, a_i, and c_i directly, we instead model their respective probit transformations ν_i, α_i, and γ_i. The probit transform is the inverse cumulative distribution function of the normal distribution, so that, for instance,

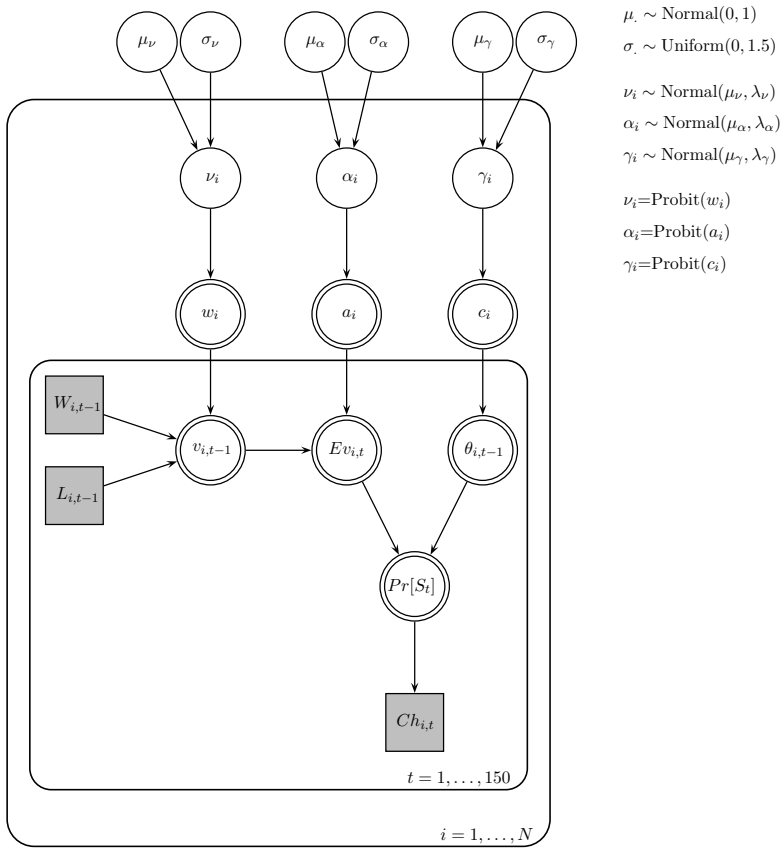


Figure A.10: Bayesian graphical EV model for a hierarchical analysis.

a rate of $\alpha_i = 0.5$ maps onto a probit value of $\nu_i = 0$. The probit scale covers the entire real line, and a standard normal distribution on the probit scale corresponds to a uniform distribution on the rate scale (Rouder & Lu, 2005, p. 588). We assume that for a group of participants, the individual probit rates ν_i, α_i, and γ_i are drawn from group-level normal distributions with respective normal means μ_ν, μ_α, and μ_γ and respective normal standard deviations σ_ν, σ_α, and σ_γ.

The specification of the model requires prior distributions for the normal means and standard deviations of the group-level distributions. We used standard normal priors on $\mu_{(\cdot)}$, that is, $\mu_{(\cdot)} \sim N(0, 1)$ and a uniform prior from 0 to 1.5 on the standard deviations $\sigma_{(\cdot)}$, that is, $\sigma_{(\cdot)} \sim U(0, 1.5)$. The upper limit of 1.5 was determined by the following line of reasoning (see also Lodewyckx et al., 2011). When, say, $\mu_\alpha = 0$ and $\sigma_\alpha = 1$, then α_i comes from a standard normal distribution on the probit scale and a_i comes from a uniform distribution on the rate scale. Increasing the value of σ_α results in a bimodal distribution for a_i, which we deem unrealistic. As μ_α increases, so does the maximum value of σ_α that results in a just-unimodal distribution for a_i. When we assign μ_α an extreme value of 2.3 (i.e., this translates to an average a value of .99) a value of $\sigma_\alpha \approx 1.5$ is the maximum value that still guarantees a unimodal distribution on the rate scale.

A.4 Part IV Application to Experimental Data

In this section we apply the Bayesian hierarchical model as shown in Figure A.10 to a validation experiment with 165 participants. The primary goal of the experiment was to carry out a *test of specific influence* for the EV model. This means that, next to the standard condition, we included three experimental conditions, each of which designed to affect selectively one of the EV parameters w, a, or c. If the parameters of the EV model indeed correspond to the psychological processes that they are assumed to be associated with, then an experimental manipulation of "attention weight" should affect only the estimate of w, an experimental manipulation of "updating rate" should affect only the estimate of a, and an experimental manipulation of "response consistency" should affect only the estimate of c.

Method

Participants

A total of 165 students from the University of Amsterdam participated for course credit.

Stimulus materials and design

The experiment featured four conditions. In the first "standard" condition, 41 participants completed a 150-trial IGT under the usual instructions. In the second "rewards" condition, 42 participants completed a 150-trial IGT under the instruction to pay particular attention to the rewards and think of the losses as being less important. This instruction was strengthened by displaying the rewards more prominently on the screen than the losses. We expected this manipulation to decrease w and leave a and c unaffected.

In the third "updating" condition, 41 participants completed a 150-trial IGT under the usual instruction. However, in the updating condition each card selection was followed by the on-screen presentation of a sequence of five numbers; participants were required to remember this sequence, as after the next card selection they were asked about the relative position of one of the numbers (Hinson, Jameson, & Whitney, 2002). For example, presentation of the number sequence $\{1, 5, 3, 4, 2\}$ (i.e., all numbers are integers ranging from 1 to 5, drawn randomly without replacement) could be followed one card selection later by the request to "enter the number that was in the third place". We expected this manipulation to increase a and leave w and c unaffected.

In the fourth "consistency" condition, 41 participants completed a 150-trial IGT under the usual instruction. However, in the consistency condition participants were told after every 10 trials that the payoff schemes for the decks could have changed (i.e., "Beware, the rewards for each deck may have changed"). We expected this manipulation to decrease c and leave w and a unaffected.

In all four conditions, we used a computerized version of the IGT where the four cards were displayed on the screen and the participants indicated their card selection by a mouse click. In all conditions of the experiment, we used the standard IGT payoff scheme shown in Table A.1. After each card selection, the associated rewards and losses were displayed on the screen for 2 seconds. Before the start of the next selection opportunity, the mouse was re-positioned at the center of the screen.

Procedure

Participants were randomly assigned to one of the four conditions. Task instructions were presented on the screen prior to the start of the experiment. Participants were allowed to start the IGT after verbally confirming that they had understood the instructions. The experiment took less than 30 minutes to complete.

Results

Card selection

Figure A.11 shows the proportion of selected decks as a function of trial number in each of the four conditions. It is clear that our experimental manipulations affected participant's choice performance. In particular, only in the standard condition did participants learn to prefer the good deck C over the bad deck B.

Although the extent of learning in the standard condition may seem relatively modest, the IGT is a surprisingly difficult task to grasp, as is evident from a study by Caroselli et al. (2006) who found that university students often tend to prefer the bad decks.

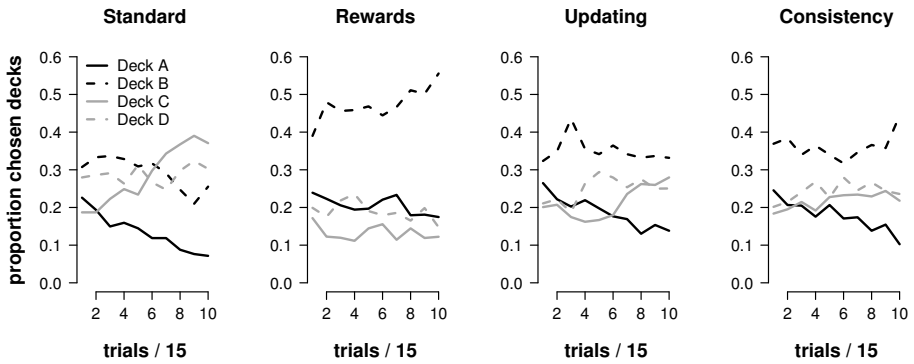


Figure A.11: The proportion of chosen decks as a function of trial number in each of the four conditions of the validation experiment. Consistent with IGT nomenclature, deck A is disadvantageous and has high-frequent loss; deck B is disadvantageous and has low-frequent loss; deck C is advantageous and has high-frequent loss; and deck D is advantageous and has low-frequent loss.

In the reward condition, participants have a strong preference for the bad deck B, a deck with relatively high rewards and an occasional large loss. The behavior is in line with the instruction to pay more attention to rewards than to losses.

In the updating condition and the consistency conditions, the participants consistently express a preference for the bad deck B, although this preference is less pronounced than in the rewards condition. In conclusion, our experimental manipulations were effective on the level of choice performance.

EV parameters: Maximum likelihood estimation

In the usual group analysis for the EV model, individual maximum likelihood estimates are averaged to produce a group estimate. Inference is then based on the group mean and its variance. For comparison purposes, we follow the same procedure here. The result

of our analysis is shown in Figure A.12, which plots the mean MLEs for the three EV parameters in each of the four different experimental conditions.

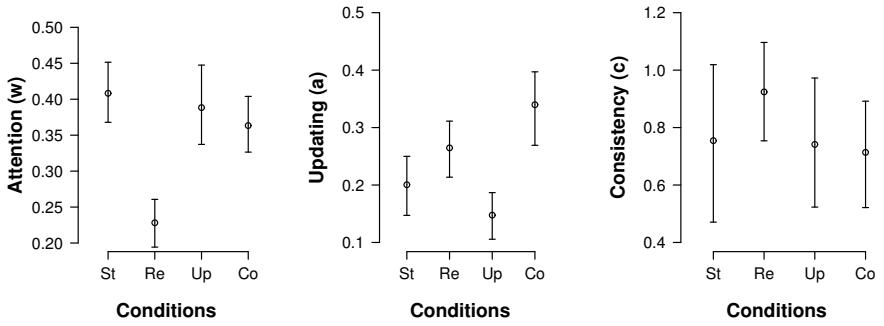


Figure A.12: Mean maximum likelihood estimates for the three EV parameters in the four experimental conditions. Error bars indicate one bootstrap standard error of the mean.

The left panel of Figure A.12 shows that, as expected, the w parameter is lower in the rewards condition than in the other three conditions, and that the w parameter does not differ between the standard condition, the updating condition, and the consistency condition. This result suggests that the w parameter is indeed uniquely associated to the attention for losses versus rewards, just as the EV model proposes.

Unfortunately, the results of the other conditions are much less clear. The middle panel and the right panel of Figure A.12 indicate that there is no reliable experimental effect on the EV parameters a and c, respectively. It may of course be argued that our experimental manipulations for a and c were too weak to produce an effect; however, the distinct patterns of choice performance for the standard condition versus the updating and consistency conditions suggests otherwise (cf. Figure A.11). This issue is presently unresolved, and more research is needed to address it.

EV parameters: Bayesian hierarchical estimation

We applied the Bayesian hierarchical EV model separately to each of the four experimental conditions. The focus of interest is on the means of the group distributions: in Figure A.10, these are indicated as μ_ν, μ_α, and μ_γ. In order to facilitate comparison with the mean MLE method, the posterior distributions for these parameters were transformed back from the probit scale to the rate scale.

Note that in the present work, we concentrate on parameter estimation rather than on model selection or hypothesis testing; this means that here we do not consider equality constraints on the model parameters across experimental conditions, such that one could formally test whether, say, μ_ν is the same or different in the four experimental conditions. The extension to model selection in Bayesian hierarchical models can be accomplished by transdimensional MCMC (e.g., Carlin & Chib, 1995; Green, 1995; Sinharay & Stern, 2005; Sisson, 2005); applications in the field of psychology are discussed in Lodewyckx et al. (2011).

Considering again the problem of parameter estimation, Figure A.13 shows that the Bayesian hierarchical estimation method and the mean MLE method yield different results. In particular, the middle panel shows that the Bayesian estimates for a are sys-

tematically lower then the mean MLEs, and the right panel shows that the Bayesian estimates for c are systematically higher than the mean MLEs.

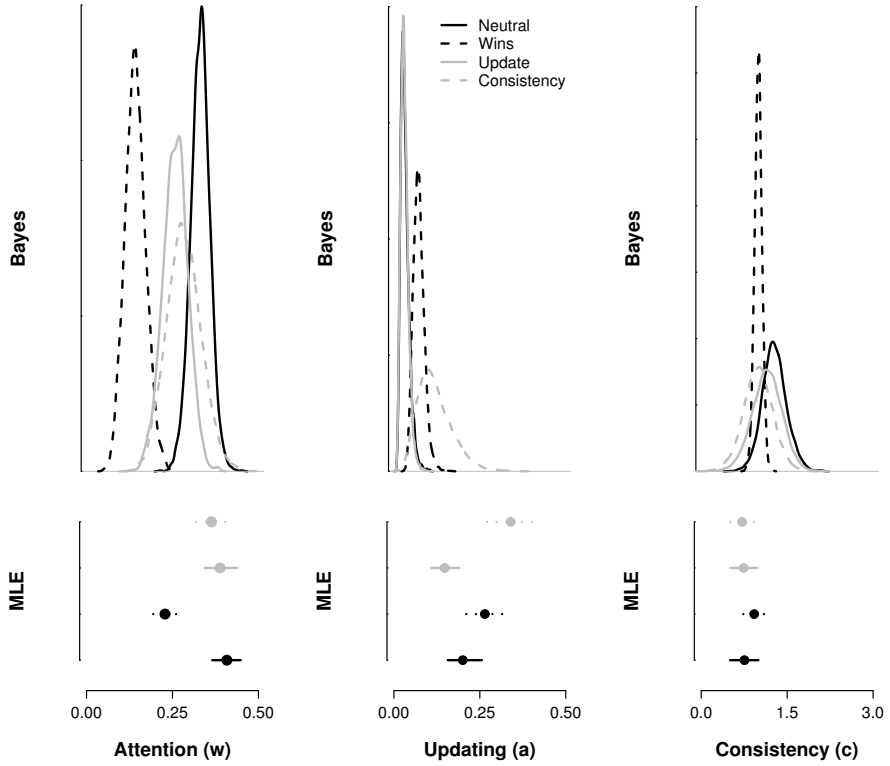


Figure A.13: Posterior distributions for the group mean of the three EV parameters in the four experimental conditions (top) compared to mean maximum likelihood estimates (bottom). For the mean maximum likelihood estimates, the horizontal error bars indicate one bootstrap standard error of the mean.

The discrepancy between the Bayesian hierarchical estimates and those provided by the mean MLE method motivates a closer inspection of the data. This inspection revealed two potential sources of contamination. The first source is that for several participants, the MLE of at least one of the parameters was estimated on the boundary of the parameter space. The situation is summarized in the first two columns of Table A.2.

When parameter point estimates are located on the boundary of the parameter space, this often signals a problem with the estimation procedure, the data, or the interaction between the data and the model. Note that the same phenomenon was observed for the parameter recovery simulations reported in Figures A.1 and A.2. We removed the first source of contamination by eliminating from the analyses all data sets for which one or more of the maximum likelihood point estimates were located on the boundary of the parameter space. The analyses for the filtered data are shown in Figure A.14, from which it is evident that results from the MLE method and the Bayesian hierarchical method are now more similar than they were for the contaminated data. In particular, the mean MLEs for a have shifted downward, and the mean MLEs for c have shifted upward. The results from the Bayesian hierarchical analysis appear to be more robust to the removal of the extreme estimates than are those from the mean MLE method.

Condition	Participant total	After removal of boundary estimates	After additional removal of cases for which BL>EV
Standard	41	30	19
Rewards	42	31	20
Updating	41	25	19
Consistency	41	27	16

Table A.2: Data Filtering for the Validation Experiment. Note. BL>EV refers to the situation in which the baseline model outperforms the EV model. See text for details.

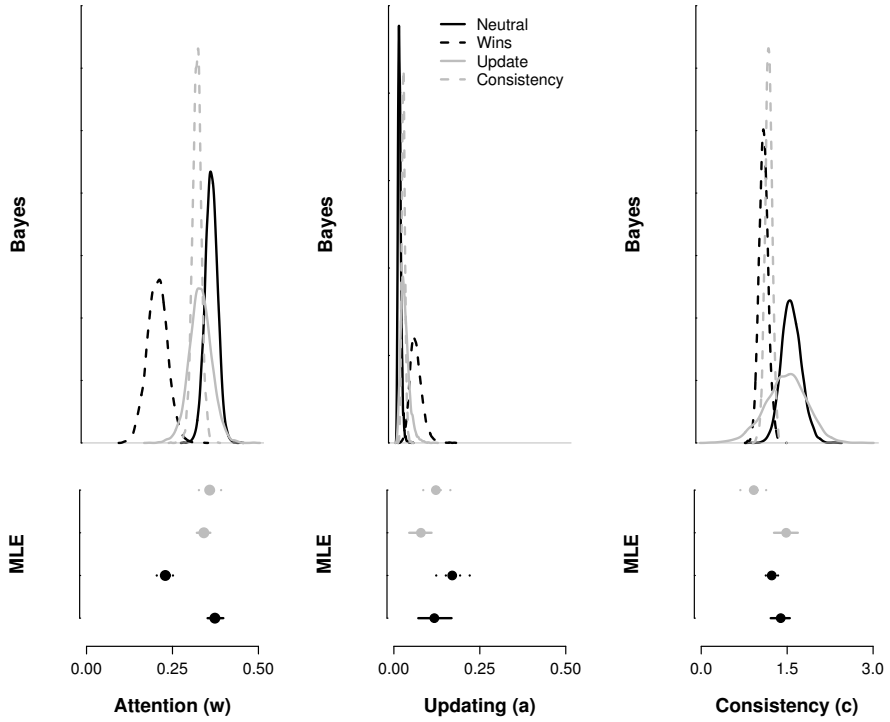


Figure A.14: Posterior distributions for the group mean of the three EV parameters in the four experimental conditions (top) compared to mean maximum likelihood estimates (bottom), after removal of participants for which at least one of the maximum likelihood point estimates was on the boundary of the parameter space. For the mean maximum likelihood estimates, the horizontal error bars indicate one bootstrap standard error of the mean.

The second source of potential contamination in the data is that a subset of participants may, for lack of effort or lack of insight, not have understood the dynamics of the IGT. In order to identify that subset, we followed Busemeyer and Stout (2002) and compared performance of the EV model to that of a *baseline model*. The baseline model is a statistical model that assumes that choices are independently and identically distributed over trials – it incorporates the frequencies with which the decks are selected, but does not incorporate any effects of learning. For example, when a participant has selected a

card from deck B in 30% of the cases, the baseline model assumes that the probabilistic forecast of the baseline model for deck B is a constant 0.3 throughout the task.

The final columns of Table A.2 shows the numbers of participants that remain once we remove participants for whom the baseline model provided a better fit than the EV model. Table A.2 shows that the two sources of contamination (i.e., parameters on the boundary and relatively poor fits of the EV model) each account for approximately 25% of participants. Figure A.15 shows that when we apply the two estimation procedures to the remaining 50% of the participants, the result of the Bayesian hierarchical estimation are again somewhat more robust than those of the mean MLE method.

It should be acknowledged that both estimation procedures lead to the same inference with respect to the effect of the experimental manipulations: a successful specific influence on the attention weight w, but no noticeable effect on updating rate a and response consistency c. Nevertheless, in other cases the inference from the Bayesian hierarchical model may differ from that of the mean MLE method. In such situations, we feel the former method is superior: it coherently combines information from different participants, summarizes uncertainty through probability distributions, and appears to be relatively robust to contamination of the data.

