Quality & Quantity (2007) 41:73–91
DOI 10.1007/s11135-005-6224-6

Comparison of Hypothesis Testing and Bayesian Model Selection

HERBERT HOIJTINK* and IRENE KLUGKIST
Department of Methodology and Statistics, University of Utrecht, P.O.Box 80140, 3508 TC Utrecht, The Netherlands

Abstract. The main goal of both Bayesian model selection and classical hypotheses testing is to make inferences with respect to the state of affairs in a population of interest. The main differences between both approaches are the explicit use of prior information by Bayesians, and the explicit use of null distributions by the classicists. Formalization of prior information in prior distributions is often difficult. In this paper two practical approaches (encompassing priors and training data) to specify prior distributions will be presented. The computation of null distributions is relatively easy. However, as will be illustrated, a straightforward interpretation of the resulting p-values is not always easy. Bayesian model selection can be used to compute posterior probabilities for each of a number of competing models. This provides an alternative for the currently prevalent testing of hypotheses using p-values. Both approaches will be compared and illustrated using case studies. Each case study fits in the framework of the normal linear model, that is, analysis of variance and multiple regression.

Key words: Bayesian model selection, encompassing prior, posterior model probability, p-value, training data

1. An Introduction to Hypothesis Testing and Bayesian Model Selection

Testing of hypotheses using p-values (see Bayarri and Berger, 2000, for a comprehensive overview) is a statistical tool that is used quite often in psychological research. The p-value can formally be defined as:

$$P(T(\mathbf{Y}) \geq t(\mathbf{y}) \mid H_0), \tag{1}$$

where $\mathbf{y}$ denotes the observed data, and $t(\mathbf{y})$ a test statistic (e.g. the student t-test or Pearson chi-square statistic) computed for the observed data. Furthermore, $\mathbf{Y}$ denotes a data matrix from the null-population described by the null-hypothesis H_0, and $T(\mathbf{Y})$ the test statistic computed for this data matrix. The probability in Equation (1) is computed over the distribution of $T(\mathbf{Y})$ under H_0. This implies that the p-value represents the proportion of $T(\mathbf{Y})$ resulting from H_0 that is larger than $t(\mathbf{y})$. According to

*Author for correspondence: E-mail: h.hoijtink@fss.uu.nl

a popular rule, a p-value smaller than 0.05 is a strong indication that H_0 should be rejected. The motivation for the latter (although the choice of the reference value 0.05 is arbitrary) is that small p-values imply that $t(\mathbf{y})$ is rarely observed if H_0 is true. Usually, the specification of the alternative hypothesis is rather vague, that is, not H_0. However, as will be illustrated later, H_0 can also be tested against a specific H_1. The latter is accomplished using a test statistic that has power against the alternative hypothesis of interest.

Developments in Bayesian statistics (see Kass and Raftery, 1995, for a comprehensive overview and references) have rendered a framework for model selection that can be used as an alternative for the testing of hypotheses. Let M_m for $m \in \{0, 1\}$ denote the null and alternative model, respectively. Bayes theorem states that

$$P(M_m|\boldsymbol{y}) = \frac{P(\boldsymbol{y}|M_m)P(M_m)}{P(|\boldsymbol{y})}, \tag{2}$$

where,

$$P(|\boldsymbol{y}) = \sum_m P(\boldsymbol{y}|M_m)P(M_m), \tag{3}$$

$P(M_m|\boldsymbol{y})$ is the posterior probability, that is after observing the data, of model m, $P(M_m)$ is the prior probability, that is before observing the data, of model m, and $P(\boldsymbol{y}|M_m)$ is the density of the data for model M_m.

Posterior model probabilities can be used to compare both nested and non-nested models, and in doing so incorporate the complexity of a model (Ockham's razor). The posterior probabilities of the null and alternative model constitute an alternative for the p-value resulting from testing a hypothesis. A difference between p-values and posterior probabilities is that the latter are computed for all models under consideration, and not only the null-model. This leads to a straightforward interpretation of posterior probabilities. Two models with (about) the same posterior probability, are (after observing the data) about equally probable. However, if one model has a posterior probability of 0.90 and the other of 0.10, the first is preferred to the latter.