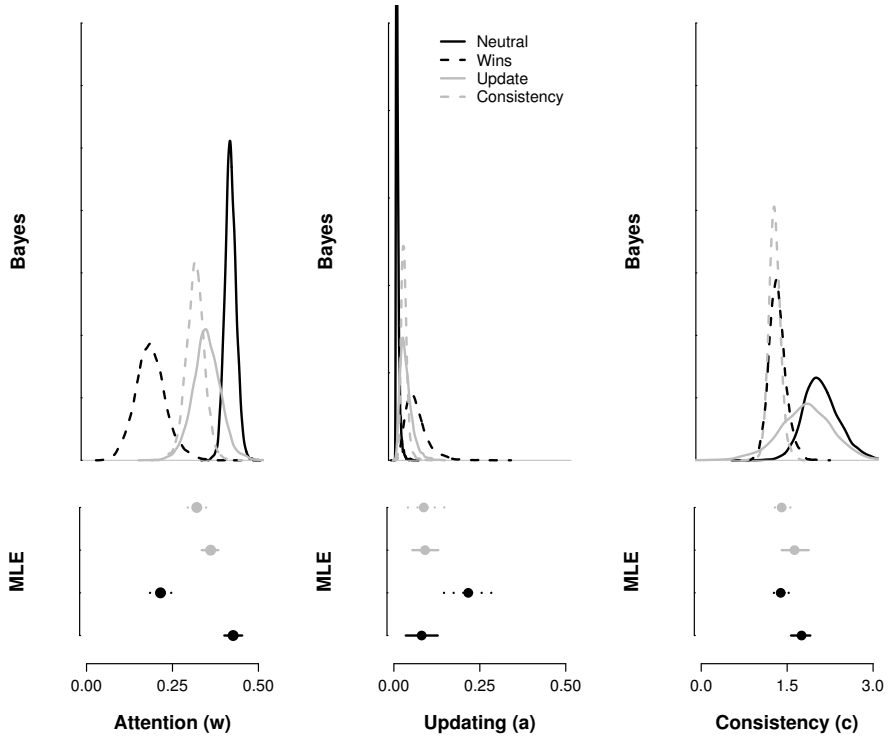


Figure A.15: Posterior distributions for the group mean of the three EV parameters in the four experimental conditions (top) compared to mean maximum likelihood estimates (bottom), after removal of (1) participants for which at least one of the maximum likelihood point estimates was on the boundary of the parameter space; and (2) participants for which the baseline model outperformed the EV model. For the mean maximum likelihood estimates, the horizontal error bars indicate one bootstrap standard error of the mean.

A.5 General Discussion

In an attempt to bridge the separate disciplines of clinical psychology and mathematical psychology, the EV model uses maximum likelihood estimation to decompose choice performance in the Iowa Gambling Task into three underlying psychological processes: the attention to losses versus rewards, the rate with which new information updates old expectancies, and the extend to which people make decisions that are consistent with their internal evaluations. The EV model has a proven track record and can be presently considered the default quantitative model for the Iowa Gambling Task.

In this article, we focused on the method of parameter estimation for the EV model. In particular, we showed that for single participants it is generally not possible to estimate the EV parameters precisely. Therefore, one should be wary of applying the EV model to the clinical diagnosis of decision making deficits on the level of single patients.

When the EV model is applied on the group level, such as when researchers compare model parameters for a group of cocaine addicts versus those for a group of normal controls, we recommend the use of the Bayesian hierarchical model. The Bayesian approach is not only more principled than the standard mean maximum likelihood approach, but the Bayesian procedure is also more robust in the face of contamination. Regardless of the estimation procedure that is used, we recommend that parameters that are on the boundary of parameter space be removed prior to the analysis.

The Bayesian hierarchical model proposed here can be applied not just to the EV model for the IGT, but much more broadly to a whole range of reinforcement learning tasks (e.g., Sutton & Barto, 1998). It is likely that tasks other than the IGT can provide a more efficient means of estimation the psychological processes of interest. For instance, it is possible that parameters are estimated more precisely when the IGT is altered to reveal foregone payoffs, that is, when the participant sees not only the result of the actual choice, but also sees the foregone payoffs from unchosen decks. The Bayesian model developed here could be used to explore a range of different task formats in order to select a format that allows researchers to extract a relatively large amount of information from a participant's choice performance.

The Expectancy Valence model for the Iowa Gambling Task has greatly facilitated the communication between the separate disciplines of clinical psychology and mathematical psychology. We hope that by taking individual differences and similarities into account in a coherent fashion, by quantifying uncertainty of parameter estimation in terms of probability distributions, and by providing the opportunity to discover new tasks with high information gain, the Bayesian hierarchical paradigm can increase this level of communication even further.

B Bayesian Inference Using WBDev: A Tutorial for Social Scientists

Abstract

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). This tutorial 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.

An excerpt of this chapter has been published as:
Wetzels, R., Lee, M.D., & Wagenmakers, E.-J., (2010). Bayesian Inference Using WBDev: A Tutorial for Social Scientists. *Behavior Research Methods, 42*, 884–897.

B.1 Introduction

Psychologists who seek quantitative models for their data face formidable challenges. Not only are data often noisy and scarce, but they may also have a hierarchical structure, they may be partly missing, they may have been obtained under an ill-defined sampling plan, and they may be contaminated by a process that is not of interest. In addition, the models under consideration may have multiple restrictions on the parameter space, especially when there is useful prior information about the subject matter at hand.

In order to address these kinds of real-world challenges, the psychological sciences have started to use Bayesian models for the analysis of their data (e.g., Lee, 2008; Rouder & Lu, 2005; Hoijtink et al., 2008). In Bayesian models, existing knowledge is quantified by *prior* probability distributions and updated upon consideration of new data to yield *posterior* probability distributions. Modern approaches to Bayesian inference include Markov chain Monte Carlo sampling (MCMC; e.g., Gamerman & Lopes, 2006; Gilks, Richardson, & Spiegelhalter, 1996), a procedure that makes it easy for researchers to construct probabilistic models that respect the complexities in the data, allowing almost any probabilistic model to be evaluated against data.

One of the most influential software packages for MCMC-based Bayesian inference is known as WinBUGS (BUGS stands for Bayesian inference Using Gibbs Sampling; Cowles, 2004; Sheu & O'Curry, 1998; D. J. Lunn et al., 2000; D. Lunn, Spiegelhalter, Thomas, & Best, 2009). WinBUGS comes equipped with an array of predefined functions (e.g., `sqrt` for square root and `sin` for sine) and distributions (e.g., the Binomial and the Normal) that allow users to combine these elementary building blocks into complex probabilistic models.

For some psychological modeling applications, however, it is highly desirable to define one's own functions and distributions. In particular, user-defined functions and distributions greatly facilitate the use of psychological process models such as ALCOVE (J. K. Kruschke, 1992), the Expectancy-Valence model for decision-making (Busemeyer & Stout, 2002), the SIMPLE model of memory (Brown, Neath, & Chater, 2007), or the Ratcliff diffusion model of response times (Ratcliff, 1978).

The ability to implement these user-defined functions and distributions can be achieved through the use of the WinBUGS Development Interface (WBDev; D. Lunn, 2003), an add-on program that allows the user to hand-code functions and distributions in the programming language Component Pascal.[1] To that end, we need BlackBox, a development environment for programs such as WinBUGS, that are written in Component Pascal.

The use of WBDev brings several advantages. For instance, complicated WBDev components lead to faster computation than their counterparts programmed in straight WinBUGS code. Moreover, some models will only work properly when implemented in WBDev. Another advantage of WBDev is that it compartmentalizes the code, resulting in scripts that are easier to understand, communicate, adjust, and debug. A final advantage of WBDev is that it allows the user to program functions and distributions that are simply not available in WinBUGS, but may be central components of psychological models (Donkin, Averell, Brown, & Heathcote, 2009; Vandekerckhove, Tuerlinckx, & Lee, 2011).

This tutorial aims to stimulate psychologists to use WBDev by providing four thoroughly documented examples; for both functions and distributions, we provide a simple and a more complex example. All examples are relevant to psychological research.[2]

[1] More information can be found at: `http://en.wikipedia.org/wiki/Component_Pascal`.

[2] There already exist two concise tutorials on how to write functions and distributions in WBDev,

Our tutorial is geared towards researchers who have experience with computer programming and WinBUGS. A gentle introduction to the WinBUGS program is provided by Ntzoufras (2009) and Lee and Wagenmakers (2009). Despite these prerequisites we have tried to keep our tutorial accessible for social scientists in general.

We start our tutorial by discussing the WBDev implementation of a simple *function* that involves the addition of variables. We then turn to the implementation of a complicated function that involves the Expectancy-Valence model (Busemeyer & Stout, 2002; Wetzels, Vandekerckhove, et al., in press). Next, we discuss the WBDev implementation of a simple *distribution*, first focusing on the Binomial distribution, and then turning to the implementation of a more complicated distribution that involves the shifted Wald distribution (Heathcote, 2004; W. Schwarz, 2001). For all of these examples, we explain the crucial parts of the WBDev scripts and the WinBUGS code. The thoroughly commented code is available online at `www.ruudwetzels.com`. For each example, our explanation of the WBDev code is followed by application to data and the graphical analysis of the output.

B.2 Installing WBDev (BlackBox)

Before we can begin hard-coding our own functions and distributions we need to download and install three programs: WinBUGS, WBDev and BlackBox.[3] To make sure all programs function properly, they have to be installed in the order given below.

1. Install WinBUGS

WinBUGS is the core program that we will use. Download the latest version from `http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml` (WinBUGS14.exe). Install the program in the default directory `./Program Files/WinBUGS14`.[4] Make sure to register the software by obtaining the registration key and following the instructions—WinBUGS will not work without it.

2. Install WinBUGS Development Interface (WBDev)

Download WBDev from `http://www.winbugsdevelopment.org.uk/` (WBDev.exe). Unzip the executable in the WinBUGS directory `./Program Files/WinBUGS14`. Then start WinBUGS, open the"wbdev01_09_04.txt" file and follow the instructions at the top of the file. During the process, WBDev will create its own directory `/WinBUGS14/WBDev`.

3. Install BlackBox Component Builder

BlackBox can be downloaded from `http://www.oberon.ch/blackbox.html`. At the time of writing, the latest version is 1.5. Install BlackBox in the default directory: `./Program Files/BlackBox Component Builder 1.5`. Go to the WinBUGS directory and select all files (press "Ctrl+A") and copy them (press "Ctrl+C"). Next, open the BlackBox directory and paste the copied files in there (press "Ctrl+V"). Select " `Yes to all`" if asked about replacing files. Once this is done, we will be able to open BlackBox and run

written by David Lunn and Chris Jackson. The examples in those tutorials require advanced programming skills and they are not directly relevant for psychologists.

[3] At the time of writing, all programs are available without charge. Note that these programs only work under the Microsoft Windows operating system.

[4] On the Windows Vista operating system, install the program in the directory "c:/WinBUGS14".

WinBUGS from inside BlackBox. This completes installation of the software, and we can start to write our own functions and distributions.

B.3 Functions

The mathematical concept of a function expresses a dependence between variables. The basic idea is that some variables are given (the input) and with them, other variables are calculated (the output). Sometimes, complex models require many arithmetic operations to be performed on the data. Because such calculations can become computationally demanding using straight WinBUGS code, it can be convenient to use WBDev and implement these calculations as a function. The first part of this section will explain a problem without using WBDev. We then show how to use WBDev to program a simple and a more complex function.

Example 1: A Rate Problem

A binary process has two possible outcomes. It might be that something either happens or does not happen, or that something either succeeds or fails, or takes one value rather than the other. An inference that often is important for these sorts of processes concerns the underlying rate at which the process takes one value rather than the other. Inferences about the rate can be made by observing how many times the process takes each value over a number of trials.

Suppose that someone plays a simple card game and can either win or lose. We are interested in the probability that the player wins a game. To study this problem, we formalize it by assuming the player plays n games and wins k of them. These are known, or observed, data. The unknown variable of interest is the probability θ that the player wins any one specific game. Assuming the games are statistically independent (i.e., that what happened on one game does not influence the others, so that the probability of winning is the same for all of the games), the number of wins k follows a Binomial distribution, which is written as

$$k \sim \text{Binomial}\big(\theta, n\big), \tag{B.1}$$

and can be read "the success count k out of a total of n trials is Binomially distributed with success rate θ". In this example, we will assume a success count of 9 ($k = 9$) and a trial total of 10 ($n = 10$).

A rate problem: The model file

A so-called model file is used to implement the model into WinBUGS. The model file for inferring θ from an observed n and k looks like this:

```
model
{
      # prior on the rate parameter theta
      theta ~ dunif(0,1)

      # observed wins k out of total games n
      k ~ dbin(theta,n)
}
```

The twiddles symbol ($\sim$) means "is distributed as". Because we use a Uniform distribution between 0 and 1 as a prior on the rate parameter θ, we write `theta ~ dunif(0,1)`. This indicates that, a priori, each value of θ is equally likely. Furthermore, k is Binomially distributed with parameters θ and n (i.e., `k ~ dbin(theta,n)`). Note that `dunif` and `dbin` are two of the predefined distributions in WinBUGS. All the distributions that are predefined in WinBUGS are listed in the distributions section in the WinBUGS manual, which can be accessed by clicking the help menu and selecting User manual (or by pressing F1). The hash symbol (#) is used for comments. The lines starting with this symbol are not executed by WinBUGS.

Copy the text into an empty file and save it as "model_rateproblemfunction.txt" in the directory from where we want to work. There are now various ways in which to proceed. One way is to work from within WinBUGS; another way is to control WinBUGS by calling it from a more general purpose program. Here, we use the statistical programming language R (R Development Core Team, 2004) to call WinBUGS, but widely-used alternative research programming environments such as MATLAB are also available (Lee & Wagenmakers, 2009).

A rate problem: The R script

The next step is to construct an R-script to call BlackBox from R.[5] When we run the script "rscript_rateproblemfunction.R", WinBUGS starts, the MCMC sampling is conducted, WinBUGS closes, and we return to R. The object that WinBUGS has returned to R is called "rateproblem", and this object contains all the information about the Bayesian inference for θ.

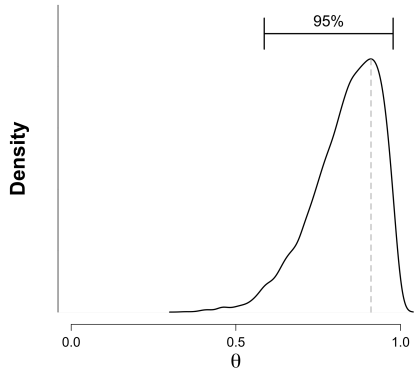


Figure B.1: The posterior distribution of the rate parameter θ after observing 9 wins out of 10 games. The dashed gray line indicates the mode of the posterior distribution at $\theta = .90$. The 95% credible interval extends from .59 to .98.

In particular, the "rateproblem" object contains a single sequence of consecutive draws from the posterior distribution of θ, a sequence that is generally known as an MCMC *chain*. We use the samples from the MCMC chain to estimate the posterior distribution of θ. To arrive at the posterior distribution, the samples are not plotted as a time series but as a distribution. In order to estimate the posterior distribution of θ, we applied the standard density estimator in R. Figure B.1 shows that the mode of the distribution is very close to .90, just as we expected. The posterior distribution is relatively spread

[5] All the scripts can be found on the website of the first author: `http://www.ruudwetzels.com`.

out over the parameter space, and the 95% credible interval extends from .59 to .98, indicating the uncertainty about the value of θ. Had we observed 900 wins out of a total of 1000 games the posterior of θ would be much more concentrated around the mode of .90, as our knowledge about the true value of θ would have greatly increased.

Example 2: ObservedPlus

In this section we examine the rate problem again, but now we change the variables. Suppose we learn that before we observed the current data, 10 games had already been played, resulting in a single win. To add this information, we design a function that adds 1 to the number of observed wins, and 10 to the number of total games. So, when we use $k = 9$ and $n = 10$ as before, we end up with

$$k_{new} = k_{old} + 1 = 9 + 1 = 10 \tag{B.2}$$

and

$$n_{new} = n_{old} + 10 = 10 + 10 = 20. \tag{B.3}$$

Hence, when we use our new function, the mode of the posterior distribution should no longer be .90 but .50 ($10/20 = .50$). Of course, this particular problem does not require the use of WBDev, and could easily be handled using plain WinBUGS code. It is the simplicity of the present problem, however, that makes it suitable as an introductory WBDev example.

In order to apply WBDev to the above problem, we are going to build a function called "ObservedPlus", using the template "VectorTemplate.odc". This template is located in the folder "...\$BlackBoxComponentBuilder1.5\WBdev\Mod$".

ObservedPlus: The WBDev script

The script file "ObservedPlus.odc" shows text in three colors. The parts that are colored black should not be changed. The parts in red are comments and these are not executed by BlackBox. The parts in blue are the most relevant parts of the code, because these are the parts that can be changed to create the desired function.

The templates for coding the functions and distributions—written by David Lunn and Chris Jackson—come bundled with the WBDev software. These templates support the development of new functions and distributions, such that researchers can focus on the specific functions they wish to implement without having to worry about programming Component Pascal code from scratch.

We now give a detailed explanation of the ObservedPlus WBDev function. The numbers **(*X*)** correspond to the numbers in the ObservedPlus WBDev script. For this simple example, we show some crucial parts of the WBDev scripts below.

(*1*) `MODULE WBDevObservedPlus;`

The name of the module is typed here. We have named our module ObservedPlus.

The name of the module (so the part after MODULE WBDev...) has to start with a capital letter.

(*2*) `args := "ss";`

Here we must define specific arguments about the input of the function. We can choose between scalars (s) and vectors (v). A scalar means that the input is a

single number. When we want to use a variable that consists of more numbers (for example a time series) we need a vector. This line has to correspond with the constants defined at **(*3*)**. In our example, we use two scalars, the number of successes `k` and the total number of observations `n`.

(*3*) `in = 0; ik = 1;`

Because of what has been defined at **(*2*)**, WBDev already knows that there should be two variables here. We name them `in` and `ik`, with `in` at the first spot (with number 0) and `ik` at the second spot (with number 1). WBDev always starts counting at 0 and not at 1.

Note that we did not name our variables `n` and `k`, but `in` and `ik`. This is because it is more insightful to use `n` and `k` later on, and it is not possible to give two or more variables the same name. Finally, note that the positions of the constants correspond to the positions of the input of the variables into the function in the model file. We will return to this issue later.

(*4*) `n, k:  INTEGER;`

The variables that are used in the calculations need to be defined. Both variables are defined as integers, because the Binomial distribution only allows integers as input: Counts of successes and the total games that are played can only be positive integers.

```
(*5*) n := SHORT(ENTIER(func.arguments[in][0].Value()));
      k := SHORT(ENTIER(func.arguments[ik][0].Value()));
```

This code takes the input values (`in` and `ik`) and gives them a name. We defined two variables in **(*4*)**, and we are now going to use them. What the script says here is: Take the input values `in` and `ik` and store them in the integer variables `n` and `k`. Because the input variables are not automatically assumed to be integers, we have to transform them and make sure the program recognizes them as integers. So, in other words, the first line says that `n` is the same as the first input variable of the function (see Figure B.2), and the second line says that `k` is the same as the second input variable of the function.

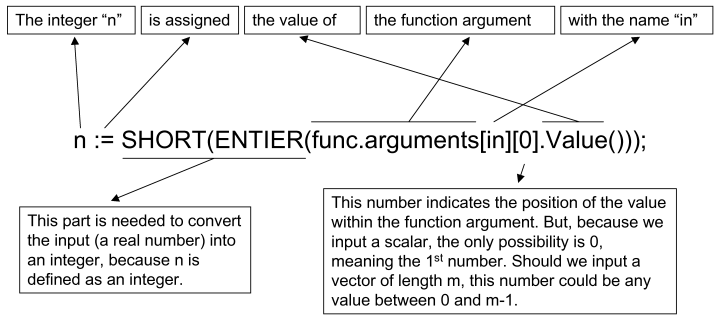


Figure B.2: A detailed explanation of part **(*5*)** of “ObservedPlus.odc”.

```
(*6*) n:=n+10;
      k:=k+1;
      values[0] := n;
      values[1] := k;
```

This is the part of the script where we do the actual calculations. At the end of this part, we fill the output array `values` with the new `n` and `k`.

```
(*7*) END WBDevObservedPlus.
```

Finally, we need to make sure that the name of the module at the end is the same as the name at the top of the file. The last line has to end with a period. Hence, the last line of the script is "`ENDWBDevObservedPlus.`".