'Level alpha' and 'level beta' are classical terms used to denote the probability of an error of the first and second kind, respectively, that is, the error probabilities *before* observing the data. Posterior probabilities can be seen as conditional error probabilities (Cohen, 1994; Sellke et al., 2001), that is, *after* observing the data. The conditional error of the first kind is the probability that M_0 is incorrectly rejected after observing the data. The latter is equal to the probability of M_0 after observing the data. Similarly $P(M_1|\boldsymbol{y})$ can be interpreted as the conditional error of the second kind.

The goal of this paper is to introduce Bayesian model selection, to compare it to hypothesis testing, and to give illustrations of its use in psychological research. In the next section two simple hypotheses and models are used to provide a first comparison of hypothesis testing and model selection. In the two sections that follow, both approaches will be illustrated and compared in the context of one-way analysis of variance followed by pairwise comparisons of means, and, selection of the best of a number of non-nested regression models. The paper will be concluded with a discussion.

2. A First Comparison of Hypothesis Testing and Bayesian Model Selection

Consider the data $\boldsymbol{y} = [0, 0.5, 1.5, 2]$ i.e. the sample size N is four. Assume that the data come from a normal population with variance $\sigma^2 = 1$ and unknown mean μ. The common null hypothesis is $H_0 : \mu_0 = 0$ the corresponding alternative hypothesis is $H_1 : \mu_1 \neq 0$. The null and alternative models are $M_0 : \mu_0 = 0$ and $M_1 : \mu_1 \neq 0$. The null hypothesis can be tested using the Z-test with $t(\boldsymbol{y}) = \frac{\bar{y} - \mu_0}{\sigma/\sqrt{N}}$, where $\bar{y}$ denotes the sample average of $\boldsymbol{y}$. The resulting (two-sided) p-value is 0.048, and indicates that the observed value of the test statistic is not very common if H_0 is true. Frick (1996) and Wainer (1999) present examples where it is sufficient to determine whether data are consistent with H_0 or not, and, where the specification of H_1 beyond 'not H_0' is irrelevant. If this is all a researcher wants to know, p-values are an excellent tool to provide answers to the research questions. In fact, in this situation posterior probabilities are *not* an alternative for the use of p-values, because specification of H_1 beyond 'not H_0' is necessary. Stated otherwise, the (prior) distribution of μ_1 under H_1 has to be specified in order to be able to compute posterior probabilities (this topic will return in Sections 3 and 4).

Often researchers are not satisfied with the p-value based conclusion that the observed data are not in agreement with the null-population. They also want to know which of the many alternatives for which $\mu_1 \neq 0$ are in the ball park. Implicitly, they want to know whether the distance between μ_1 and μ_0 is so large that it is relevant, that is, interesting from a practical or scientific point of view. Since p-values only indicate whether data are likely given H_0 or not (an illustration follows below), effect size measures are usually used as a tool to translate the p-value based conclusion 'not H_0' into the research conclusion *relevant* H_1. This can be seen as an informal (it involves a subjective evaluation of the effect size) attempt to quantify the evidence that the data come from a relevant H_1. In Section 3.4 it will be shown that the set of models can contain formal counterparts of the informal *relevant* H_1.

The following situation illustrates that p-values only indicate whether or not data are likely given H_0, and that posterior probabilities consider both M_0 and M_1. The data and the null model and hypothesis are the same as before. However, $H_1 : \mu_1 = 2$ and $M_1 : \mu_1 = 2$. A test statistic that has power against the alternative hypothesis is the likelihood ratio statistic

$$t(\boldsymbol{y}) = -2 \log \frac{P(\boldsymbol{y}|H_0)}{P(\boldsymbol{y}|H_1)}, \tag{4}$$

where $P(\boldsymbol{y} \mid H_m) = P(\boldsymbol{y} \mid M_m)$ as given in Equation (5). Under H_0, $T(\boldsymbol{Y})$ has a $N(-16, 64)$ null distribution. For the data at hand $t(\boldsymbol{y}) = 0$, the corresponding one-sided p-value (1) is 0.024. The choice for a one-sided test is motivated by the specification of the alternative hypothesis: it would be awkward to choose the alternative if $\overline{y}$ is, for example, -2, which in a two-sided setup would lead to a small p-value.