Now we need to compile the function by pressing "Ctrl+k". Syntax errors cause WBDev to return an error message. Name this file "ObservedPlus.odc" and save it in the directory "...\$\backslash BlackBoxComponentBuilder1.5\backslash WBdev\backslash Mod$".

We are not entirely ready to use the function yet. WBDev needs to know that there exists a function called ObservedPlus; WBDev also needs to know what the input looks like (i.e., how many inputs are there, what order are they presented, and are they scalars and vectors?), and what the output is. To accomplish this, open the file "functions.odc" in the directory "...$\backslash BlackBoxComponentBuilder1.5\backslash WBdev\backslash Rsrc$". Add the line: `v<-"ObservedPlus"(s,s) "WBDevObservedPlus.Install"` and then save the file. The next time that WBDev is started, it knows that there is a function named ObservedPlus which has two scalars as input, and a vector as output. The function is now ready to be used in a model file.

ObservedPlus: The model file

In order to use the newly scripted function ObservedPlus we use a model file that is similar to the model file used in the earlier rate problem example.

```
model
{
      # Uniform prior on the rate parameter
      theta ~ dunif(0,1)

      # use the function to get the new n and the new k
      data[1:2] <- ObservedPlus(n,k)

      # define the new n and new k as variables
      newn <- data[1]
      newk <- data[2]

      # the new observed data
      newk ~ dbin(theta,newn)
}
```

We assume a Uniform prior on θ (i.e., `theta` $\sim$ `dunif(0,1)`). The function ObservedPlus takes as input the total number of games n and the number of wins k. From

them, the new n and new k can be calculated (i.e., `data[1:2] <- ObservedPlus(n,k)`). Note that functions require the use of the assignment operator (`<-`) instead of the twiddles symbol (∼). Remember that in the WBDev function the location of *in* was 0 and the location of *ik* was 1. Because that order was used, the input has to have n first and then k.

Next, `newn` is the first number in the vector `data` and `newk` is the second (i.e., `newn <- data[1]`, `newk <- data[2]`). Remember that when scripting in WBDev, the first element has index 0, but in the model file the first element has index 1. Finally, we use our new variables to do inference on the rate parameter θ (i.e., `newk ~ dbin(theta,newn)`).

Copy the text from the model file into an empty text file and name this file "model_observedplus.txt". Copy this file to the location of the model file that was used in the rate problem example.

ObservedPlus: The R script

To run this model from R, we can use the script of the original rate problem. The only thing that needs to be changed is the name of the model file. This should now be "model_observedplus.txt". Change this name and run the R-script.

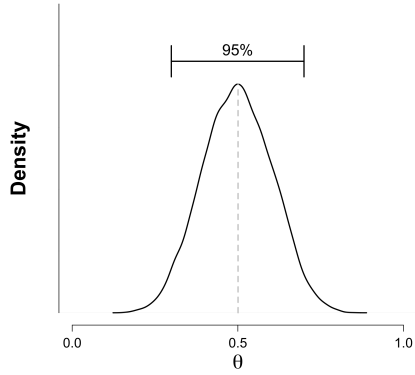


Figure B.3: The posterior distribution of the rate parameter θ, after using the function ObservedPlus. The dashed gray line indicates the mode of the posterior distribution at $\theta = .50$. The 95% credible interval extends from .30 to .70.

Figure B.3 shows the posterior distribution of θ. The mode of the distribution is .50, because $k_{new} = 10$ and $n_{new} = 20$. Again, because the total number of games played is fairly small, the posterior distribution of θ is relatively spread out (the 95% credible interval ranges from .30 to .70), reflecting our uncertainty about the true value of θ.

Example 3: The Expectancy-Valence Model

In the example described above, we could have used plain WinBUGS code instead of writing a script in BlackBox. But sometimes it can be very useful to write a BlackBox script instead of plain WinBUGS code, especially if the model under consideration is relatively complex. Implementing such a model into WBDev can speed up the computation time for inference substantially. The reason for this speed-up is that WBDev scripts are pre-compiled while the WinBUGS model files are interpreted at runtime. The present example, featuring the Exptectancy-Valence model to understand risk-seeking behavior in decision making, provides a concrete demonstration of this general point.

Suppose a psychologist wants to study decision making of clinical populations under controlled conditions. A task that is often used for this purpose is the "Iowa gambling task", developed by Bechara and Damasio (IGT; Bechara et al., 1994, 1997).

In the IGT, participants have to discover, through trial and error, the difference between risky and safe decisions. In the computerized version of the IGT, the participant starts with $2000 in play money. The computer screen shows players four decks of cards (A, B, C, and D), and then they have to select a card from one of the decks. Each card is associated with either a reward or a loss. The default payoff scheme is presented in Table B.1.

	Bad Decks		Good Decks	
	A	B	C	D
reward per trial	100	100	50	50
number of losses per 10 cards	5	1	5	1
loss per 10 cards	1250	1250	250	250
net profit per 10 cards	−250	−250	250	250

Table B.1: Rewards and Losses in the IGT. Cards from decks A and B yield higher rewards than cards from decks C and D, but they also yield higher losses. The net profit is highest for cards from decks C and D.

At the start of the IGT, participants are told that they should maximize net profit. During the task, they are presented with a running tally of the net profit, and the task finishes after 250 card selections.

The Expectancy-Valence (EV) model proposes that choice behavior in the IGT comes about through the interaction of three latent psychological processes. Each of these processes is vital for successful performance, typified by a gradual increase in preference for the good decks over the bad decks. First, the model assumes that the participant, after selecting a card from deck k, $k \in \{1, 2, 3, 4\}$ on trial t, calculates the resulting net profit or valence. This valence v_k is a combination of the experienced reward $W(t)$ and the experienced loss $L(t)$:

$$v_k(t) = (1 - w)W(t) + wL(t). \tag{B.4}$$

Thus, the first parameter of the Expectancy Valence model is w, the *attention weight* for losses relative to rewards, $w \in [0, 1]$.

On the basis of the sequence of valences v_k experienced in the past, the participant forms an expectation Ev_k of the valence for deck k. In order to learn, new valences need to update the expected valence Ev_k. If the experienced valence v_k is higher or lower than expected, Ev_k needs to be adjusted upward or downward, respectively. This intuition is captured by the equation

$$Ev_k(t + 1) = Ev_k(t) + a(v_k(t) - Ev_k(t)), \tag{B.5}$$

in which the *updating rate* $a \in [0, 1]$ determines the impact of recently experienced valences.

The EV model also uses a reinforcement learning method called softmax selection or Boltzmann exploration (Kaelbling et al., 1996; Luce, 1959) to account for the fact that participants initially explore the decks, and only after a certain number of trials decide

to always prefer the deck with the highest expected valence.

$$\Pr[S_k(t+1)] = \frac{\exp\left(\theta(t)Ev_k\right)}{\sum_{j=1}^{4}\exp\left(\theta(t)Ev_j\right)}. \tag{B.6}$$

In this equation, $1/\theta(t)$ is the "temperature" at trial t and $Pr(S_k)$ is the probability of selecting a card from deck k. In the EV model, the temperature is assumed to vary with the number of observations according to

$$\theta(t) = (t/10)^c, \tag{B.7}$$

where c is the *response consistency* or sensitivity parameter. In fits to data, this parameter is usually constrained to the interval $[-5, 5]$. When c is positive, response consistency θ increases (i.e., the temperature $1/\theta$ decreases) with the number of observations. This means that choices will be more and more guided by the expected valences. When c is negative, choices will become more and more random as the number of card selections increases.

In sum, the EV model decomposes choice behavior in the Iowa gambling task to three components or parameters:

1. An attention weight parameter w that quantifies the weighting of losses versus rewards.

2. An updating rate parameter a that quantifies the memory for rewards and losses.

3. A response consistency parameter c that quantifies the level of exploration.

The EV model: the WBDev script

To implement the EV model as a function in WBDev it is useful to first describe what data is observed and passed on to WinBUGS. In this example, we examine the data of one participant who has completed a 250-trial IGT. Hence, the observed data are an index of which deck was chosen at each trial, and the sequence of wins and losses that the participant incurred. Please see the script-file called "EV.odc".

(*1*) We name our module EV.

(*2*) In the EV example, we use 3 scalars for the 3 parameters and 3 vectors for the wins, losses and index at each trial.

(*3*) We start with the data vectors (the order is arbitrary, but needs to correspond to the one used in the model file) and we name these constants `iwins`, `ilosses` and `iindex`. After that, the function has as input the parameters of the EV-model, `iw`, `ia` and `ic`.

(*4*) In this section we define all the variables that we need to use in our calculations. Several mathematical functions are already available in WBDev. Information about these functions can be found by right-clicking the word "Math" in the script and then by clicking "documentation".

(*5*) Here we take our input EV parameters and assign them to the variables that we defined in part **(*4*)**.

(*6*) This is the part of the script where we do the actual calculations. At the end of this part, we fill the output variable called "values", with the output of our EV-function, the probability of choice for a deck.

(*7*) Make sure that the name of the module at the end is the same as the name at the top of the file. The last line has to end with a period.

(*11*) The DrawSample(.) procedure returns a pseudo-random number from the new distribution.

Name this file "EV.odc" and save it in the directory "...*BlackBoxComponentBuilder*1.5\ *WBdev**Mod*". Open the file"functions.odc" in the directory "...\ *BlackBox Component Builder 1.5*\ *WBdev*\ *Rsrc*". Add the line: `v <- "EV"(v,v,v,s,s,s) "WBDevEV.Install"` and then save the file. The next time that WBDev is started, it knows that there is a function named EV which has three vectors and three scalars as input, and a vector as output.

The EV-model: the model file

In order to use the EV-model we need to implement the graphical model in WinBUGS. The following model file is used in this example:

```
model
{
      # EV parameters are assigned prior distributions
      w ~ dunif(0,1)
      a ~ dunif(0,1)
      c ~ dunif(-5,5)

      # data from the EV function
      evprobs[1:1000] <- EV(wi[],lo[],ind[],w,a,c)

      # only use the information from the chosen deck
      # see explanation below
      for (i in 1:250)
      {
            p.EV[i,1] <- evprobs[deckA[i]]
            p.EV[i,2] <- evprobs[deckB[i]]
            p.EV[i,3] <- evprobs[deckC[i]]
            p.EV[i,4] <- evprobs[deckD[i]]
            ind[i] ~ dcat(p.EV[i,])
      }
}
```

The parameters of the model, w, a, c are assigned Uniform prior distributions. w and a are bounded between 0 and 1 and c is bounded between -5 and 5 (i.e., `w` $\sim$ `dunif(0,1), a` $\sim$ `dunif(0,1), c` $\sim$ `dunif(-5,5)`). The wins and the losses from the 250-trials are stored in the vectors *wi* and *lo*. The indices from the decks that were chosen are stored in the vector *ind*. Together with the EV parameters they are input for the EV function that calculates the probability per choice (i.e., `evprobs[1:1000] <- EV(wi[],lo[],ind[],w,a,c)`).

Note that this function calculates 1000 probabilities for a 250-trial dataset. This is because the probability for each deck is calculated, not only for the chosen deck but for all decks. So at each trial, four probabilities are calculated and for 250 trials this totals 1000 probabilities. However, we are only interested in the probability of the chosen deck.

To handle this problem, we make four vectors, `deckA`, `deckB`, `deckC` and `deckD` which are rows of length 250. Each vector contains a sequence of numbers where the number at position t is calculated by adding four to the number at position $t-1$ ($x_t = x_{t-1} + 4$). The vector `deckA` starts with number 1, `deckB` starts with number 2, `deckC` starts with number 3 and `deckD` starts with number 4. Using these vectors, we can disentangle the probabilities for each deck at each trial, `evprobs[deckA[i]]` corresponds to the probabilities of choosing deck 1 at each trial i, `evprobs[deckB[i]]` to the probabilities of choosing deck 2 at each trial i, `evprobs[deckC[i]]` to the probabilities of choosing deck 3 at each trial i and `evprobs[deckD[i]]` to the probabilities of choosing deck 4 at each trial.

Finally, we state that the choice for a deck at trial i (the observed data vector `ind`) is Categorically distributed (i.e., `ind[i]` $\sim$ `dcat(p.EV[i,])`). The Categorical distribution (which is a special case of the Multinomial distribution) is the probability distribution for the choice of a card deck. This distribution is a generalization of the Bernoulli distribution for a categorical random variable (i.e., the choice for one of the four decks at each trial of the IGT). Copy the text from the model into an empty file and save it as "model_ev.txt" in the directory from where we want to work.

The EV model: The R script

To run this model and to supply WinBUGS with the data, we use the R-script called "rscript_expectancyvalence.r". Change the working directory (in the R-script) to the directory where the model file is located on the computer. This script contains fictitious data from a person who completed a 250-trial IGT.

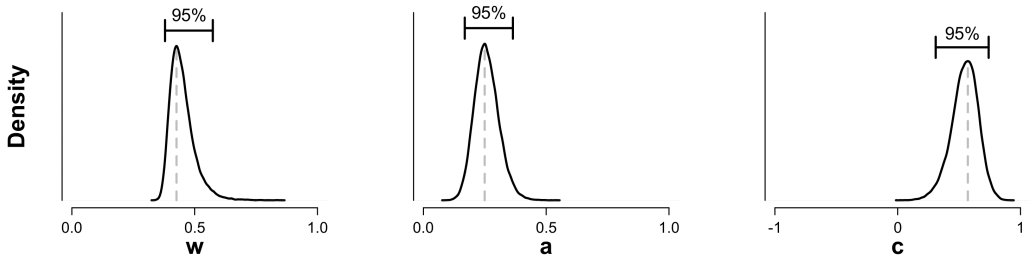


Figure B.4: The posterior distributions of the three EV parameters, w, a and c. The dashed gray lines indicate the modes of the posterior distributions at $w = .43$, $a = .25$ and $c = 0.58$. The 95% credible intervals for w, a and c extend from .38 to .57, from .17 to .36 and from 0.31 to 0.74, respectively.

Figure B.4 shows that the posterior mode of the attention weight parameter w is .43, the posterior mode of the update parameter a is .25 and the posterior mode of the consistency parameter c is 0.58.

On an average computer, it takes about 85 seconds to generate these posterior distributions. Had we used plain WinBUGS instead of WBDev code to compute these distributions, the calculation time would have taken approximately 15 minutes. Hence, implementing the function into WBDev speeds up the analysis by a factor 10.

B.4 Distributions

Statistical distributions are invaluable in psychological research. For example, in the simple rate problem discussed earlier, we use the Binomial distribution to model our data. WinBUGS comes equipped with an array of predefined distributions, but it does not include all distributions that are potentially useful for psychological modeling. Using WBDev, researchers can augment WinBUGS to include these desired distributions.

The next section explains how to write a new distribution, starting with the Binomial distribution as a simple introduction, and then considering the more complicated shifted Wald distribution.

Example 4: Binomial distribution

The Binomial distribution is already hard-coded in WinBUGS. But, because it is a very well-known and relatively simple distribution, it serves as a useful first example.

To program a distribution in WBDev, we can use the distribution template that is already in the BlackBox directory. This file is located in the folder: "...\$BlackBoxComponentBuilder1.5\WBdev\Mod". In order to program the distribution, we first need to write out the log likelihood function:

$$\begin{aligned}\log\left(Pr(K=k)\right) &= \log\left(f(k;n,\theta)\right)\\ &= \log\left(\binom{n}{k}\theta^k(1-\theta)^{n-k}\right)\\ &= \log\binom{n}{k} + \log(\theta^k) + \log(1-\theta)^{n-k}\\ &= \log(n!) - \log(k!) - \log(n-k)! + k\log(\theta) + (n-k)\log(1-\theta).\end{aligned} \tag{B.8}$$

Binomial distribution: The WBDev script

Here we describe the WDDev script for the Binomial distribution (See the file "BinomialTest.odc")

(*1*) We name our module BinomialTest.

(*2*) The parameters of the input of the Binomial distribution, *theta* and n.

(*3*) Here global variables can be declared. With global is meant that it is loaded only once, while the value of the variable may be needed many times. This part of the template does not need to be changed for this example.

(*4*) We have to declare what type of arguments are the input of the distribution. In this case these are two scalars (i.e. two single numbers), *theta* and n.

(*5*) This describes whether the distribution is discrete or continuous. When the distribution is discrete, isDiscrete should be set to TRUE. When the distribution is continuous, it should be set to FALSE. For the Binomial distribution isDiscrete is set to TRUE.

The other thing that is defined in this part of the script is if the cumulative distribution is to be provided. If so, canIntegrate should be set to TRUE. If this is set to true, an algorithm should be provided at **(*11*)**. We set canIntegrate to FALSE because we did not implement the cumulative distribution.

(*6*) This part of the code should define the natural bounds of the distribution. In our case, we take 0 as a lower bound and n as an upper bound, because k can never be larger than n.

(*7*) As the name implies, this is the part where the full log likelihood of the distribution is defined. This is an implementation of the log likelihood as defined in Equation B.8.

(*8*) Sometimes WinBUGS can ignore the normalizing constants. When that is the case, WinBUGS calls LogPropLikelihood(.). In our example, we refer back to the full log likelihood function.

(*9*) Occasionally, WinBUGS can make use of the LogPrior(.) procedure, which is proportional to the real log-prior function. In other words, this procedure omits the additive constants on the log scale. In our example, we just refer back to the full log likelihood function.

(*10*) This is the part where the cumulative distribution is defined when in part **(*7*)** canIntegrate is set to TRUE. Because we set this to FALSE, we do not define anything in this section.

(*11*) The DrawSample(.) procedure returns a pseudo-random number from the new distribution.

(*12*) The last thing that needs to be done is to make sure that the name of the module at the end is the same as the name at the top of the file. The last line has to end with a period.

Save this file as "BinomialTest.odc" and copy this file into the appropriate BlackBox directory, "...\$BlackBoxComponentBuilder1.5\WBdev\Mod$".

Open the distribution file "distributions.odc" in the directory "...\ *BlackBox Component Builder 1.5*\ *WBdev*\ *Rsrc*". Add the line `s ~ "BinomialTest"(s,s) "WBDevBinomialTest.Install"` and then save it.

Binomial distribution: The model file

To use the scripted Binomial distribution, we write a model file that is very similar to the model file used in the rate problem example. We only need to change the name of the distribution from dbin to BinomialTest.

```
model
{
      # prior on rate parameter theta
      theta~dunif(0,1)

      # observed wins k out of total games n
      k~BinomialTest(theta,n)

      # compute the posterior predictive of k
      postpred.k~BinomialTest(theta,n)
}
```

This example is essentially the same statistical problem as the first example, the rate problem. Ten games are played (i.e., $n = 10$) and nine games are won (i.e., $k = 9$). We assume a Uniform prior on θ (i.e., `theta ~ dunif(0,1)`). The observed wins k are distributed as our newly made BinomialTest with rate parameter *theta* and total games n (i.e., `k~BinomialTest(theta,n)`). With *theta* and k defined, this completes the model for BinomialTest. The Drawsample feature of the function ([**(*11*)**]), produces the posterior predictive values for k (i.e., `postpred.k~BinomialTest(theta,n)`). Save this file as "model_rateproblemdistribution.txt" and copy it to the working directory.