Using (2) the posterior model probabilities can be computed. For the setup at hand

$$P(\boldsymbol{y} \mid M_m) = \prod_{i=1}^{4} \frac{1}{\sqrt{2\pi}} \exp -\frac{1}{2}(y_i - \mu_m)^2. \tag{5}$$

The necessity to specify prior model probabilities is a step typical for Bayesian model selection. One option is to use a non-informative prior: $P(M_1) = P(M_2) = 0.5$. The implication is that differences in posterior model probabilities are caused by differences in the degree to which the data support the model, that is, $P(\boldsymbol{y} \mid M_m)$ for $m \in \{0, 1\}$. The resulting posterior probabilities are 0.5 for both the null and the alternative model i.e. after observing the data both models are equally likely.

The conclusions obtained using p-values and posterior probabilities are strikingly different. Loosely speaking the first rejects H_0, while the latter does not prefer M_0 over M_1. Since $\overline{y} = 1$ is nicely centered between the population means of the null and alternative hypothesis, the conclusion based on the posterior probabilities is correct. This simple example illustrates a drawback of p-values: the size of a p-value may point to H_1, not because the data are likely given H_1, but because the data are unlikely given H_0. This pitfall can to some extend be avoided if the evaluation of a p-value includes an evaluation of the effect size (here $\overline{y} = 1$) in relation to what is specified by H_0 and H_1. However, although the resulting approach is useful and valuable, it is not longer a formal testing procedure since it includes a subjective evaluation of the effect size. Posterior probabilities are computed for each model in the set of models under consideration. Using posterior probabilities the pitfall that 'not H_0' does not necessarily imply H_1 is avoided.

The computation of posterior probabilities is not completely unproblematic. The bottle-neck is the specification of prior distributions (for each model) for the parameters that are not fixed at a specific value. The main criticism is that these specifications are subjective, researcher dependent, and thus, that the conclusions are subjective and researcher dependent (Sober, 2002, pp. 22–24; Howson, 2002, pp. 53–61). In the next two sections two approaches to obtain prior distributions that are not based on (subjective) prior knowledge will be outlined. The first approach will be illustrated using a one-way analysis of variance followed by pairwise comparisons of means. Encompassing priors that are non-informative with respect to the untransformed parameters of the one-way model will be formulated. The second approach will be illustrated using the selection of the best of a number of non-nested regression models. Here priors will be formulated using training data. Using these examples it will be shown that Bayesian model selection is a viable alternative for hypotheses testing.

3. Encompassing Priors

3.1. ANALYSIS OF VARIANCE: DATA, HYPOTHESES AND p-VALUES

One way analysis of variance is often used in psychological research. A common design is the situation where a researcher has a control group (C), and two experimental groups ($E1$ and $E2$) that differ in the 'treatment' that has been received. If the persons participating in the experiment have been randomized over the three groups, the dependent variable can be the outcome of a test given to the persons after the treatment, or, in a design with a pre-test and a post-test, the gain-score.

In this section a simple one way analysis of variance will be executed. Hypotheses will be formulated and tested, and the resulting p-values will be discussed. In the next section, models will be formulated, and posterior probabilities computed and discussed. The data that will be analyzed come from Toothaker (1993, p. 3). He describes an experiment from developmental psychology: "A total of 42 first-grade students were randomly assigned to one of three groups, $n = 14$ per group. Subjects in the first group were informed that there was another child alone in the next room who had been warned not to climb on a chair. This group was called indirect responsibility (C). In addition to the story told to the subjects in (C), subjects in the second group were informed that when the adult left, they were in charge and to take care of anything that happened. This group was called direct responsibility one ($E1$). Subjects started a simple task, and the adult left the room. Next, there was a loud crash from the next room and a short time of crying and sobbing. In the third group, direct responsibility two ($E2$), subjects had the same instructions as ($E1$), but there were also

Table I. The one way analysis of variance data

Data	C	$E1$	$E2$
	3	3	4
	2	5	3
	1	4	5
	4	3	5
	3	5	3
	2	3	3
	3	4	4
	4	3	4
	4	3	4
	2	5	4
	2	5	5
	3	4	2
	1	1	3
	2	3	1
Mean	2.57	3.64	3.57
SD	1.02	1.15	1.15

calls for help. From behind a one-way mirror, two raters gave ratings on helping behavior. The scale was from 1 (no help) to 5 (went to the next room)." The data from this experiment and per group the sample mean and standard deviation are presented in Table I.