Binomial distribution: The R script

The last thing that we need to do is to start R and open the appropriate R-script "rscript_rateproblemdistribution.r". Change the working directory (in the R-script) to the directory where the model file is located on the computer. After running the code, the results should be similar to those shown in Figure B.1.

After having observed those data, the prediction of future data can be of interest. The so-called *posterior predictive distribution* gives the relative probability of different outcomes after the data have been observed. First, a sample is drawn from the joint posterior distribution. Next, data are generated using the posterior sample.

In this example these different outcomes can be $k = 1, 2, ..., 10$. The posterior predictive is often used for checking the assumptions of a model. If a model describes the data well, then the posterior predictive generated under the model should resemble the observed data. Large differences between the observed data and the posterior predictive imply that the model is not suitable for the data at hand. Figure B.5 shows the posterior predictive of k. The median of the posterior predictive is $k = 9$, which corresponds to the observed data.

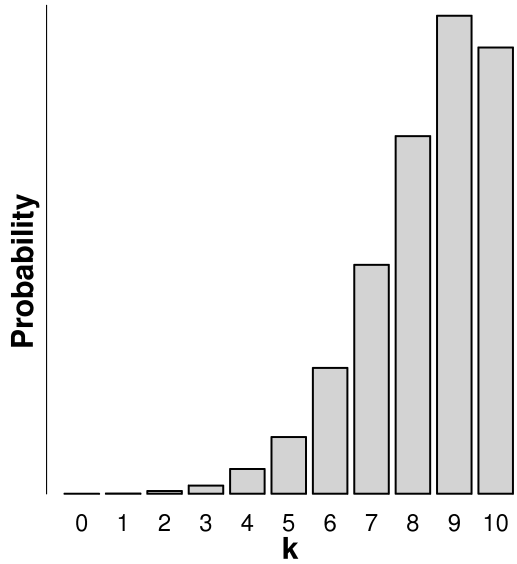


Figure B.5: The posterior predictive of k, the number of wins out of 10 games. The median of the posterior predictive is $k = 9$.

Example 5: Shifted Wald distribution

Many psychological models use response times (RTs) to infer latent psychological properties and processes (Luce, 1986). One common distribution used to model RTs is the inverse Gaussian or Wald distribution (Wald, 1947). This distribution represents the density of the first passage times of a Wiener diffusion process toward a single absorbing boundary, as shown in Figure B.6, using three parameters.

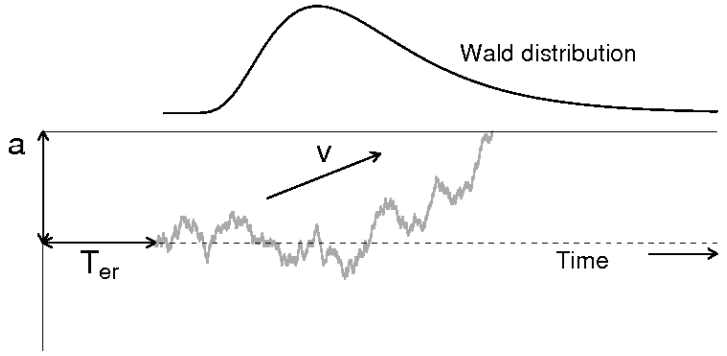


Figure B.6: A diffusion process with one boundary. The shifted Wald parameter a reflects the separation between the starting point of the diffusion process and the absorbing barrier, v reflects the drift rate of the diffusion process and T_{er} is a positive-valued parameter that shifts the entire distribution.

The parameter v reflects the drift rate of the diffusion process. The parameter a reflects the separation between the starting point of the diffusion process and the absorbing barrier. The third parameter, T_{er}, is a positive-valued parameter that shifts the entire distribution. The probability density function for this shifted Wald distribution is given by:

$$f(t|v,a,T_{er}) = \frac{a}{\sqrt{2\pi(t-T_{er})^3}} \exp\left\{ -\frac{[a-v(t-T_{er})]^2}{2(t-T_{er})} \right\}, \tag{B.9}$$

which is unimodel and positively skewed. Because of these qualitative properties, it is a good candidate for fitting empirical RT distributions. As an illustration, Figure B.7 shows changes in the shape of the shifted Wald distribution as a result of changes in the shifted Wald parameters v, a, and T_{er}.

The shifted Wald parameters have a clear psychological interpretation (e.g., Heathcote, 2004; Luce, 1986; W. Schwarz, 2001, 2002). Participants are assumed to accumulate noisy information until a predefined threshold amount is reached and a response is initiated. Drift rate v quantifies task difficulty or subject ability, response criterion a quantifies response caution, and the shift parameter T_{er} quantifies the time needed for non-decision processes (Matzke & Wagenmakers, 2009). Experimental paradigms in psychology for which it is likely that there is only a single absorbing boundary include saccadic eye movement tasks with few errors (R. H. S. Carpenter & Williams, 1995), go/no-go tasks (Gomez, Ratcliff, & Perea, 2007) or simple reaction time tasks (Luce, 1986, pp. 51–57). Here we show how to implement the shifted Wald distribution in WBDev.

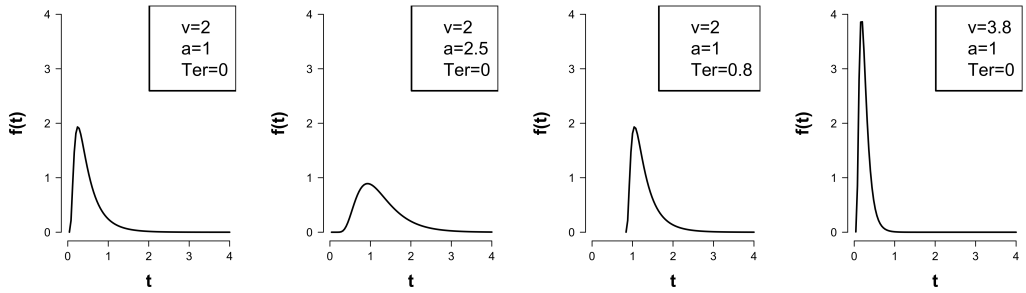


Figure B.7: Changes in the shape of the shifted Wald distribution as a result of changes in the parameters v, a and T_{er}. Each panel shows the shifted Wald distribution with different combinations of parameters.

Shifted Wald distribution: The WBDev script

Please see the supplementary materials for the WBDev code. Open BlackBox, and open the file "ShiftedWald.odc".

(*1*) We name our module ShiftedWald.

(*2*) The parameters of the distribution, which, in this case are the drift rate v, response caution a and shift T_{er}.

(*4*) We have to declare what type of arguments are the input of the distribution. In this case these are the three scalar parameters of the shifted Wald distribution.

(*6*) This part of the code should define the natural bounds of the distribution. In our case, we take Ter as a lower bound and INF (meaning $+\infty$) as an upper bound.

Save this file as "ShiftedWald.odc" and copy this file into the appropriate BlackBox directory, "...\$BlackBoxComponentBuilder1.5\WBdev\Mod$".

Open the distribution file "distributions.odc" in the directory "*...\ BlackBox Component Builder 1.5\ WBdev\ Rsrc*". Add the line `s ~ "ShiftedWald"(s,s,s) "WBDevShiftedWald.Install"` and then save it.

Shifted Wald distribution: The model file

Once we implemented the WBDev function in BlackBox, we can use the function ShiftedWald in the model. The model file is as follows:

```
model
{
      # prior distributions for shifted Wald parameters
      # drift rate
      v ~ dunif(0,10)

      # boundary separation
      a ~ dunif(0,10)

      # Non-decision time
```

```
      Ter ~ dunif(0,1)

      # data are shifted Wald distributed
      for (i in 1:nrt)
      {
            rt[i] ~ ShiftedWald(v,a,Ter)
      }
}
```

The priors for v and a, are Uniform distributions that range from 0 to 10 (i.e., `v` $\sim$ `dunif(0,10)`, `i.e., a` $\sim$ `dunif(0,10)`). The prior for T_{er} is a Uniform distribution that ranges from 0 to 1 (i.e., `Ter` $\sim$ `dunif(0,1)`. With the priors in place, we can use our ShiftedWald function to estimate the posterior distributions for the three model parameters v, a and T_{er} (i.e., `rt[i]` $\sim$ `ShiftedWald(v,a,Ter)`). Save the lines as a text file and name it "model_shiftedwaldind.txt".

Shifted Wald distribution: The R script

Now, open the R-script "rscript_shiftedwald_individual.r", for the individual analysis into an R-file and run it. Change the directory of the location of the model file and the location of the copy of BlackBox to the appropriate directories. The R-script loads a real data set from a lexical decision task (Wagenmakers, Ratcliff, Gomez, & McKoon, 2008). Nineteen participants had to quickly decide whether a visually presented letter string was a word (e.g., table) or a nonword (e.g., drapa). We will fit the response times of correct "word" responses of the first participant to the shifted Wald distribution. The response time data can be downloaded from `www.ruudwetzels.com`.

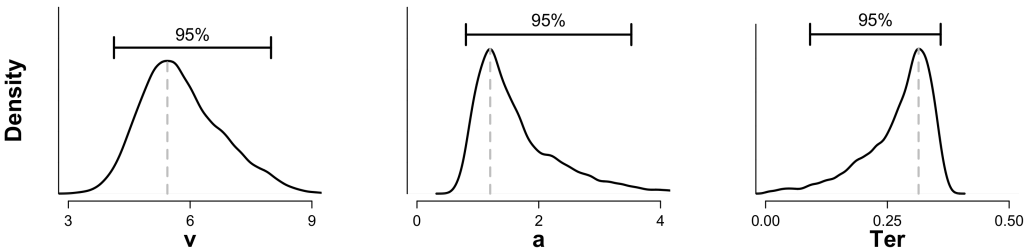


Figure B.8: The posterior distribution of the three Wald parameters v, a and T_{er}. The dashed gray lines indicate the modes of the posterior distributions at $v = 5.57$, $a = 1.09$ and $T_{er} = .33$. The 95% credible intervals for v, a and T_{er} extend from 4.12 to 8.00, from 0.80 to 3.52 and from .09 to .36, respectively.

Figure B.8 shows the posterior distribution of the three shifted Wald parameters, v, a and T_{er}. One thing that stands out is that the posterior distributions of the shifted Wald parameters are very spread out across the parameter space. The 95% credible intervals for v, a and T_{er} extend from 4.12 to 8.00, from 0.80 to 3.52 and from .09 to .36, respectively. It seems that data from only one participant are not enough to yield very accurate estimates of the shifted Wald parameters. In the following section we show how our estimates will improve when we use a hierarchical model and analyze all participants simultaneously.

Shifted Wald distribution: A hierarchical extension

In an experimental setting, the problem of few data per participant can be addressed by hierarchical modeling (Farrell & Ludwig, 2008; Gelman & Hill, 2007; Rouder, Sun, Speckman, Lu, & Zhou, 2003; Shiffrin et al., 2008). In our shifted Wald example, each subject is assumed to generate their data according to the shifted Wald distribution, but with different parameter values. We extend the individual analysis and assume that the parameters for each subject are governed by a group Normal distribution. This means that all individual participants are assumed to have their shifted Wald parameters drawn from the same group distribution, allowing the data from all the partipants to be used for inference, without making the unrealistic assumption that participants are identical copies of each other.

The model file that implements the hierarchical shifted Wald analysis is shown below:

```
model
{
      # prior distributions for group means:
      v.g ~ dunif(0,10)
      a.g ~ dunif(0,10)
      Ter.g ~ dunif(0,1)

      # prior distributions for group standard deviations:
      sd.v.g ~ dunif(0,5)
      sd.a.g ~ dunif(0,5)
      sd.Ter.g ~ dunif(0,1)

      # transformation from group standard deviations to group
      # precisions (i.e., 1/var, which is what WinBUGS expects
      # as input to the dnorm distribution):
      lambda.v.g <- pow(sd.v.g,-2)
      lambda.a.g <- pow(sd.a.g,-2)
      lambda.Ter.g <- pow(sd.Ter.g,-2)

      # data come From a shifted Wald distribution
      for (i in 1:ns)   #subject loop
      {
            # individual parameters drawn from group level
            # normals censored to be positive using the
            # I(0,) command:
            v.i[i] ~ dnorm(v.g,lambda.v.g)I(0,)
            a.i[i] ~ dnorm(a.g,lambda.a.g)I(0,)
            Ter.i[i] ~ dnorm(Ter.g,lambda.Ter.g)I(0,)

            # for each participant,
            # data are shifted Wald distributed
            for (j in 1:nrt[i])
            {
                  rt[i,j] ~ ShiftedWald(v.i[i],a.i[i],Ter.i[i])
            }
      }
}
```

The hierarchical analysis of the reaction time data proceeds as follows. The prior for the group means is a Uniform distribution, ranging from 0 to 10 (i.e., `v.g` $\sim$ `dunif(0,10)`, `a.g` $\sim$ `dunif(0,10)`) or from 0 to 1 (i.e., `Ter.g` $\sim$ `dunif(0,1)`). The standard de-

viations are drawn from a Uniform distribution ranging from 0 to 5 (i.e., `sd.v.g` $\sim$ `dunif(0,5)`, `sd.a.g` $\sim$ `dunif(0,5)`) or from 0 to 1 (i.e., `sd.Ter.g` $\sim$ `dunif(0,5)`). Next, the standard deviations have to be transformed to precisions (i.e., `lambda.v.g <- pow(sd.v.g,-2)`, `lambda.a.g <- pow(sd.a.g,-2)`, `lambda.Ter.g <- pow(sd.Ter.g, -2)`). Then, the individual parameters $v.i$, $a.i$ and $Ter.i$ are drawn from Normal distributions with corresponding group means and group precisions (i.e., `v.i[i]`$\sim$`dnorm(v.g ,lambda.v.g)I(0,)`, `a.i[i]` $\sim$ `dnorm(a.g, lambda.a.g)I(0,)`, `Ter.i[i]` $\sim$ `dnorm (Ter.g, lambda.Ter.g)I(0,)`). The `I(0,)` command indicates that the distribution is left-censored at 0. For each individual, the data are distributed according to a shifted Wald distribution with their own individual parameters. Save the model file as a text file and name it: "model_shiftedwaldhier.txt".

When we run this model using the R-script for the hierarchical analysis, we first focus on the group mean parameters $v.g$, $a.g$ and $T_{er}.g$. Figure B.9 shows the posterior distributions of the shifted Wald group-mean parameters. The distributions indicate that there is relatively little uncertainty about the parameter values. The posterior distributions of the group-mean parameters are concentrated around their modes $v.g = 4.27$, $a.g = 0.97$, and $T_{er}.g = 0.36$. The 95% credible intervals for $v.g$, $a.g$ and $T_{er}.g$ extend from 3.80 to 4.70, from 0.85 to 1.10 and from .34 to .38, respectively.

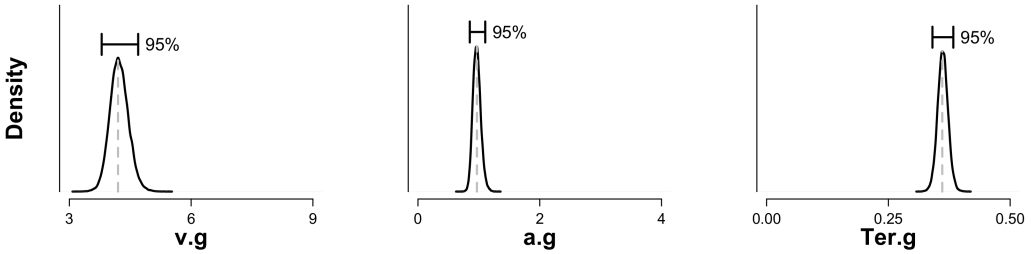


Figure B.9: The posterior distribution of the three "group-level" shifted Wald parameters $v.g$, $a.g$ and $T_{er}.g$. The dashed gray lines indicate the modes of the posterior distributions at $v.g = 4.27$, $a.g = .97$ and $T_{er}.g = .36$. The 95% credible intervals for $v.g$, $a.g$ and $T_{er}.g$ extend from 3.80 to 4.70, from 0.85 to 1.10 and from .34 to .38, respectively.

It is informative to consider the influence of the hierarchical extension on the individual estimates for the shifted Wald parameters. Specifically, we can examine the posterior distributions for the same subject that we analyzed in the individual shifted Wald analysis, but now in the hierarchical setting.

The hierarchical extension leads to a practical improvement, through a speed up of the MCMC estimation process. However, the hierarchical extension also leads to a theoretical improvement because compared to the individual analysis, the posterior distributions appear much less spread out. This shows that the hierarchical model leads to a better understanding of the model parameters.

To underscore this point, Figure B.10 shows the posterior distributions of the individual shifted Wald parameters, for both the hierarchical analysis and the individual analysis. It is clear that the posterior distributions of the shifted Wald parameters are less spread out in the hierarchical analysis than in the individual analysis. Also, the parameter estimates from the hierarchical analysis are slightly different than those from the individual analysis. In particular, they seem to have moved towards their common group mean. This effect is called *shrinkage*, and is a standard and important property of hierarchical models (Gelman et al., 2004).

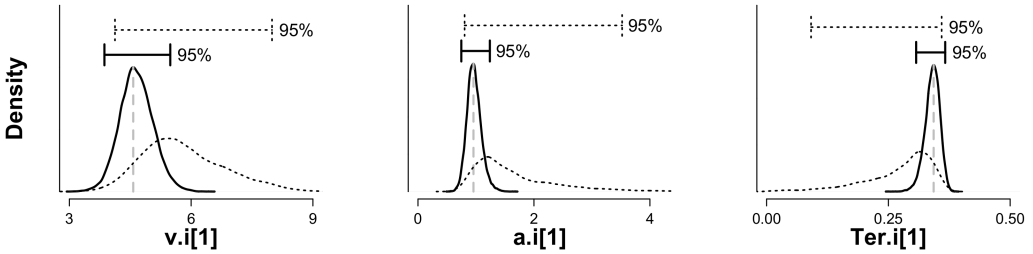


Figure B.10: The posterior distribution of the three individual shifted Wald parameters $v.i$, $a.i$ and $T_{er}.i$ from the hierarchical analysis (solid lines) and the individual analysis (dotted lines). The dashed gray lines indicate the modes of the posterior distributions from the hierarchical analysis at $v.i[1] = 4.57$, $a.i[1] = 0.96$ and $T_{er}.i[1] = .34$. The 95% credible intervals in the hierarchical model for $v.i[1]$, $a.i[1]$ and $T_{er}.i[1]$ extend from 3.86 to 5.49, from 0.75 to 1.24 and from .31 to .37,respectively.

In sum, the WBDev implementation of the shifted Wald distribution enables researchers to infer shifted Wald parameters from reaction time data. Not only does WinBUGS allow straightforward analyses on individual data, it also makes it easy to add hierarchical structure to the model (Lee, 2008; Rouder & Lu, 2005). This can greatly improve the quality of the posterior estimates, and is often a very sensible and informative way of analyzing data.

B.5 Discussion

In this paper we have shown how the WinBUGS Development Interface (WBDev) can be used to help psychological scientists model their sparse, noisy, but richly structured data. We have shown how a relatively complex function such as the Expectancy-Valence model can be incorporated in a fully Bayesian analysis of data. Furthermore, we have shown how to implement statistical distributions, such as the shifted Wald distribution, that have specific application in psychological modeling, but are not part of a standard set of statistical distributions.

The WBDev program is set up for Bayesian modeling, and is equipped with modern sampling techniques such as Markov chain Monte Carlo. These sampling techniques allow researchers to construct quantitative Bayesian models that are non-linear, highly structured, and potentially very complicated. The advantages of using WBDev together with WinBUGS are substantial. WinBUGS code can sometimes lead to slow computation and complex models might not work at all. Scripting some components of the model in WBDev can considerably speed up computation time. Furthermore, compartmentalizing the scripts can make the model easier to understand and debug. Moreover, WinBUGS facilitates statistical communication between researchers who are interested in the same model. The most basic advantage, however, is that WBDev allows the user to program functions and distributions that are simply unavailable in WinBUGS.