Hypotheses testing usually consists of two steps. In the first step $H_0: \mu_C = \mu_{E1} = \mu_{E2}$ is tested. In the second step three pairwise comparisons are tested: $H_{0a}: \mu_C = \mu_{E1}$, $H_{0b}: \mu_C = \mu_{E2}$ and $H_{0c}: \mu_{E1} = \mu_{E2}$. Since in the second step three hypotheses are tested, the risk of an error of the first kind (incorrectly rejecting one or more of the pairwise null hypotheses) is larger than level alpha. A wide variety of procedures has been developed to control the 'family wise error rate' that is, the probability of one or more errors of the first kind. The interested reader is referred to Toothaker (1993) and Ramsey (2002) for a comprehensive overview. Here the Hayter modification of Fisher's LSD procedure (Ramsey 2002) will be used. This is a protected procedure i.e. the pairwise comparisons are only tested if H_0 is significant.

The p-value of H_0 (obtained using a F-test for the equality of independent means) is smaller than the commonly used alpha-level of 0.05 (see, Table II). Stated otherwise, it is unlikely that the data result from a population described by H_0. As illustrated in Section 2, the conclusion 'not H_0' does not necessarily imply H_1 or '*relevant* H_1'. However, inspection of

Table II. P-values and posterior probabilities

Hypothesis	p-value	Model	Post. prob.	Non sharp
H_0	0.025	M_0	0.028	0.051
H_{0a}	0.015	M_{1a}	0.025	0.045
H_{0b}	0.022	M_{1b}	0.037	0.073
H_{0c}	0.866	M_{1c}	0.610	0.680
		M_2	0.301	0.151

the means in Table I shows that μ_C is quite (subjective evaluation of the authors) different from the other means (about 1 on a scale running from 1 to 5). For the example at hand it seems save to conclude that 'not H_0' implies a '*relevant* H_1'.

Usually, a significant H_0 is followed by pairwise comparisons of the means. The p-values obtained using the Hayter modification of Fisher's LSD procedure are displayed in Table II. As can be seen the mean of the control group is significantly different from the means of both experimental groups. The means of the experimental groups are not significantly different. It seems save to conclude that the mean of the control group is smaller (both significantly and relevantly) than the means of both experimental groups.

Pairwise comparisons among means may lead to inconsistent results (Dayton, 2003). Consider, for example, the situation where H_{0a}, H_{0b} and H_{0c}, have p-values of 0.045, 0.055 and 0.70, respectively. This results in a set of conclusions that can not all be true: $\mu_C \neq \mu_{E1}$, $\mu_C = \mu_{E2}$ and $\mu_{E1} = \mu_{E2}$. There are two drawbacks to the use of p-values (that, as will be illustrated in the next section, are not shared by posterior probabilities) that make it hard to properly deal with inconsistencies: (i) testing non-nested hypotheses is an underdeveloped area of statistics; and, (ii) the use of a fixed alpha-level.

To elaborate on (i). Suppose, that the overall F-test indicates that H_0 is not true, and that the pairwise comparisons indicate that not all the means are different (the situation described in the previous paragraph). This leaves the question which pair of means is the same. However, this question cannot be answered using p-values: none of H_{0a}, H_{0b} and H_{0c} is the obvious null hypothesis; and, comparison of non-nested hypotheses using p-values is not a well-developed area of statistics. As will be illustrated in the next section, comparison of non-nested models using posterior probabilities is straightforward and easy.

To elaborate on (ii). The wish to control the probability of an error of the first kind at a level of 5% leads to the strict rule that only p-values smaller than 0.05 lead to a rejection of the corresponding null hypothesis.

As illustrated above, this may lead to inconsistent results. Common sense dictates that p-values of 0.045 and 0.055 (or, if you like, 0.049 and 0.051) should be evaluated in the same way. However, in practice many researchers (and journals) are 'happy' with 0.045, and 'disappointed' with 0.055. This problem can be avoided if researchers are willing to trade control of the error of the first kind before analysis of the data, for conditional error probabilities, that is, the probability of wrong decisions after observing the data. As will be illustrated in the next section, the latter can be achieved using posterior probabilities without needing an arbitrary criterion like '0.05'.

3.2. MODEL SELECTION, ENCOMPASSING PRIORS AND POSTERIOR PROBABILITIES

Model selection usually proceeds in three steps. First, all the models that are of interest have to be specified. For the situation at hand there are five models (note, that a comma implies that the preceding or succeeding parameter is not equal to one of the other parameters):

$$M_0 : \mu_C = \mu_{E1} = \mu_{E2}$$
$$M_{1a} : \mu_C = \mu_{E1}, \mu_{E2} \quad M_{1b} : \mu_C = \mu_{E2}, \mu_{E1} \quad M_{1c} : \mu_C, \mu_{E1} = \mu_{E2}$$
$$M_2 : \mu_C, \mu_{E1}, \mu_{E2}.$$

The second step in model selection is the specification of prior probabilities for each of the models. As is clear from Equation (2) the size of the posterior model probabilities is influenced by the prior model probabilities. In the examples in this paper, prior probabilities are chosen to be the same for each model. The implication is that differences in posterior model probabilities are caused by differences in the degree to which the data support the model, that is, $P(\boldsymbol{y} \mid M_m)$, for $m \in \{0, 1a, 1b, 1c, 2\}$.

If a model contains unknown parameters the prior distribution of the unknown parameters has to be specified. This is the third step in model selection. It is necessary because if not all parameters are fixed at a specific value, $P(\boldsymbol{y} \mid M_m)$ can only be computed if the prior distribution $g(\cdot)$ of the model parameters is specified:

$$P(\boldsymbol{y} \mid M_m) = \int_{\boldsymbol{\theta}} P(\boldsymbol{y} \mid \boldsymbol{\theta}) g(\boldsymbol{\theta} \mid M_m) d\boldsymbol{\theta}. \tag{6}$$

Note that, $\boldsymbol{\theta}$ denotes the parameters of the model at hand, that is, the three population means and the within group variance σ^2, and that

$$P(\boldsymbol{y} \mid \mu_C, \mu_{E1}, \mu_{E2}, \sigma^2) = \prod_{A=C,E1,E2} \prod_{i \in A} \frac{1}{\sqrt{2\pi}\sigma} \exp \frac{1}{2} \frac{(y_i - \mu_A)^2}{\sigma^2}. \tag{7}$$

Looking at (6), it is clear that the posterior model probabilities (2) depend on the prior distributions of the parameters for each model. In the

sequel it will be explained how prior distributions are chosen in the encompassing prior approach.

Specification of prior distributions is the bottle-neck of Bayesian model selection. The use of subjective prior distributions is often criticized because the resulting inferences are also subjective. See, for example, Sober (2002, pp. 22–24) and Howson (2002, pp. 53–61) for interesting discussions. However, irrespective of whether subjective inference is appreciated or not, the specification of subjective prior distributions is far from easy. Consider, for example, M_2. It is rather common (see for example et al. 1995, pp. 71–76) to use independent normal priors for the population means, and a scaled inverse chi-square prior for the within group variance. However, specifying the means and variances of the normal priors, and the scale and degrees of freedom of the scaled inverse chi-square prior such that these priors represent a researchers subjective prior knowledge, is a task that is beyond most researchers (and statisticians).

Both the criticism and the specification problem motivated a search for prior distributions that are not subjective and easily specified. The basic idea of encompassing priors is to specify a distribution that is non-informative (vague, diffuse) with respect to the untransformed means and variance for the encompassing model, that is, the model within which all other models are nested. Subsequently, the priors for the nested models are derived from the encompassing prior. For the application at hand, M_2 is the encompassing model. It contains four parameters: three population means and a within group variance. The other models can be obtained via restrictions on the parameter space of M_2 and are thus nested within this encompassing model. Using independent distributions for each model parameter, the prior distribution of M_2 can be denoted by $g(\mu_C \mid M_2)g(\mu_{E1} \mid M_2)g(\mu_{E2} \mid M_2)g(\sigma^2 \mid M_2)$. The prior distributions of the nested models follow directly from this encompassing prior:

$$g(\mu_C, \mu_{E1}, \mu_{E2}, \sigma^2 \mid M_m) = \frac{g(\mu_C \mid M_2)g(\mu_{E1} \mid M_2)g(\mu_{E2} \mid M_2)g(\sigma^2 \mid M_2)I_{M_m}}{\int g(\mu_C \mid M_2)g(\mu_{E1} \mid M_2)g(\mu_{E2} \mid M_2)g(\sigma^2 \mid M_2)I_{M_m} d\mu_C, \mu_{E1}, \mu_{E2}, \sigma^2}, \quad (8)$$

where the indicator function I has the value 1 if the argument is true, that is, if the parameter values are in accordance with the restrictions imposed by model M_m, and 0 otherwise.