It should be mentioned that WinBUGS has an open-source alternative called OpenBUGS[6] (Thomas, OHara, Ligges, & Sturtz, 2006), a program that is also written in Component Pascal and also uses the BlackBox framework. Because OpenBUGS is open-source, users are allowed to add new functions and distributions. At the time of writing, OpenBUGS is

[6] `http://www.mathstat.helsinki.fi/openbugs/`.

supported by many researchers who continually adjust and improve the program; hence, OpenBUGS is a promising alternative to WinBUGS and WBDev. In order to take full advantage of the added flexibility of OpenBUGS, however, the user needs to have substantial knowledge of both Component Pascal and Bayesian inference. Instead, the WinBUGS and WBDev environment is more restricted, and this limitation makes it relatively easy for users to develop their own functions and distributions.

Once a core psychological model is implemented through WBDev, it is straightforward to take into account variability across participants or items using a hierarchical, multilevel extension (i.e., models with random effects for subjects or items). This approach allows a researcher to model individual differences as smooth variations in parameters of a certain cognitive model, reaching an optimal compromise between the extremes of complete pooling (i.e., treating all participants as identical copies) and complete independence (i.e., treating each participant as a fully independent unit). In general, statistical models that are implemented in WinBUGS can be easily implemented to deal with the complexities that plague empirical data, including when data are missing, the sampling plan is unclear or participants originate from different subgroups. For these reasons, we believe the fully Bayesian analysis of highly structured models is likely to be a driving force behind future theoretical and empirical progress in the psychological sciences.

C Appendix to Chapter 4: "Calculating the Bayes Factor Using R"

```
## Run the script below to compute the Bayes factors for the examples.
## Matlab code and info on how to compute the Bayes factors
## on Rouder's website (http://pcl.missouri.edu/bf-reg) can be found at
## www.ruudwetzels.com

## Functions to compute the Bayes Factor for (partial) correlation

jzs_corbf=function(r,n){
        int=function(r,n,g){
                (1+g)^((n-2)/2)*(1+(1-r^2)*g)^(-(n-1)/2)*
                g^(-3/2)*exp(-n/(2*g))};
        bf10=sqrt((n/2))/gamma(1/2)*integrate(int,lower=0,upper=Inf,r=r,n=n)$value;
return(bf10)}

jzs_partcorbf=function(r0,r1,p0,p1,n){
        int=function(r,n,p,g){
                (1+g)^((n-1-p)/2)*(1+(1-r^2)*g)^(-(n-1)/2)*
                g^(-3/2)*exp(-n/(2*g))};
        bf10=integrate(int,lower=0,upper=Inf,r=r1,p=p1,n=n)$value/
        integrate(int,lower=0,upper=Inf,r=r0,p=p0,n=n)$value;
return(bf10)}

###MacLean et al. 2010 (correlation)

r=-0.3616346;
n=54;
jzs_corbf(r,n) #BF10=3.85

###Kanai et al. in press (correlation)

r=0.4844199;
n=40;
jzs_corbf(r,n) #BF10=17.87

### Lleras et al. in press (partial correlation)

n=40;
p0=1;
p1=2;
r0=sqrt(0.6084);
r1=sqrt(0.6084408);
jzs_partcorbf(r0,r1,p0,p1,n) #BF10=0.13
```

D Appendix to Chapter 5: "Calculating the Bayes Factor Using R"

```
###
# For all functions:
# y is the response vector
# x is the design matrix
# R-scripts with the ANOVA examples can be found at
# www.ruudwetzels.com
###

## (1) Function to compute the Bayes Factor
## with Zellner's g-prior prior

zellner.g = function(y, x, g){
  output = matrix(,1,2)
  colnames(output) = c('BF_10','g/(g+1)')
  n = length(y)
  r2 = summary( lm(y ~ x) )$r.squared
  k = dim(x)[2] - 1
  output[1] = ( 1+g )^( (n-k-1)/2 )*( 1+g*(1-r2) )^(-(n-1)/2)
  output[2] = g / (g+1)
  return(output)}

## (2) Function to compute the Bayes Factor
## with Jeffreys-Zellner-Siow prior

zellnersiow = function(y, x){
  output = matrix(,1,2)
  colnames(output) = c('BF_10','g/(g+1)')
  n = length(y)
  r2 = summary( lm(y ~ x) )$r.squared
  k = dim(x)[2] - 1

  BF.integral = function(g, n = n, k = k, r2 = r2){
    (1+g)^((n-k-1)/2)*(1+g*(1-r2))^(-(n-1)/2)*g^(-3/2)*exp(-n/(2*g))}
    output[1] = ((n/2)^(1/2)/gamma(1/2)) * integrate(BF.integral,0,Inf,n=n,k=k,r2=r2)$value

  shrinkage.integral=function(g,n=n,k=k,r2=r2){
    (1+g)^((n-k-1-2)/2)*(1+g*(1-r2))^(-(n-1)/2)*g^(1-3/2)*exp(-n/(2*g))}
    g.=integrate(shrinkage.integral,0,Inf,n=n,k=k,r2=r2)$value

  output[2] = g. / integrate(BF.integral,0,Inf,n=n,k=k,r2=r2)$value
  return(output)}

## (3) Function to compute the Bayes Factor
## with Liang et al. hyper-g prior

hyper.g = function(y, x, a){
  output = matrix(,1,2)
  colnames(output) = c('BF_10','g/(g+1)')
  n = length(y)
  r2 = summary( lm(y ~ x) )$r.squared
  k = dim(x)[2]-1

  BF.integral=function(g, n=n, k=k, a=a, r2=r2){
```

```
  (1+g)^((n-k-1-a)/2)*(1+g*(1-r2))^(-(n-1)/2)}

output[1]=((a-2)/2)*integrate(BF.integral,0,Inf,n=n,a=a,k=k,r2=r2)$value
output[2]=(2/(k+a))*(f21hyper((n-1)/2,2,(k+a)/2+1,r2)/f21hyper((n-1)/2,1,(k+a)/2,r2))
return(output)}
```

E Appendix to Chapter 7: "Bem: a Robustness Analysis"

Abstract

In this online appendix 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.

In our manuscript "Why psychologists must change the way they analyze their data: The case of psi" we presented a Bayesian re-analysis of the data from Bem (2011). In particular, we analyzed each of Bems experiments using the default Bayesian t-test (Rouder et al., 2009). The results showed that there was no evidence for precognition to speak of. Table E.1 shows the results.

As explained in our main manuscript, the Bayes factor BF_{01} quantifies the evidence for H_0 (i.e., no precognition) versus H_1 (i.e., precognition). In order to calculate this Bayes factor, we need to specify a probability distribution for effect size, given H_1. That is, what effect sizes do we expect, should precognition really exist?

In our main manuscript, we used the default option that reflects a lack of knowledge about precognition —a Cauchy distribution on effect size that is centered around zero with scale parameter or probable error $r = 1$, that is, $\delta \sim \text{Cauchy}(0,1)$. This distribution is shown as the red line in Figure E.1.

However, one might argue that this default distribution is not appropriate, or, at least, that is sensible to examine other prior distributions on effect size as well. This was

Table E.1: The results of 10 crucial tests for the experiments reported in Bem (2011), reanalyzed using the default Bayesian t-test.

Exp	df	t	p	BF_{01}	Evidence category (in favor of $H_.$)
1	99	2.51	0.01	0.61	Anecdotal (H_1)
2	149	2.39	0.009	0.95	Anecdotal (H_1)
3	96	2.55	0.006	0.55	Anecdotal (H_1)
4	98	2.03	0.023	1.71	Anecdotal (H_0)
5	99	2.23	0.014	1.14	Anecdotal (H_0)
6	149	1.80	0.037	3.14	Substantial (H_0)
6	149	-1.74	0.041	3.49	Substantial (H_0)
7	199	-1.31	0.096	7.61	Substantial (H_0)
8	99	1.92	0.029	2.11	Anecdotal (H_0)
9	49	2.96	0.002	0.17	Substantial (H_1)

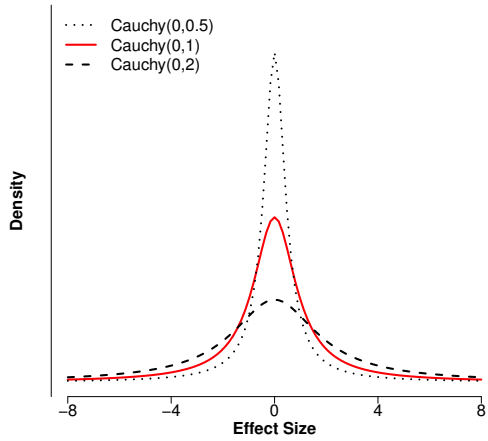


Figure E.1: Three examples of a Cauchy distribution. The solid line indicates the prior that underlies the default Bayesian t-test.

suggested independently by Patrizio Tressoldi (by Email) and Eric Kvaalen (on www.newscientist.com). In particular, one might argue that previous work has shown effect sizes in precognition and psi to be relatively small (e.g., Storm et al., 2010). Therefore, one could argue that instead of assuming $\delta \sim$ Cauchy(0,1), we might want to assume a Cauchy distribution that is more narrowly peaked, for instance $\delta \sim$ Cauchy(0,0.5), a distribution shown as the dotted line in Figure E.1. Naturally, one might then wonder whether and to what extent a change in the scale parameter of the Cauchy distribution fundamentally alters our conclusions.

In order to examine this possibility we conducted a robustness analysis in which we systematically varied the scale parameter r from 0 to 3 to quantify the effect that this has on the Bayes factor BF_{01}. The results are shown in Figure E.2.

Note that Figure E.2 plots the Bayes factor such that the scale of evidence in favor of H_0 is visually equivalent to the scale of evidence in favor of H_1. Also note that when $r = 0, H_0 = H_1$, and the Bayes factor indicates that the evidence is perfectly ambiguous (i.e., $BF_{01} = 1$).

The different panels in Figure E.2 indicate that our choice for the default prior does not affect our conclusions. In fact, the red dot —the result of our default test— seems to provide a relatively accurate summary of the evidence. Yes, it is true that for very small values of r the evidence is occasionally in favor of H_1, but —and this is the crucial point— only for the bottom right panel is the evidence clearly in favor of H_1. That is, in the bottom right panel the maximum Bayes factor is almost 1/10, meaning that the observed data are about 10 times more likely under H_1 than they are under H_0, given of course that the prior scale parameter r is chosen a posteriori, something that greatly biases the Bayes factor in favor of H_1.

For 7 out of the remaining 9 other panels, even the maximum Bayes factor indicates only "anecdotal" evidence (i.e., evidence worth "no more than a bare mention", that is, the data are less than 3 times more likely under H_1 than under H_0). This leaves the top-left two panels, for which the maximum Bayes factor does reach the criterion for

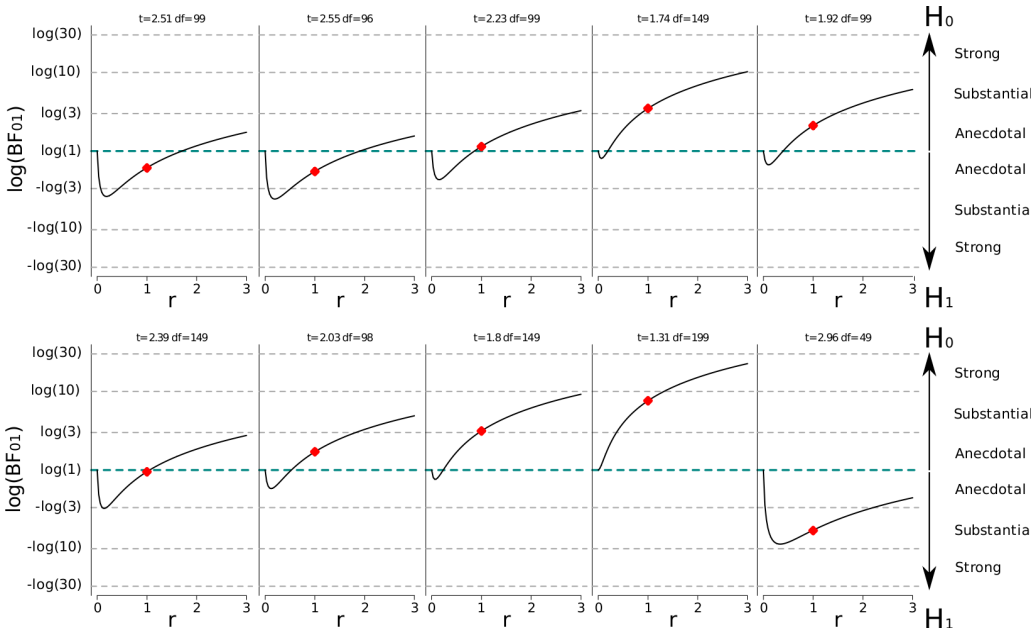


Figure E.2: A robustness analysis for the data from Bem (2011). The Bayes factor BF_{01} is plotted as a function of the scale parameter r of the Cauchy prior for effect size under H_1. The dot indicates the result from the default prior, the horizontal thick line in the middle of the plot indicates complete ambiguous evidence, and the horizontal grey lines demarcate the different qualitative categories of evidence (see our main manuscript). Importantly, the results in favor of H_1 are never compelling, except perhaps for the bottom right panel.

"substantial" evidence; however, it does so only just, and only for very specific values of the scale parameter. Again, the default test (indicated by the red dot) seems to provide a reasonable indication of the evidence.

In sum, we conclude that our results are robust to different specifications of the scale parameter for the effect size prior under H_1. This reinforces our general argument that p-values may strongly overstate the evidence against H_1.

F Appendix to Chapter 9: "Results from a Confirmatory Replication Study of Bem (2011)"

F.1 Introduction

In 2011, Dr. Bem published an article in the *Journal of Personality and Social Psychology*, the flagship journal of social psychology, in which he claimed that people can look into the future (Bem, 2011). In his first experiment, "precognitive detection of erotic stimuli", participants were instructed as follows: "(...) on each trial of the experiment, pictures of two curtains will appear on the screen side by side. One of them has a picture behind it; the other has a blank wall behind it. Your task is to click on the curtain that you feel has the picture behind it. The curtain will then open, permitting you to see if you selected the correct curtain." In the experiment, the location of the pictures was random and chance performance is therefore 50%. Nevertheless, Bem's participants scored 53.1%, significantly higher than chance; however, the effect was present only for erotic pictures, and not for neutral pictures, positive pictures, negative pictures, and romantic-but-not-erotic pictures. Bem also claimed that the psi effects are more pronounced for extraverts, and that for certain erotic pictures women show psi but men do not.

We set out to replicate Bem's experiment in a purely confirmatory fashion. First we detailed our method, design, and planned analyses in a document that we posted online before a single participant was tested.[1] As outlined in the online document, our replication focused on Bem's key findings; therefore, we tested only women, used only neutral and erotic pictures, and included a standard extraversion questionnaire. We also tested each participant in two contiguous sessions. Each session featured the same pictures, but presented them in a different random order.[2] The idea is that individual differences in psi –if these exist– lead to a positive correlation between performance on session 1 and session 2. Performance is quantified by the proportion of times that the participant chooses the curtain that hides the picture. Each session featured 60 trials, with 45 neutral pictures and 15 erotic pictures.

A vital part of the online document concerns the *a priori* specification of our analyses. First we outlined our main analysis tool, the Bayes factor t-test:

> "Data analysis proceeds by a series of Bayesian tests. For the Bayesian t-tests, the null hypothesis H_0 is always specified as the absence of a difference. Alternative hypothesis 1, H_1, assumes that effect size is distributed as Cauchy(0,1); this is the default prior proposed by Rouder et al. (2009). Alternative hypothesis 2, H_2, assumes that effect size is distributed as a half-normal distribution with positive mass only and the 90^{th} percentile at an effect size of 0.5; this is the "knowledge-based prior" proposed by Bem et al.

[1] See `http://confrep.blogspot.nl/` and `http://dl.dropbox.com/u/1018886/Advance_Information_on_Experiment_and_Analysis.pdf`.

[2] The online document detailed a method of yoking picture location and picture type. Due to a miscommunication with the programmer, yoking was not properly implemented and presentation of picture location and picture type was instead just random.

(submitted).[3] We will compute the Bayes factor for H_0 vs. H_1 (BF_{01}) and for H_0 vs. H_2 (BF_{02})."

The next six sections re-iterate the predictions from our online document and present the resulting Bayes factors. In the end, we tested 100 participants who each contributed two sessions. Because we use the Bayes factor we did not have the specify the number of participants in advance.

F.2 Results From a Confirmatory Study

Performance: Neutral vs. Erotic, Session 1

Confirmatory test 1: Based on the data of session 1 only: Does performance for erotic pictures differ from performance for neutral pictures? To address this question we compute a paired t test (Wetzels et al., 2009) and monitor BF_{01} and BF_{02} as the data come in. Figure F.1 shows the results.

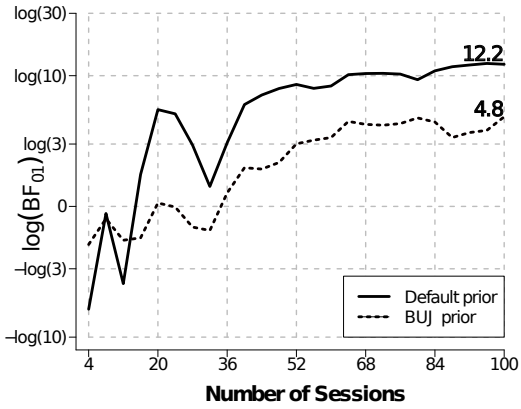


Figure F.1: Performance for erotic pictures does not differ from performance for neutral pictures (data from session 1 only). The logarithm of the Bayes factor is monitored as the data come in; $\log(BF_{01})$ is shown as a solid line, $\log(BF_{02})$ is shown as a dashed line.

Performance: Erotic vs. Chance, Session 1

Confirmatory test 2: Based on the data of session 1 only: Does performance for erotic pictures differ from chance (in this study 50%)? To address this question we compute a one-sample t test and monitor BF_{01} and BF_{02} as the data come in. Figure F.2 shows the results.

[3]Current addendum: this paper has since been published (i.e, Bem et al., 2011).

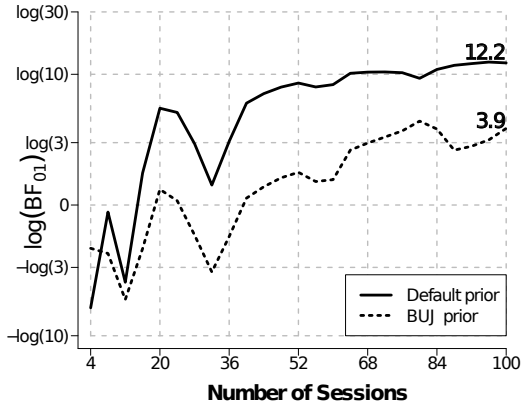


Figure F.2: Performance for erotic pictures does not differ from chance (data from session 1 only).The logarithm of the Bayes factor is monitored as the data come in; $\log(BF_{01})$ is shown as a solid line, $\log(BF_{02})$ is shown as a dashed line.

Correlation Extraversion and Performance Erotic Pictures

Confirmatory test 3: Based on the data of session 1 only: Is there a positive correlation between extraversion scores and performance for erotic pictures? This possibility was suggested by Bem (2011), and we assess this claim using the default Bayesian test for correlation proposed by Jeffreys (1961). Figure F.3 shows the results.