Using the encompassing prior approach it can be shown that for specific models and diffuse encompassing priors, the model selection is hardly affected by the choice of the prior (Klugkist et al., 2005). This is however not the case for the models considered in this section. Although the exact functional form of the encompassing prior is rather irrelevant, the scale factor has a large effect on the answer. To choose the scale factor for

the encompassing prior, the following criterion is used: the prior should not exclude regions of the parameter space with substantial posterior probability, but it should also not include too many values that are completely out of range for the data at hand. Specification of the encompassing prior is often facilitated if the data are taken into consideration. As can be seen in Table I, the response format is such that all data are in the interval 1–5. This implies that the means of the control and both experimental groups also have to be in this interval. Furthermore, with this response format, σ^2 can not be larger than 4 (if half of the persons in a group responds 1, and the other half 5). Consequently, non-informative priors for the means and variance are obtained if $g(\mu_C \mid M_2) = g(\mu_{E1} \mid M_2) = g(\mu_{E2} \mid M_2) = U[1, 5]$ and $g(\sigma^2 \mid M_2) = U[0, 4]$, where U denotes a uniform distribution with lower and upper bound as indicated. The resulting encompassing prior distribution is proper, the same holds for the prior distributions of the other models that can be derived from the encompassing prior using Equation (8). If encompassing priors are used, Equation (6) is the likelihood $P(\boldsymbol{y} \mid \boldsymbol{\theta})$ averaged over intervals specified by the prior distributions. Stated otherwise, the posterior probabilities are 'averaged likelihoods' transformed to probabilities using Bayes theorem.

After specification of the set of models, prior probabilities and the encompassing prior distribution, the posterior probability of each model can be computed using Equation (2) with $P(\boldsymbol{y} \mid M_m)$ as specified in (6). Due to the integration involved, the computation of (6) is not always easy. The interested reader is referred to Carlin and Chib (1995) and Kass and Raftery (1995) for an inventory of methods for the computation of (6). Newton and Raftery (1994) propose to approximate the integral using importance sampling based on a sample of parameter vectors from both the posterior distribution of the model at hand, and, an imaginary prior distribution. Their method is used for all analyzes executed in this section.

The posterior probabilities for the models under consideration in the example at hand are given in the fourth column of Table II. Models M_0, M_{1a} and M_{1b} have small posterior probabilities, that is, after observing the data it is rather unlikely that these models are true. The posterior probabilities of models M_{1c} and M_2 indicate that there is a posterior probability of about 0.9 (0.61 + 0.30) that the mean of the control group differs from the means of both experimental groups, and, that there is a posterior probability of about 0.3 that all three means are different.

3.3. A COMPARISON OF p-VALUES AND POSTERIOR PROBABILITIES

In this section the limitations of p-values discussed previously will be summarized. It will also be elaborated how these limitations can be overcome using posterior probabilities.

1. As indicated in Section 2 a small p-value does not necessarily imply that H_0 has to be rejected in favor of the alternative hypothesis. As several authors have indicated before, the p-value usually overstates the amount of evidence against the null hypothesis (Berger and Sellke, 1987; Cohen, 1994). The reason for the latter is that $P(T(\boldsymbol{Y}) > t(\boldsymbol{y}) \mid H_0) \neq P(H_0 \mid \boldsymbol{y})$. This phenomenon can also be observed in the example at hand. The p-value for testing H_0 vs. H_1 is 0.025, which is usually considered to be a substantive amount of evidence *against* H_0. However, if only M_0 and M_2 are considered, the posterior probability of the corresponding model is 0.085 (0.028/(0.028+0.301)), which does not imply a straightforward rejection of M_0.
2. In contrast to p-values, the comparison of non-nested models is not a problem if posterior probabilities are used. Looking at Table II it is evident that M_{1c} is superior to M_{1a} and M_{1b}.
3. In contrast to p-values posterior probabilities that differ only slightly in size are not evaluated differently because of the requirement to adhere to a criterion (like an alpha-level of 0.05) that has to be fixed *before* the analysis. Posterior probabilities can be interpreted as conditional error probabilities (Cohen, 1994; Sellke et al., 2001). To elaborate on the latter, consider only M_0 and M_2. The posterior probabilities if only these two models are considered are 0.085 and 0.915, respectively. Stated otherwise, the probability *after* observing the data that M_0 is true is 0.085, that is, the probability of incorrectly rejecting M_0 after observing the data (the conditional error of the first kind) is 0.085. It is interesting that this probability is larger than the common level alpha of 0.05. However, since there is no pre-specified criterion that has to be used to evaluate posterior probabilities, each researcher can make his own decision whether 0.085 is small enough to discard M_0.
4. It is relatively easy to define and evaluate models that are *relevantly* different from each other using posterior probabilities. In the next section this topic will be elaborated and illustrated using the data at hand.

3.4. THE NIL HYPOTHESIS AND NON-SHARP NULL MODELS

As mentioned before, both Frick (1996) and Wainer (1999) present examples where it is sufficient to determine whether data are consistent with H_0 or not. The null hypotheses they consider are of the kind 'an average is zero' and 'the difference between two averages is zero', that is, so called 'nil-hypotheses' (Cohen, 1994), or sharp or point null hypotheses (Berger and Sellke 1987). The examples given by Frick (1996) and Wainer (1999) are convincing. However, there are also examples where the evaluation of a sharp H_0 does not make sense. Cohen (1994) clearly takes the position that 'the difference between two averages' is never zero, and thus, that it

is meaningless to test the corresponding null hypothesis. Apparently many researchers agree with Cohen, because the evaluation of a significant null hypothesis is often supported by the evaluation of an effect size measure in order to be able to replace the conclusion H_1 by *relevant* H_1.

Model M_0 introduced in Section 3.1 could be called a sharp model. Models M_{1a}, M_{1b} and M_{1c} contain 'sharp elements'. The criticism on 'nil-hypotheses' also applies to these models. Whether or not M_2 is *relevantly* better than M_0 can only be determined using effect size measures in addition to the posterior probabilities. However, it also possible to reformulate these models such that effect sizes are included. All a researcher has to do is determine which difference between two means is considered to be relevant. Here a difference of 0.4 (which is 10% of the distance of the scale on which helping behavior is rated) is considered to be relevant. Furthermore, the prior expectation that helping behavior should increase from C via $E1$ to $E2$ is used during the construction of M_{1a}, M_{1c} and M_2. Note that M_{1b} is not in agreement with this prior expectation. This leads to the following counterparts of the models used sofar:

$$M_0 : |\mu_C - \mu_{E1}| \leq 0.4, \quad |\mu_C - \mu_{E2}| \leq 0.4, \quad |\mu_{E1} - \mu_{E2}| \leq 0.4,$$

$$M_{1a} : |\mu_C - \mu_{E1}| \leq 0.4, \quad \mu_{E2} - \mu_C \geq 0.4, \quad \mu_{E2} - \mu_{E1} \geq 0.4,$$

$$M_{1b} : \mu_{E1} - \mu_C \geq 0.4, \quad |\mu_C - \mu_{E2}| \leq 0.4, \quad \mu_{E1} - \mu_{E2} \geq 0.4,$$

$$M_{1c} : \mu_{E1} - \mu_C \geq 0.4, \quad \mu_{E2} - \mu_C \geq 0.4, \quad |\mu_{E1} - \mu_{E2}| \leq 0.4,$$

$$M_2 : \mu_{E1} - \mu_C \geq 0.4, \quad \mu_{E2} - \mu_C \geq 0.4, \quad \mu_{E2} - \mu_{E1} \geq 0.4.$$

Using Equations (2), (6) and (7) the posterior probabilities of these models can be computed. Note that, the prior distributions for the parameters of these models can still be derived using Equation (8), even if the encompassing model itself is no longer part of the models under investigation. As can be seen in the last column of Table II, model M_{1c} 'the average in group C is *relevantly* different from the averages in groups $E1$ and $E2$' has a higher posterior probability than the counterpart discussed in Section 3.2 (one but last column of Table II). Model M_2 'all the averages are relevantly different' has a substantially lower posterior probability. Stated otherwise, M_{1c} has by far the highest posterior probability, but the other models can not completely be ruled out.

Giving a difference between two means of more than 0.4 the label *relevant* is of course subjective. However, nothing prevents other researchers

from using a different number. As long as the threshold value used is reported, the meaning of the corresponding posterior probabilities is clear. Comparison of hypotheses corresponding to the five models using p-values can, as far as the authors know, not be found in the literature. The main problems here are null models where the parameters of interest are *not* fixed at zero, and, a test-statistic that has power against the alternative hypotheses specified. A point of departure to solve these problems might be developments in testing hypotheses with inequality constraints among the parameters. The interested reader is referred to Robertson et al. (1988), who give a comprehensive overview of order restricted inference, that is, hypotheses that are more informative than the traditional null and alternative hypotheses.