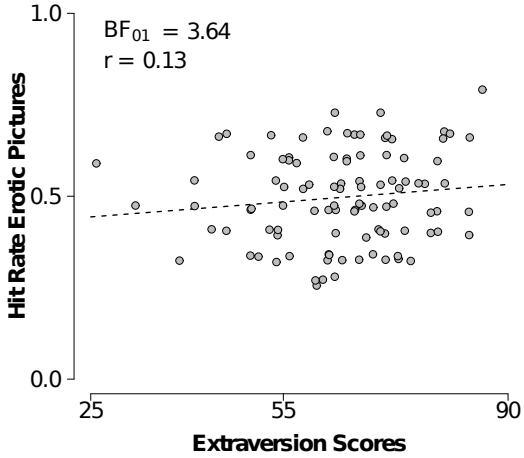


Figure F.3: There is no relation between extraversion scores and performance on erotic pictures. The correlation is 0.13, and the Bayes factor in favor of the null is 3.64. Note that this is not an order restricted test. Data are jittered to prevent visual overlap.

Correlation Between Performance Session 1 and Session 2

Confirmatory test 4: If participants have ESP, this trait should be related from session 1 to session 2. In other words, individual differences in ESP express themselves statistically as a positive correlation between performance on erotic pictures for session 1 and session 2. This prediction will again be tested using the default Bayesian test for correlation proposed by (Jeffreys, 1961). Figure F.4 shows the results.

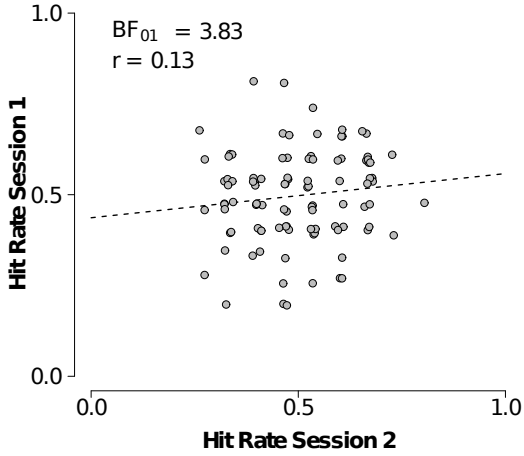


Figure F.4: There is no relation between performance on session 1 and session 2. The correlation is 0.13, and the Bayes factor in favor of the null is 3.83. Note that this is not an order restricted test. Data are jittered to prevent visual overlap.

Performance: Neutral vs. Erotic, Both Sessions Combined

Confirmatory test 5: Based on the data both sessions combined: Does performance for neutral pictures differ from performance for erotic pictures? To address this question we compute a paired t test and monitor BF_{01} and BF_{02} as the data come in. Figure F.5 shows the results.

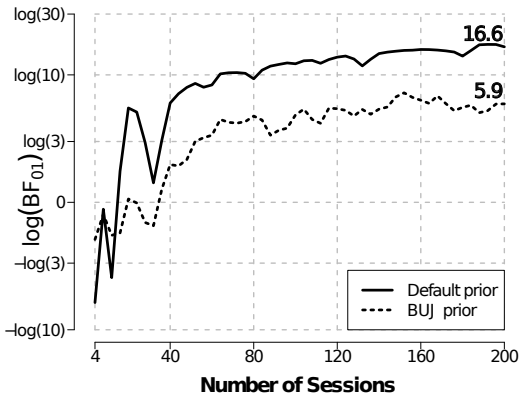


Figure F.5: Performance for erotic pictures does not differ from performance for neutral pictures (data from sessions 1 and 2).The logarithm of the Bayes factor is monitored as the data come in; $\log(BF_{01})$ is shown as a solid line, $\log(BF_{02})$ is shown as a dashed line.

Performance: Erotic vs. Chance, Both Sessions Combined

Confirmatory test 6: Based on the data both sessions combined: Does performance for neutral pictures differ from performance for erotic pictures? To address this question we compute a paired t test and monitor BF_{01} and BF_{02} as the data come in. Figure F.6 shows the results.

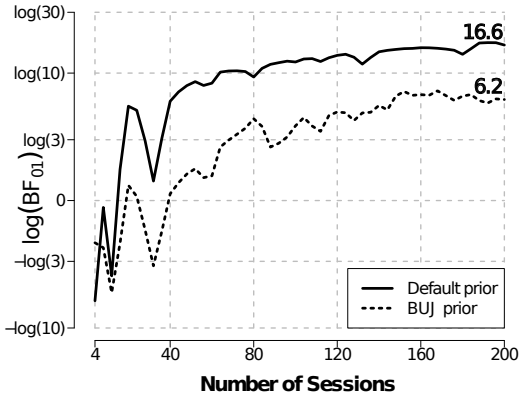


Figure F.6: Performance for erotic pictures does not differ from performance for neutral pictures (data from sessions 1 and 2).The logarithm of the Bayes factor is monitored as the data come in; $\log(BF_{01})$ is shown as a solid line, $\log(BF_{02})$ is shown as a dashed line.

F.3 Conclusion

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.

References

Abdi, H. (2003). Partial regression coefficients. *Encyclopedia of Social Sciences Research Methods. Thousand Oaks, CA: Sage*.

Abramowitz, M., & Stegun, I. (1972). *Handbook of mathematical functions*. New York: Dover Publications.

American Psychological Association. (2010). Publication Manual of the American Psychological Association (6th ed.). *Washington, DC, American Psychological Association*.

Bakker, M., van Dijk, A., & Wicherts, J. M. (in press). The rules of the game called psychological science. *Perspectives on Psychological Science*.

Ball, R. (2005). Experimental designs for reliable detection of linkage disequilibrium in unstructured random population association studies. *Genetics*, *170*, 859–873.

Bartlett, M. S. (1957). A comment on D. V. Lindley's statistical paradox. *Biometrika*, *44*, 533–534.

Batchelder, W. H. (2007). Cognitive psychometrics: Combining two psychological traditions. *CSCA Lecture, Amsterdam, The Netherlands, October 2007*.

Batchelder, W. H., & Riefer, D. M. (1999). Theoretical and empirical review of multinomial process tree modeling. *Psychonomic Bulletin & Review*, *6*, 57–86.

Bayarri, M. J., & Berger, J. (1991). Comment. *Statistical Science*, *6*, 379–382.

Bayarri, M. J., & García-Donato, G. (2007). Extending conventional priors for testing general hypotheses in linear models. *Biometrika*, *94*, 135–152.

Bechara, A., Damasio, A. R., Damasio, H., & Anderson, S. (1994). Insensitivity to future consequences following damage to human prefrontal cortex. *Cognition*, *50*, 7–15.

Bechara, A., Damasio, H., Tranel, D., & Damasio, A. R. (1997). Deciding advantageously before knowing the advantageous strategy. *Science*, *275*, 1293–1295.

Begley, C. G., & Ellis, L. M. (2012). Raise standards for preclinical cancer research. *Nature*, *483*, 531–533.

Bem, D. J. (2000). Writing an empirical journal article. In R. Sternberg (Ed.), *Guide to publishing in psychology journals* (pp. 3–16). Cambridge: Cambridge University Press.

Bem, D. J. (2003). Writing the empirical journal article. In J. M. Darley, M. P. Zanna, & H. L. Roediger III (Eds.), *The compleat academic: A career guide* (pp. 171–201). Washington, DC: American Psychological Association.

Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, *100*, 407–425.

Bem, D. J., Utts, J., & Johnson, W. O. (2011). Must psychologists change the way they analyze their data? *Journal of Personality and Social Psychology*, *101*, 716–719.

Berger, J. (2006). The case for objective bayesian analysis. *Bayesian Analysis*, *1*, 385–402.

Berger, J. O., & Delampady, M. (1987). Testing precise hypotheses. *Statistical Science*, *2*, 317–352.

Berger, J. O., & Jefferys, W. H. (1992). The Application of Robust Bayesian Analysis to Hypothesis Testing and Occam's Razor. *Statistical Methods and Applications*, *1*, 17–32.

Berger, J. O., & Pericchi, L. R. (1996). The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association*, *91*, 109–122.

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*, 112–139.

Berger, J. O., & Wolpert, R. L. (1988). *The likelihood principle (2nd ed.)*. Hayward (CA): Institute of Mathematical Statistics.

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

Berry, D. A., & Fristedt, B. (1985). *Bandit problems: Sequential allocation of experiments*. London: Chapman & Hall.

Billingsley, P. (2008). *Probability and measure*. New York: Wiley.

Bones, A. K. (2012). We knew the future all along: Scientific hypothesizing is much more accurate than other forms of precognition–a satire in one part. *Perspectives on Psychological Science*, *7*, 307–309.

Box, G. E. P., & Tiao, G. C. (1973). *Bayesian inference in statistical analysis*. Reading: Addison–Wesley.

Brown, G., Neath, I., & Chater, N. (2007). A temporal ratio model of memory. *Psychological Review*, *114*, 539–576.

Busemeyer, J. R., & Stout, J. C. (2002). A contribution of cognitive decision models to clinical assessment: Decomposing performance on the Bechara gambling task. *Psychological Assessment*, *14*, 253–262.

Carlin, B. P., & Chib, S. (1995). Bayesian model choice via Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society, Series B*, *57*, 473–484.

Carlin, B. P., & Louis, T. A. (2000). *Bayes and empirical Bayes methods for data analysis (2nd ed.)*. London: Chapman & Hall.

Caroselli, J. S., Hiscock, M., Scheibel, R. S., & Ingram, F. (2006). The simulated gambling paradigm applied to young adults: An examination of university students' performance. *Applied Neuropsychology*, *13*, 203–212.

Carpenter, R. H. S., & Williams, M. L. L. (1995). Neural computation of log likelihood in control of saccadic eye movements. *Nature*, *377*, 59–62.

Carpenter, S. (2012). Psychology's bold initiative. *Science*, *335*, 1558–1560.

Casella, G., & George, E. I. (1992). Explaining the Gibbs sampler. *The American Statistician*, *46*, 167–174.

Casella, G., & Moreno, E. (2006). Objective Bayesian variable selection. *Journal of the American Statistical Association*, *101*, 157–167.

Chipman, H. (1996). Bayesian variable selection with related predictors. *Canadian Journal of Statistics*, *24*, 17–36.

Christensen–Szalanski, J. J. J., & Willham, C. F. (1991). The hindsight bias: A meta–analysis. *Organizational Behavior and Human Decision Processes*, *48*, 147–168.

Clyde, M. (1999). Bayesian model averaging and model search strategies (with discussion). *Bayesian Statistics*, *6*, 157–185.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, *49*, 997–1003.

Cohen, J., Cohen, P., West, S., & Aiken, L. (2003). *Applied Multiple Regression / Correlation Analysis for the Behavioral Sciences*. Mahwah, NJ: Erlbaum.

Consonni, G., & Veronese, P. (2008). Compatibility of prior specifications across linear models. *Statistical Science*, *23*, 332–353.

Cortina, J. M., & Dunlap, W. P. (1997). On the logic and purpose of significance testing. *Psychological Methods*, *2*, 161–172.

Cowles, M. K. (2004). Review of WinBUGS 1.4. *The American Statistician*, *58*, 330–336.

Crone, E. A., & van der Molen, M. W. (2004). Developmental changes in real–life decision–making: Performance on a gambling task previously shown to depend on the ventromedial prefrontal cortex. *Developmental Neuropsychology*, *25*, 251–279.

Cui, W., & George, E. (2008). Empirical Bayes vs. fully Bayes variable selection. *Journal of Statistical Planning and Inference*, *138*, 888–900.

Cumming, G. (2008). Replication and p intervals: p values predict the future only vaguely, but confidence intervals do much better. *Perspectives on Psychological Science*, *3*, 286-300.

Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature*, *441*, 876–879.

Dawid, A. P. (1984). Statistical theory: The prequential approach. *Journal of the Royal Statistical Society A*, *147*, 278–292.

Dawid, A. P., & Lauritzen, S. L. (2001). Compatible prior distributions. In E. George (Ed.), *Bayesian methods with applications to science, policy, and official statistics* (pp. 109–118). Luxembourg: Monographs of Official Statistics.

DeGroot, M., & Schervish, M. (2002). *Probability and statistics*. Boston: Addison-Wesley Boston.

Dellaportas, P., Forster, J., & Ntzoufras, I. (in press). Joint specification of model space and parameter space prior distributions. *Statistical Science*.

Del Negro, M., & Schorfheide, F. (2008). Forming priors for DSGE models (and how it affects the assessment of nominal rigidities). *Journal of Monetary Economics*, *55*(7), 1191–1208.

Dennis, S. J., Lee, M. D., & Kinnell, A. (2008). Bayesian analysis of recognition memory: The case of the list–length effect. *Journal of Memory and Language*, *59*, 361–376.

Diaconis, P. (1978). Statistical problems in ESP research. *Science*, *201*, 131–136.

Diaconis, P. (1991). Comment. *Statistical Science*, *6*, 386.

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

Dickey, J. M., & Lientz, B. P. (1970). The weighted likelihood ratio, sharp hypotheses about chances, the order of a Markov chain. *The Annals of Mathematical Statistics*, *41*, 214–226.

Diener, E., Ng, W., Harter, J., & Arora, R. (2010). Wealth and happiness across the world: Material prosperity predicts life evaluation, whereas psychosocial prosperity predicts positive feeling. *Journal of Personality and Social Psychology*, *99*, 52–61.

Dienes, Z. (2008). *Understanding psychology as a science: An introduction to scientific and statistical inference*. Basingstoke: Palgrave Macmillan.

Dienes, Z. (2011). Bayesian versus Orthodox statistics: Which side are you on? *Perspectives on Psychological Science*, *X*, X–X.

Dixon, P. (2003). The p–value fallacy and how to avoid it. *Canadian Journal of Experimental Psychology*, *57*, 189–202.

Donkin, C., Averell, L., Brown, S., & Heathcote, A. (2009). Getting more from accuracy and response time data: Methods for fitting the linear ballistic accumulator model. *Behavior Research Methods*, *41*, 1095–1110.

Draper, N., & Smith, H. (1998). *Applied Regression Analysis*. New York: Wiley–Interscience.

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, *70*, 193–242.

Erdfelder, E. (2010). A Note on Statistical Analysis. *Experimental Psychology*, *57*, 1–4.

Estes, W. K. (1950). Toward a statistical theory of learning. *Psychological Review*, *57*, 94–107.

Estes, W. K. (1956). The problem of inference from curves based on group data. *Psychological Bulletin*, *53*, 134–140.

Estes, W. K. (2002). Traps in the route to models of memory and decision. *Psychonomic Bulletin & Review*, *9*, 3–25.

Faraway, J. (2002). *Practical regression and ANOVA using R.* Retrieved from `http://cran.r-project.org/doc/contrib/Faraway-PRA.pdf` (Available at http://cran.r-project.org/doc/contrib/Faraway-PRA.pdf)

Farrell, S., & Ludwig, C. (2008). Bayesian and maximum likelihood estimation of hierarchical response time models. *Psychonomic Bulletin & Review*, *15*, 1209–1217.

Feller, W. (1970). *An introduction to probability theory and its applications: Vol. I.* New York: John Wiley & Sons.

Feller, W. (1971). *An introduction to probability theory and its applications: Vol. ii.* New York: John Wiley & Sons.

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

Fisher, R. A. (1935). *The design of experiments.* Edinburgh: Oliver and Boyd.

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

Frank, M. C., & Saxe, R. (in press). Teaching replication to promote a culture of reliable science. *Perspectives on Psychological Science*.

Frick, R. W. (1996). The appropriate use of null hypothesis testing. *Psychological Methods*, *1*, 379–390.

Gallistel, C. (2009). The importance of proving the null. *Psychological Review*, *116*(2), 439–453.

Gamerman, D., & Lopes, H. F. (2006). *Markov chain Monte Carlo: Stochastic simulation for Bayesian inference.* Boca Raton, FL: Chapman & Hall/CRC.

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

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

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models.* Cambridge: Cambridge University Press.

George, E., & McCulloch, R. (1997). Approaches for Bayesian variable selection. *Statistica Sinica*, *7*, 339–373.

Gigerenzer, G. (1993). The Superego, the Ego, and the Id in statistical reasoning. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 311–339). Hillsdale (NJ): Erlbaum.

Gigerenzer, G. (1998). We need statistical thinking, not statistical rituals. *Behavioral and Brain Sciences*, *21*, 199–200.

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (Eds.). (1996). *Markov chain Monte Carlo in practice.* Boca Raton (FL): Chapman & Hall/CRC.

Gilks, W. R., Thomas, A., & Spiegelhalter, D. J. (1994). A language and program for complex Bayesian modelling. *The Statistician*, *43*, 169–177.

Gill, J. (2002). *Bayesian methods: A social and behavioral sciences approach.* Boca Raton (FL): CRC Press.

Goldacre, B. (2009). *Bad science.* London: Fourth Estate.

Gomez, P., Ratcliff, R., & Perea, M. (2007). A model of the go/no–go task. *Journal of Experimental Psychology: General*, *136*, 389-413.

Gönen, M., Johnson, W. O., Lu, Y., & Westfall, P. H. (2005). The Bayesian two–sample *t* test. *The American Statistician*, *59*, 252–257.

Good, I. J. (1983). *Good thinking: The foundations of probability and its applications.* Minneapolis: University of Minnesota Press.

Good, I. J. (1985). Weight of evidence: A brief survey. In J. M. Bernardo, M. H. DeGroot, D. V. Lindley, & A. F. M. Smith (Eds.), *Bayesian statistics 2* (pp. 249–269). New York: Elsevier.

Grahe, J., Reifman, A., Herman, A., Walker, M., Oleson, K., Nario–Redmond, M., et al. (in press). Harnessing the undiscovered resource of student research projects. *Perspectives on Psychological Science*.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, *82*, 711–732.

Griffiths, T. L., Kemp, C., & Tenenbaum, J. B. (2008). Bayesian models of cognition. In R. Sun (Ed.), *Cambridge handbook of computational cognitive modeling* (pp. 59–100). New York: Cambridge University Press.

Hagen, R. L. (1997). In praise of the null hypothesis statistical test. *American Psychologist*, *52*, 15–24.

Heathcote, A. (2004). Fitting Wald and ex–Wald distributions to response time data: An example using functions for the S–PLUS package. *Behavior Research Methods, Instruments, & Computers*, *36*, 678–694.

Heathcote, A., Brown, S., & Mewhort, D. J. K. (2000). The power law repealed: The case for an exponential law of practice. *Psychonomic Bulletin & Review*, *7*, 185–207.

Hedges, L. (1981). Distribution Theory for Glass's Estimator of Effect size and Related Estimators. *Journal of Educational and Behavioral Statistics*, *6*, 107.

Hinson, J. M., Jameson, T. L., & Whitney, P. (2002). Somatic markers, working memory, and decision making. *Cognitive, Affective, & Behavioral Neuroscience*, *2*, 341–353.

Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian model averaging: A tutorial. *Statistical Science*, *14*, 382-417.

Hoffman, L., & Rovine, M. J. (2007). Multilevel models for the experimental psychologist: Foundations and illustrative examples. *Behavior Research Methods*, *39*, 101–117.

Hoijtink, H., Klugkist, I., & Boelen, P. (2008). *Bayesian evaluation of informative hypotheses.* New York: Springer.

Howard, G., Maxwell, S., & Fleming, K. (2000). The proof of the pudding: An illustration of the relative strengths of null hypothesis, meta-analysis, and Bayesian analysis. *Psychological Methods*, *5*, 315–332.

Hume, D. (1748). *An enquiry concerning human understanding.*

Hyman, R. (2007). Evaluating parapsychological claims. In R. J. Sternberg, H. L. Roediger III, & D. F. Halpern (Eds.), *Critical thinking in psychology* (pp. 216–231). Cambridge: Cambridge University Press.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*, 696–701.

Ishwaran, H., & Rao, J. (2003). Detecting differentially expressed genes in microarrays using bayesian model selection. *Journal of the American Statistical Association*, *98*, 438–455.

Jahfari, S., Waldorp, L., Wildenberg, W. van den, Scholte, H., Ridderinkhof, K., & Forstmann, B. (2011). Effective connectivity reveals important roles for both the hyperdirect (fronto-subthalamic) and the indirect (fronto-striatal-pallidal) fronto-basal ganglia pathways during response inhibition. *The Journal of Neuroscience*, *31*(18), 6891–6899.

Jasny, B. R., Chin, G., Chong, L., & Vignieri, S. (2011). Again, and again, and again... *Science*, *334*, 1225.

Jaynes, E. T. (2003). *Probability theory: The logic of science*. Cambridge, UK: Cambridge University Press.

Jefferys, W. H. (1990). Bayesian analysis of random event generator data. *Journal of Scientific Exploration*, *4*, 153–169.

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

Jennison, C., & Turnbull, B. W. (1990). Statistical approaches to interim monitoring of medical trials: A review and commentary. *Statistical Science*, *5*, 299–317.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth–telling. *Psychological Science*, *23*, 524–532.

Kaelbling, L. P., Littman, M. L., & Moore, A. W. (1996). Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, *4*, 237–285.

Kanai, R., Bahrami, B., Roylance, R., & Rees, G. (in press). Online social network size is reflected in human brain structure. *Proceedings of the Royal Society B: Biological Sciences*.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 377–395.

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*, 928–934.

Kaufman, C. G., & Sain, S. R. (2010). Bayesian functional anova modeling using gaussian process prior distributions. *Bayesian Analysis*, *5*, 123–150.

Kennedy, J. E. (2001). Why is psi so elusive? A review and proposed model. *The Journal of Parapsychology*, *65*, 219–246.

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review*, *2*, 196–217.

Kerridge, D. (1963). Bounds for the frequency of misleading Bayes inferences. *The Annals of Mathematical Statistics*, *34*, 1109–1110.

Killeen, P. R. (2005). An alternative to null–hypothesis significance tests. *Psychological Science*, *16*, 345–353.

Killeen, P. R. (2006). Beyond statistical inference: A decision theory for science. *Psychonomic Bulletin & Review*, *13*, 549–562.

Killeen, P. R. (2007). Replication statistics as a replacement for significance testing: Best practices in scientific decision–making. In J. W. Osborne (Ed.), *Best practices in quantitative methods.* Thousand Oaks, CA: Sage Publications.

Kim, S., & Cohen, A. (1998). On the Behrens-Fisher Problem: A Review. *Journal of Educational and Behavioral Statistics*, *23*, 356–377.

Kleibergen, F. (2004). Invariant Bayesian inference in regression models that is robust against the Jeffreys–Lindley's paradox. *Journal of Econometrics*, *123*, 227–258.

Klugkist, I. (2008). Encompassing prior based model selection for inequality constrained analysis of variance. In H. Hoijtink, I. Klugkist, & P. Boelen (Eds.), *Bayesian evaluation of informative hypotheses.* (pp. 53–83). New York: Springer.

Klugkist, I., & Hoijtink, H. (2007). The Bayes factor for inequality and about equality constrained models. *Computational Statistics and Data Analysis*, *51*, 6367–6379.

Klugkist, I., Kato, B., & Hoijtink, H. (2005). Bayesian model selection using encompassing priors. *Statistica Neerlandica*, *59*, 57–69.

Klugkist, I., Laudy, O., & Hoijtink, H. (2005). Inequality constrained analysis of variance: A Bayesian approach. *Psychological Methods*, *10*, 477–493.

Kolmogorov, A. (1956). *Foundations of the theory of probability*. New York: Chelsea Publishing Company.

Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S. F., & Baker, C. I. (2009). Circular analysis in systems neuroscience: The dangers of double dipping. *Nature Neuroscience*, *12*, 535–540.

Kruschke, J. (In Press). What to believe: Bayesian methods for data analysis. *Trends in Cognitive Sciences*.

Kruschke, J. K. (1992). ALCOVE: An exemplar–based connectionist model of category learning. *Psychological Review*, *99*, 22–44.

Kruschke, J. K. (2010a). Bayesian data analysis. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*, 658–676.

Kruschke, J. K. (2010b). *Doing Bayesian data analysis: A tutorial introduction with R and BUGS*. Burlington: Academic Press.

Kruschke, J. K. (2011). Bayesian assessment of null values via parameter estimation and model comparison. *Perspectives on Psychological Science*, *X*, X–X.

Kuo, L., & Mallick, B. (1998). Variable selection for regression models. *Sankhya: The Indian Journal of Statistics, Series B*, *60*, 65–81.

Laudy, O. (2006). *Bayesian inequality constrained models for categorical data*. Unpublished doctoral dissertation, Utrecht University.

Lauritzen, S. L. (1996). *Graphical models*. Oxford: Clarendon.

Leamer, E. (1978). Regression selection strategies and revealed priors. *Journal of the American Statistical Association*, *73*, 580–587.

Lee, M. D. (2008). Three case studies in the Bayesian analysis of cognitive models. *Psychonomic Bulletin & Review*, *15*, 1–15.

Lee, M. D. (2011). How cognitive modeling can benefit from hierarchical Bayesian models. *Journal of Mathematical Psychology*, *1*, 1–7.

Lee, M. D., & Wagenmakers, E.-J. (2005). Bayesian statistical inference in psychology: Comment on Trafimow (2003). *Psychological Review*, *112*, 662–668.

Lee, M. D., & Wagenmakers, E.-J. (2009). A course in Bayesian graphical modeling for cognitive science. Unpublished course materials, retrieved September 10, 2009, from E–J Wagenmakers' website: `http://www.ejwagenmakers.com`.

Lee, M. D., & Webb, M. R. (2005). Modeling individual differences in cognition. *Psychonomic Bulletin & Review*, *12*, 605–621.

Lewis, S. M., & Raftery, A. E. (1997). Estimating Bayes factors via posterior simulation with the Laplace–Metropolis estimator. *Journal of the American Statistical Association*, *92*, 648–655.

Liang, F., Paulo, R., Molina, G., Clyde, M., & Berger, J. (2008). Mixtures of g priors for Bayesian variable selection. *Journal of the American Statistical Association*, *103*, 410–423.

Lindley, D. (1980). L.J. Savage-his work in probability and statistics. *The Annals of Statistics*, *8*, 1–24.

Lindley, D. (1997). Some comments on Bayes factors. *Journal of Statistical Planning and Inference*, *61*, 181–189.

Lindley, D. V. (1957). A statistical paradox. *Biometrika*, *44*, 187–192.

Lindley, D. V. (1972). *Bayesian statistics, a review*. Philadelphia (PA): SIAM.

Lindley, D. V. (1993). The analysis of experimental data: The appreciation of tea and wine. *Teaching Statistics*, *15*, 22–25.

Lindley, D. V. (2000). The philosophy of statistics. *The Statistician*, *49*, 293–337.

Lleras, A., Porporino, M., Burack, J., & Enns, J. (2011). Rapid resumption of interrupted

search is independent of age-related improvements in visual search. *Journal of Experimental Child Psychology*, *109*, 58–72.

Lodewyckx, T., Kim, W., Tuerlinckx, F., Kuppens, P., Lee, M. D., & Wagenmakers, E.-J. (2011). A tutorial on Bayes factor estimation with the product space method. *Journal of Mathematical Psychology*, *55*, 331–347.

Loftus, G. R. (1996). Psychology will be a much better science when we change the way we analyze data. *Current Directions in Psychological Science*, *5*, 161–171.

Luce, R. D. (1959). *Individual choice behavior*. New York: Wiley.

Luce, R. D. (1986). *Response times*. New York: Oxford University Press.

Lunn, D. (2003). WinBUGS Development Interface (WBDev). *ISBA Bulletin*, *10*, 10–11.

Lunn, D., Spiegelhalter, D., Thomas, A., & Best, N. (2009). The BUGS project: Evolution, critique and future directions. *Statistics in Medicine*, *28*, 3049–3067.

Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS – a Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, *10*, 325–337.

Mackay, C. (1852/1932). *Extraordinary popular delusions and the madness of crowds (2nd ed.)*. Boston: Page. Original second edition published 1852.

MacLean, K., Ferrer, E., Aichele, S., Bridwell, D., Zanesco, A., Jacobs, T., et al. (2010). Intensive meditation training improves perceptual discrimination and sustained attention. *Psychological Science*, *21*, 829–839.

Maruyama, Y. (2009). A Bayes factor with reasonable model selection consistency for anova model. *Arxiv preprint arXiv:0906.4329*.

Masson, M. E. J. (2011). A tutorial on a practical Bayesian alternative to null–hypothesis significance testing. *Behavior Reseach Methods*, *43*, 679–690.

Matzke, D., & Wagenmakers, E.-J. (2009). Psychological interpretation of ex–Gaussian and shifted Wald parameters: A diffusion model analysis. *Psychonomic Bulletin & Review*, *16*, 798-817.

Miller, I., & Miller, M. (2004). *John E. Freund's Mathematical Statistics with Applications*. New Jersey: Prentice Hall.

Mitchell, T., & Beauchamp, J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, *83*, 1023–1032.

Moreno, E., Bertolino, F., & Racugno, W. (1999). Default Bayesian analysis of the Behrens–Fisher problem. *Journal of Statistical Planning and Inference*, *81*, 323–333.

Morey, R. D., Pratte, M. S., & Rouder, J. N. (2008). Problematic effects of aggregation in zROC analysis and a hierarchical modeling solution. *Journal of Mathematical Psychology*, *52*, 376–388.

Morey, R. D., Rouder, J. N., Pratte, M. S., & Speckman, P. L. (2011). Using MCMC chain outputs to efficiently estimate Bayes factors. *Journal of Mathematical Psychology*, *55*, 368–378.

Morey, R. D., Rouder, J. N., & Speckman, P. L. (2008). A statistical model for discriminating between subliminal and near–liminal performance. *Journal of Mathematical Psychology*, *52*, 21–36.

Mulder, J., Klugkist, I., van de Schoot, R., Meeus, W. H. J., Selfhout, M., & Hoijtink, H. (2009). Bayesian model selection of informative hypotheses for repeated measurements. *Journal of Mathematical Psychology*, *53*, 530–546.

Mussweiler, T. (2006). Doing Is for Thinking! *Psychological Science*, *17*, 17–21.

Myung, I. J. (2003). Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology*, *47*, 90–100.

Myung, I. J., Forster, M. R., & Browne, M. W. (2000). A special issue on model selection. *Journal of Mathematical Psychology*, *44*.

Myung, I. J., & Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin & Review*, *4*, 79–95.

Myung, J. I., Karabatsos, G., & Iverson, G. J. (2008). A statistician's view on Bayesian evaluation of informative hypotheses. In H. Hoijtink, I. Klugkist, & P. Boelen (Eds.), *Bayesian evaluation of informative hypotheses.* (pp. 309–327). New York: Springer.

Navarro, D. J., Griffiths, T. L., Steyvers, M., & Lee, M. D. (2006). Modeling individual differences using Dirichlet processes. *Journal of Mathematical Psychology*, *50*, 101–122.

Neuroskeptic. (in press). The nine circles of scientific hell. *Perspectives on Psychological Science*.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*, 175–220.

Nickerson, R. S. (2000). Null hypothesis statistical testing: A review of an old and continuing controversy. *Psychological Methods*, *5*, 241–301.

Nosek, B. A., Spies, J. R., & Motyl, M. (in press). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*.

Nosofsky, R. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *115*, 39–57.

Ntzoufras, I. (2009). *Bayesian modeling using WinBUGS.* Hoboken, NJ: Wiley.

O'Hagan, A. (1995). Fractional Bayes factors for model comparison. *Journal of the Royal Statistical Society B*, *57*, 99–138.

O'Hagan, A., & Forster, J. (2004). *Kendall's advanced theory of statistics vol. 2B: Bayesian inference (2nd ed.).* London: Arnold.

O'Hara, R., & Sillanpää, M. (2009). A review of Bayesian variable selection methods: What, how and which. *Bayesian Analysis*, *4*, 85–118.

Osherovich, L. (2011). Hedging against academic risk. *Science–Business eXchange*, *4*.

Pearson, K. (1920). Notes on the history of correlation. *Biometrika*, *13*, 25–45.

Pitt, M. A., Myung, I. J., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review*, *109*, 472–491.

Poirier, D. J. (2006). The growth of Bayesian methods in statistics and economics since 1970. *Bayesian Analysis*, *1*, 969–980.

Press, S., Chib, S., Clyde, M., Woodworth, G., & Zaslavsky, A. (2003). *Subjective and objective Bayesian statistics: Principles, models, and applications.* Hoboken, New Jersey: Wiley-Interscience.

Price, G. R. (1955). Science and the supernatural. *Science*, *122*, 359–367.

Prinz, F., Schlange, T., & Asadullah, K. (2011). Believe it or not: How much can we rely on published data on potential drug targets? *Nature Reviews Drug Discovery*, *10*, 712–713.

Proschan, M. A., & Presnell, B. (1998). Expect the unexpected from conditional expectation. *The American Statistician*, *52*, 248–252.

Qian, S., & Shen, Z. (2007). Ecological Applications of Multilevel Analysis of Variance. *Ecology*, *88*, 2489–2495.

R Development Core Team. (2004). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `http://www.R-project.org` (ISBN 3–900051–00–3)

Raftery, A. E. (1995). Bayesian model selection in social research. In P. V. Marsden (Ed.), *Sociological methodology* (pp. 111–196). Cambridge: Blackwells.

Rao, M. (1988). Paradoxes in conditional probability. *Journal of Multivariate Analysis*, *27*, 434–446.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*, 59–108.

Ratcliff, R., & Tuerlinckx, F. (2002). Estimating parameters of the diffusion model: Approaches to dealing with contaminant reaction times and parameter variability. *Psychonomic Bulletin & Review*, *9*, 438–481.

Richard, F. D., Bond, C. F. J., & Stokes-Zoota, J. J. (2003). One hundred years of social psychology quantitatively described. *Review of General Psychology*, *7*, 331–363.

Robert, C. (1993). A note on jeffreys-lindley paradox. *Statistica Sinica*, *3*, 601–608.

Rodgers, J., & Nicewander, W. (1988). Thirteen ways to look at the correlation coefficient. *The American Statistician*, *42*, 59–66.

Roediger, H. L. (2012). Psychology's woes and a partial cure: The value of replication. *APS Observer*, *25*.

Rosenthal, R. (1990). How are we doing in soft psychology? *American Psychologist*, *45*, 775–777.

Rosenthal, R., & Rubin, D. (1982). A simple, general purpose display of magnitude of experimental effect. *Journal of educational psychology*, *74*, 166–169.

Rossell, D., Baladandayuthapani, V., & Johnson, V. E. (2008). Bayes factors based on test statistics under order restrictions. In H. Hoijtink, I. Klugkist, & P. A. Boelen (Eds.), *Bayesian evaluation of informative hypotheses.* (pp. 111–129). New York: Springer.

Rouder, J. N., & Lu, J. (2005). An introduction to Bayesian hierarchical models with an application in the theory of signal detection. *Psychonomic Bulletin & Review*, *12*, 573–604.

Rouder, J. N., Lu, J., Morey, R. D., Sun, D., & Speckman, P. L. (2008). A hierarchical process dissociation model. *Journal of Experimental Psychology: General*, *137*, 370–389.

Rouder, J. N., Lu, J., Speckman, P. L., Sun, D., & Jiang, Y. (2005). A hierarchical model for estimating response time distributions. *Psychonomic Bulletin & Review*, *12*, 195–223.

Rouder, J. N., Lu, J., Sun, D., Speckman, P., Morey, R., & Naveh-Benjamin, M. (2007). Signal detection models with random participant and item effects. *Psychometrika*, *72*, 621-642.

Rouder, J. N., & Morey, R. D. (2011). A Bayes–factor meta analysis of Bem's ESP claim. *Psychonomic Bulletin & Review*, *18*, 682–689.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t–tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*, 225–237.

Rouder, J. N., Sun, D., Speckman, P., Lu, J., & Zhou, D. (2003). A hierarchical bayesian statistical framework for response time distributions. *Psychometrika*, *68*, 589–606.

Roverato, A., & Consonni, G. (2004). Compatible prior distributions for DAG models. *Journal of the Royal Statistical Society B*, *66*, 47–61.

Royall, R. M. (1997). *Statistical evidence: A likelihood paradigm.* London: Chapman & Hall.

Sarewitz, D. (2012). Beware the creeping cracks of bias. *Nature*, *485*, 149.

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods*, *1*, 115–129.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461–464.

Schwarz, W. (2001). The ex–Wald distribution as a descriptive model of response times. *Behavior Research Methods, Instruments, & Computers*, *33*, 457–469.

Schwarz, W. (2002). On the convolution of inverse gaussian and exponential random variables. *Communications in Statistics, Theory and Methods*, *31*, 2113–2121.

Schweder, T., & Hjort, N. L. (1996). Bayesian synthesis or likelihood synthesis–what does Borel's paradox say? *Report International Whaling Commission*, *46*, 475–479.

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

Scott, J., & Berger, J. (2010). Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *The Annals of Statistics*, *38*, 2587–2619.

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

Sen, S., & Churchill, G. (2001). A statistical framework for quantitative trait mapping. *Genetics*, *159*, 371–287.

Shafer, G. (1982). Lindley's paradox. *Journal of the American Statistical Association*, *77*, 325–351.

Sheu, C.-F., & O'Curry, S. L. (1998). Simulation–based Bayesian inference using BUGS. *Behavioral Research Methods, Instruments, & Computers*, *30*, 232–237.

Shiffrin, R. M., Lee, M. D., Kim, W., & Wagenmakers, E.-J. (2008). A survey of model evaluation approaches with a tutorial on hierarchical Bayesian methods. *Cognitive Science*, *32*, 1248–1284.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False–positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*, 1359–1366.

Singpurwalla, N. D., & Swift, A. (2001). Network reliability and Borel's paradox. *The American Statistician*, *55*, 213–218.

Sinharay, S., & Stern, H. S. (2005). An empirical comparison of methods for computing Bayes factors in generalized linear mixed models. *Journal of Computational and Graphical Statistics*, *14*, 1-21.

Sisson, S. A. (2005). Transdimensional Markov chains: A decade of progress and future perspectives. *Journal of the American Statistical Association*, *100*, 1077-1089.

Spiegelhalter, D. J. (1998). Bayesian graphical modelling: A case–study in monitoring health outcomes. *Applied Statistics*, *47*, 115–133.

Spiegelhalter, D. J., Thomas, A., Best, N., & Lunn, D. (2003). *WinBUGS version 1.4 user manual*. Cambridge, UK: Medical Research Council Biostatistics Unit.

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

Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance–or vice versa. *Journal of the American Statistical Association*, *54*, 30–34.

Sterling, T. D., Rosenbaum, W. L., & Weinkam, J. J. (1995). Publication decisions revisited: The effect of the outcome of statistical tests on the decision to publish and vice versa. *The American Statistician*, *49*, 108–112.

Steyvers, M., Lee, M. D., & Wagenmakers, E.-J. (2008). A Bayesian analysis of human decision–making on bandit problems. *Journal of Mathematical Psychology*, *53*, 168–179.

Stigler, S. M. (1986). *The history of statistics: The measurement of uncertainty before 1900*. Cambridge, MA: Harvard University Press.

Stigler, S. M. (1989). Francis Galton's account of the invention of correlation. *Statistical Science*, *4*, 73–86.

Stone, C. J., Hansen, M. H., Kooperberg, C., & Truong, Y. K. (1997). Polynomial splines and their tensor products in extended linear modeling (with discussion). *The Annals of Statistics*, *25*, 1371–1470.

Storm, L., Tressoldi, P. E., & Di Risio, L. (2010). Meta–analysis of free–response studies, 1992–2008: Assessing the noise reduction model in parapsychology. *Psychological Bulletin*, *136*, 471–485.

Stout, J. C., Busemeyer, J. R., Lin, A., Grant, S. J., & Bonson, K. R. (2004). Cognitive modeling analysis of decision–making processes in cocaine abusers. *Psychonomic Bulletin & Review*, *11*, 742–747.

Strawderman, W. (1971). Proper Bayes minimax estimators of the multivariate normal mean. *The Annals of Mathematical Statistics*, *42*, 385–388.

Strube, M. J. (2006). SNOOP: A program for demonstrating the consequences of premature and repeated null hypothesis testing. *Behavior Research Methods*, *38*, 24–27.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. Cambridge (MA): The MIT Press.

Thomas, A., OHara, B., Ligges, U., & Sturtz, S. (2006). Making BUGS open. *R news*, *6*, 12–17.

Thompson, B. (2002). What future quantitative social science research could look like: Confidence intervals for effect sizes. *Educational Researcher*, *31*, 25–32.

Toutenburg, H., & Shalabh. (2009). *Statistical analysis of designed experiments*. New York: Springer Verlag.

Utts, J. (1991). Replication and meta–analysis in parapsychology (with discussion). *Statistical Science*, *6*, 363–403.

Vandekerckhove, J., Tuerlinckx, F., & Lee, M. D. (2008). A Bayesian approach to diffusion process models of decision–making. *Proceedings of the 30th Annual Conference of the Cognitive Science Society*, 1429–1434.

Vandekerckhove, J., Tuerlinckx, F., & Lee, M. D. (2011). Hierarchical diffusion models for two–choice response time. Submitted. *Psychological Methods*, *16*, 44–62.

Vanpaemel, W. (2010). Prior sensitivity in theory testing: An apologia for the bayes factor. *Journal of Mathematical Psychology*, *54*, 491–498.

Verdinelli, I., & Wasserman, L. (1995). Computing Bayes factors using a generalization of the Savage–Dickey density ratio. *Journal of the American Statistical Association*, *90*, 614–618.

Vul, E., Harris, C., Winkielman, P., & Pashler, H. (2009). Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Perspectives on Psychological Science*, *4*, 274–290.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, *14*, 779–804.

Wagenmakers, E.-J. (2009). Methodological and empirical developments for the Ratcliff diffusion model of response times and accuracy. *European Journal of Cognitive Psychology*, *21*, 641–671.

Wagenmakers, E.-J., & Grünwald, P. (2006). A Bayesian perspective on hypothesis testing. *Psychological Science*, *17*, 641–642.

Wagenmakers, E.-J., Grünwald, P., & Steyvers, M. (2006). Accumulative prediction error and the selection of time series models. *Journal of Mathematical Psychology*, *50*, 149–166.

Wagenmakers, E.-J., Lee, M. D., Lodewyckx, T., & Iverson, G. (2008). Bayesian versus

frequentist inference. In H. Hoijtink, I. Klugkist, & P. A. Boelen (Eds.), *Bayesian evaluation of informative hypotheses.* (pp. 181–207). New York: Springer Verlag.

Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., & Grasman, R. P. P. P. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the Savage-Dickey method. *Cognitive psychology*, *60*, 158–189.

Wagenmakers, E.-J., Ratcliff, R., Gomez, P., & McKoon, G. (2008). A diffusion model account of criterion shifts in the lexical decision task. *Journal of Memory and Language*, *58*, 140–159.

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., & van der Maas, H. L. J. (in press). Why psychologists must change the way they analyze their data: The case of psi. *Journal of Personality and Social Psychology*.

Wainer, H. (1999). One cheer for null hypothesis significance testing. *Psychological Methods*, *4*, 212–213.

Wald, A. (1947). *Sequential analysis.* New York: Wiley.

Wasserman, L. (2000). Bayesian model selection and model averaging. *Journal of Mathematical Psychology*, *44*, 92–107.

Wasserman, L. (2004). *All of statistics: A concise course in statistical inference.* New York: Springer.

Weaver, R. (2008). Parameters, predictions, and evidence in computational modeling: A statistical view informed by ACT-R. *Cognitive Science*, *32*, 1349–1375.

Westfall, P., & Gönen, M. (1996). Asymptotic properties of anova Bayes factors. *Communications in Statistics-Theory and Methods*, *25*, 3101–3123.

Wetzels, R., Lee, M., & Wagenmakers, E.-J. (in press). Bayesian inference using WBDev: A tutorial for social scientists. *Behavior Research Methods*.

Wetzels, R., Matzke, D., Lee, M., Rouder, J., Iverson, G., & Wagenmakers, E.-J. (2011). Statistical evidence in experimental psychology: An empirical comparison using 855 t tests. *Perspectives on Psychological Science*, *6*, 291–298.

Wetzels, R., Raaijmakers, J., Jakab, E., & Wagenmakers, E.-J. (2009). How to quantify support for and against the null hypothesis: A flexible WinBUGS implementation of a default Bayesian t–test. *Psychonomic Bulletin & Review*, *16*, 752–760.

Wetzels, R., Vandekerckhove, J., Tuerlinckx, F., & Wagenmakers, E.-J. (in press). Bayesian parameter estimation in the Expectancy Valence model of the Iowa gambling task. *Journal of Mathematical Psychology*, *54*, 14–27.

Wilkinson, L., & the Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*, 594–604.

Wiseman, R., & Schlitz, M. (1997). Experimenter effects and the remote detection of staring. *Journal of Parapsychology*, *61*, 197-207.

Wolpert, R. (1995). Comment–inference from a deterministic population dynamics model for bowhead whales. *Journal of the American Statistical Association*, *90*, 426–427.

Wood, S., Busemeyer, J., Koling, A., Cox, C. R., & Davis, H. (2005). Older adults as adaptive decision makers: Evidence from the Iowa gambling task. *Psychology and Aging*, *20*, 220–225.

Yechiam, E., & Busemeyer, J. R. (2005). Comparison of basic assumptions embedded in learning models for experience–based decision making. *Psychonomic Bulletin & Review*, *12*, 387–402.

Yechiam, E., Busemeyer, J. R., Stout, J. C., & Bechara, A. (2005). Using cognitive models to map relations between neuropsychological disorders and human decision–making deficits. *Psychological Science*, *16*, 973–978.

Yechiam, E., Kanz, J. E., Bechara, A., Stout, J. C., Busemeyer, J. R., Altmaier, E. M., et al. (2008). Neurocognitive deficits related to poor decision making in people behind bars. *Psychonomic Bulletin & Review*, *15*, 44–51.

Yechiam, E., Stout, J. C., Busemeyer, J. R., Rock, S. L., & Finn, P. R. (2005). Individual differences in the response to foregone payoffs: An examination of high functioning drug abusers. *Journal of Behavioral Decision Making*, *18*, 97–110.

Yong, E. (2012). Bad copy. *Nature*, *485*, 298–300.

Zellner, A. (1986). On assessing prior distributions and bayesian regression analysis with g-prior distributions. *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, 233–243.

Zellner, A. (1987). *An introduction to Bayesian inference in econometrics*. Malabar, FL: RE Krieger Pub. Co.

Zellner, A., & Siow, A. (1980). Posterior odds ratios for selected regression hypotheses. In J. M. Bernardo, M. H. DeGroot, D. V. Lindley, & A. F. M. Smith (Eds.), *Bayesian statistics* (pp. 585–603). Valencia: University Press.

Zeugner, S., & Feldkircher, M. (2009). Benchmark priors revisited: On adaptive shrinkage and the supermodel effect in bayesian model averaging. *IMF Working Papers*, *9*, 202.

Nederlandse Samenvatting

Deze dissertatie bestaat uit twee delen. In het eerste deel bespreken we Bayesiaanse alternatieven voor veelgebruikte frequentistische nulhypothesetoetsen. Daarbij gaan we ook in op de potentiële voordelen van deze toetsen ten opzichte van hun frequentistische tegenhangers. Vervolgens bespreken we in het tweede gedeelte hoe Bayesiaanse methodes het empirisch onderzoek in de sociale wetenschappen vooruit kan helpen.

Deel I: Bayesiaanse Model Selectie: Theoretisch

In het eerste hoofdstuk van het eerste deel, *Hoofdstuk twee*, stellen we een Bayesiaanse t test voor. Deze Savage-Dickey (SD) t test is geïnspireerd door de Jeffreys-Zellner-Siow (JZS) t test. De SD t test behoudt de belangrijkste eigenschappen van de JZS t test, maar is breder toepasbaar. De SD t test stelt onderzoekers bijvoorbeeld in staat om eenzijdig te toetsen en is ook toepasbaar in situaties waarbij de twee groepen niet dezelfde variantie hebben.

In *Hoofdstuk drie* bespreken we hoe de zogenaamde encompassing prior (EP) benadering, een methode die wordt toegepast bij Bayesiaanse model selectie met ongelijkheidsrestricties, ook gebruikt kan worden in situaties met gelijkheidsrestricties. Dit doen we door te kijken naar de ratio van de hoogte van de posterior en de prior verdeling, op het punt van gelijkheid (de Savage-Dickey ratio). We laten zien dat de EP benadering een veralgemenisering is van de Savage-Dickey ratio methode, en dat de EP benadering dus gebruikt kan worden voor zowel gelijkheids– als ongelijkheidsrestricties. Deze algemene EP benadering is een computationeel efficiënte methode om Bayes factors uit te rekenen voor geneste modellen. Echter, de EP benadering voor gelijkheidsrestricties leidt tot de Borel-Kolmogorov paradox.

In *Hoofdstuk vier* stellen we een Bayesiaanse hypothesetoets voor, om te toetsen op de aan– of afwezigheid van correlaties of partiële correlaties. Deze toets is een toepassing van Bayesiaanse technieken die gebruikt worden voor het selecteren van variabelen in regressiemodellen. We illustreren het gebruik van de tests door middel van drie voorbeelden uit de psychologische literatuur.

In *Hoofdstuk vijf* presenteren we een Bayesiaanse hypothese toets voor variantie analyse (ANOVA). We illustreren wat er gebeurt bij de toepassing van verschillende g-priors op de ANOVA hypothesetoets. Vervolgens illustreren we de test ook aan de hand van twee voorbeelden.

Deel II: Bayesiaanse Model Selectie: Toegepast

In het tweede gedeelte bespreken we hoe Bayesiaanse methodes het empirisch onderzoek in de sociale wetenschappen vooruit kan helpen.

Empirisch onderzoek in de psychologie heeft altijd erg veel gebruik gemaakt van frequentistische toetsen, die gedreven zijn door p waardes. Deze manier van toetsen en de conclusies die hieruit getrokken worden krijgen al geruime tijd kritiek. Een oplossing voor de problemen van hypothesetoetsen met p waardes is om naast p waardes ook effectgroottes te vermelden. Een andere oplossing is om p waardes te vervangen door Bayes factors. In *Hoofdstuk zes* vergelijken we p waardes, effectgroottes en Bayes factors met elkaar.

Hiervoor gebruiken we de resultaten van 855 recentelijk gepubliceerde t toetsen uit de psychologische literatuur. Onze vergelijking laat twee hoofdresultaten zien. Ten eerste, hoewel p waardes en standaard Bayes factors bijna altijd in overeenstemming zijn over welke hypothese er beter wordt ondersteund door de data, is er vaak geen overeenstemming over de overtuigingskracht van dit bewijs. Voor 70% van de data die een p waarde opleverde tussen de .01 en .05, gaf de standaard Bayes factor aan dat het bewijs niet erg overtuigend was. Ten tweede, effectgroottes kunnen extra informatie geven, zowel naast de Bayes factor als naast de p waarde.

Het volgende hoofdstuk, *Hoofdstuk zeven*, is een reactie op een controversieel artikel waarin wordt beweerd dat mensen in de toekomst kunnen kijken. In dit controversiële artikel deed Dr. Bem negen studies, waarin hij meer dan duizend proefpersonen onderzocht om te onderzoeken of gebeurtenissen in de toekomst het heden kunnen beïnvloeden. In dit hoofdstuk bespreken we verschillende tekortkomingen van deze studies. We laten zien dat de data-analyse gedeeltelijk exploratief was, en dat eenzijdige p waardes het bewijs tegen de nulhypothese kunnen overschatten. We heranalyseren de data door een standaard Bayesiaanse t test te gebruiken en laten zien dat het bewijs voor paranormale gaven niet of nauwelijks aanwezig is. We beargumenteren ook, dat om een sceptisch publiek te kunnen overtuigen van een dergelijke controversiële claim, confirmatief onderzoek cruciaal is. Daarbij is het gebruik van conservatieve toetsen (toetsen die niet te snel bewijs voor een effect leveren) belangrijk. Tot slot concluderen we dat de p waardes van Bem geen bewijs leveren voor de stelling dat mensen in de toekomst kunnen kijken. Ze zijn daarentegen wel een indicatie dat psychologen de manier waarop zij hun onderzoeken uitvoeren, en de manier waarop zij hun data analyseren, moeten veranderen.

In het laatste hoofdstuk van deze dissertatie, *Hoofdstuk acht*, bespreken we het doen van confirmatief onderzoek. Het waarheidsgehalte van claims die gedaan worden hangt af van de manier waarop data is verzameld en geanalyseerd. In dit hoofdstuk benadrukken we twee ongemakkelijke feiten die een bedreiging vormen voor wat wij zien als de kern van wetenschappelijk onderzoek. Het eerste feit is dat psychologen hun data-analyse over het algemeen niet vastleggen voordat ze hun data gezien hebben. Daardoor wordt het heel verleidelijk om de analyse aan te passen aan de verzamelde data. Deze gang van zaken maakt de uiteindelijke analyse erg lastig te interpreteren, want de mate waarin er geëxploreerd is, is voor reviewers of lezers niet in te schatten. Het tweede feit is dat de p waarde het bewijs tegen de nulhypothese overschat, en dat het gebruik van frequentistische toetsen (met p waardes) ook voor inflexibiliteit zorgt bij het verzamelen van data. We stellen voor dat onderzoekers hun studies van tevoren centraal aanmelden, en dat ze ook van tevoren aangeven welke analyses er uitgevoerd zullen gaan worden. Deze analyses zijn dan de enige analyses die het predicaat confirmatief mogen dragen, en alleen voor deze analyses zijn de standaard toetsen valide. Alle andere analyses krijgen het predicaat exploratief. Daarnaast stellen we voor dat onderzoekers gebruik maken van Bayes factors in plaats van p waardes bij het uitvoeren van een nulhypothesetoets. Bayes factors staan het tussentijds evalueren van de resultaten toe; een wetenschapper mag stoppen met data verzamelen als ze vindt dat haar punt gemaakt is en de data haar hypothese voldoende onderbouwen.

Deel III: Appendices

Bayesiaanse methodes kunnen ook erg bruikbaar zijn zonder dat er gekeken wordt naar Bayes factors. Om te illustreren hoe onderzoekers Bayesiaanse statistiek kunnen gebruiken om hun data te modelleren staan in de appendix onder andere twee hoofdstukken

die laten zien hoe mathematische modellen geëvalueerd kunnen worden met gebruikmaking van Bayesiaanse statistiek.

In *Appendix A* onderzoeken we de statistische eigenschappen van het zogenaamde Expectancy Valence (EV) model. We laten zien dat de resultaten van het model moeilijk te interpreteren zijn op het niveau van het individu. Vervolgens stellen we een hiërarchische extensie voor, die we ook implementeren. Dit model combineert op een coherente manier informatie van verschillende individuen om tot een goede schatting te komen. Als laatste passen we dit model toe op data van een experiment die de interpretatie van de EV parameters onderzoekt en valideert.

De laatste jaren is de populariteit van Bayesiaanse data-analyses enorm toegenomen, dat heeft onder andere te maken met de WinBUGS software. Deze software is gratis verkrijgbaar en stelt de gebruiker in staat om statistische modellen eenvoudig te implementeren. Echter, voor complexere psychologische procesmodellen kan het prettig zijn, en soms noodzakelijk, om zelf functies en verdelingen toe te voegen aan wat er al beschikbaar is in WinBUGS. Deze functionaliteit is beschikbaar via de WinBUGS Development Interface (WBDev). *Appendix B* illustreert het gebruik van WBDev door middel van voorbeelden zoals de implementatie van het EV model, als ook de shifted Wald verdeling die gebruikt wordt voor reactietijd taken.

Vervolgens presenteren we in *Appendix C en D* de R scripts om de Bayes factors voor de correlatie, de partiële correlatie en de ANOVA hypothesetoets uit te rekenen.

Daarna, in *Appendix E*, kijken we terug op de controversiële studie uit *Hoofdstuk zeven*. We bekijken hoe gevoelig de Bayesiaanse t test is als we andere prior verdelingen gebruiken. Wij laten zien dat andere zinvolle prior verdelingen geen andere kwalitatieve resultaten genereren dan de resultaten die de standaard berekening lieten zien. Dus zijn onze initiële conclusies niet gevoelig voor het gebruik van andere prior verdelingen.

Als laatste presenteren we in *Appendix F* de resultaten van het confirmatieve onderzoek naar paranormale gaven uit *Hoofdstuk acht*. Alle toetsen genereren bewijs in het voordeel van de nulhypothese. Anders gezegd, alle confirmatieve experimenten genereren bewijs *tegen* de hypothese dat mensen de toekomst kunnen voorspellen.

Dankwoord

Als eerste wil ik graag mijn twee promotoren, Eric-Jan en Han, bedanken voor alle begeleiding tijdens mijn studie– en promotietijd. Daarbij is Eric-Jan, mijn directe begeleider, het belangrijkst geweest. Ik denk niet dat er ooit iemand is geweest met wie ik het vaker oneens ben geweest dan met Eric-Jan. Maar, omdat Eric-Jan soms toch wel gelijk had, en heel misschien wel vaak gelijk had, denk ik ook niet dat er ooit iemand is geweest van wie ik zo veel heb geleerd als van Eric-Jan. Daarnaast hebben we natuurlijk ook gewoon veel gelachen, bier gedronken, en hard gewerkt. EJ, dat mijn promotietijd een geweldige tijd is geweest, heb ik voor een groot deel aan jou te danken. Je bent een inspirerende man, ik zal het samenwerken met jou erg gaan missen.

Mijn dank gaat ook uit naar de overige leden van mijn promotiecommissie, voor het lezen en beoordelen van mijn proefschrift: Paul de Boeck, Denny Borsboom, Jean-Paul Fox, Francis Tuerlinckx, Wolf Vanpaemel en Lourens Waldorp.

Vervolgens wil ik mijn collega's van de afdeling Psychologische Methodenleer bedanken voor alle discussies over psychologie, statistiek, filosofie, en de combinatie van die drie. Ik heb er vaak veel van geleerd en altijd van genoten. Ik ben, toen ik eigenlijk een beetje per toeval bij Methodenleer terecht kwam, met mijn neus in de boter gevallen.

Erg belangrijk zijn mijn vrienden geweest, van wie verreweg het grootste gedeelte niet in de wetenschap zit. Er is niets fijners dan de realitycheck die ik altijd kreeg bij het besef dat iedereen direct in slaap viel als ik ook maar een minipoging deed het over mijn werk te hebben. Bedankt daarvoor, maar ook voor de vele avonden in de kroeg, de oneindige stroom goede slechte grappen, en voor het zijn van heel goed gezelschap. Ik ben erg dankbaar dat ik jullie heb ontmoet en als vrienden heb, dat is voor mij ontzettend waardevol.

Natuurlijk is mijn familie altijd heel belangrijk geweest. Niet in het bijzonder tijdens mijn promotie als wel daarvoor, tijdens en zonder twijfel ook daarna. Bedankt mam, pap, Rik, Ellen, Joep, Oscar, Marijntje, en Manfred, dat jullie er zijn.

Als laatste maar allerbelangrijkste, bedankt Sara. Jij bent, sinds kort samen met onze kleine Caspar, het mooiste, liefste, grappigste en meest dierbare in mijn leven.