4. Training Data

4.1. MULTIPLE REGRESSION

Stevens (1992, pp. 578–585) describes the Sesame street data. These data contain measurements of the knowledge of body-parts, numbers, forms and the like, before and after the first year of the television program, for 240 children in the age range between three to five. The research question of interest in this section is whether knowledge of letters after watching Sesame street for a year is better predicted using knowledge of letters (a topic that receives a lot of attention in the series) before watching Sesame street, using the Peabody picture vocabulary test which is not related to topics presented in Sesame street, or both.

The research question at hand is representative for a class of questions that is often encountered: is one set of predictors superior to another set, or, should the sets be combined (see for example Congdon 2001, pp. 139–142, and Tabachnick and Fidell, 2001, pp. 131–139). The main results of straightforward multiple regression analyzes are presented in Table III. Not the whole data matrix was used, but $N = 122$ randomly selected children. This part of the data matrix will subsequently be called the calibration sample. Irrespective of the model used, all predictors have a positive relationship with knowledge of numbers after watching Sesame street for a year. The models with one predictor explain, 34% (*Letters*) and 27% (*Peabody*) of the variation of the dependent variable, the model with two predictors explains 43%. For each model the null hypothesis that the predictors can not be used to predict the dependent variable is rejected (each p-value is 0.00).

Two questions remain: are both one predictor models about equally good, or, is one better than the other; and, is the two predictor model better than both one predictor models, or, is the increase in explained

Table III. Analyzing three models using multiple regression and training data

Predictor	Letters	Peabody	Both
Constant	12.41	5.40	2.52
Regr. Coeff. Letters	0.83		0.63
Regr. Coeff. Peabody		0.46	0.29
R^2	0.34	0.27	0.43
p-value	0.00	0.00	0.00
R^2-Cross Validated	0.27	0.21	0.35
$P(M_m \mid x)$	0.001	0.000	0.999
$P(M_m \mid x)$	0.98	0.02	

variation not that impressive considering the fact that the model contains an extra predictor compared to the other models. For the example at hand, the second question could be investigated via the testing of nested regression models (Tabachnick and Fidell, 2001, pp. 165–170). However, in related but more elaborate situations (more than two sets of predictors, and various combinations of these sets) this does not provide a solution since mostly non-nested models have to be compared. Even for the example at hand (two non-nested models with one predictor) this is not easily done via hypothesis testing. It is not clear which are the null and alternative model, and, the actual testing of non-nested models is an underdeveloped area in statistics.

Nevertheless, both questions can be addressed using non-Bayesian methods via cross-validation (Camstra and Boomsma 1992; Stevens 1992, pp. 96–98). There are several kinds of cross-validation. In the simplest form the data matrix is randomly split into two parts: the calibration sample and the validation sample. The calibration sample is used to estimate the regression coefficients of each model of interest. Subsequently, the validation sample (here $N = 118$) is used to compute for each model the squared correlation between the dependent variable and its predicted value using the regression equations obtained in the calibration sample. The result is called the cross-validated proportion of variance explained. It is a measure of the predictive performance of a model that is not biased by the number of predictors in the model, or the step-wise or other 'trial and error' methods by which models are constructed and the regression coefficients estimated. As can be seen in Table III, the cross-validated R^2 of the model with both predictors is 0.08 higher than the model containing only Letters, and 0.14 higher than the model containing only Peabody. It appears (based on our subjective evaluation of the increase in cross-validated R^2) that the model with two

predictors gives better predictions than the models with one predictor. The decision whether the model containing only Letters is relevantly better than the model containing only Peabody is left to the reader.

Possibly the only drawback of the cross-validated R^2 is the lack of a formal comparison of the models. Questions like is one model 'significantly better' or 'more probable' than the others, have not yet been addressed. In the next section it will be explained how these questions can be answered using posterior model probabilities.

4.2. POSTERIOR PROBABILITIES COMPUTED USING TRAINING DATA

To compute posterior probabilities, for each model the prior distribution of each parameter that is not fixed at a specific value has to be specified. One of the goals of this paper is to show that prior distributions can be specified such that they do not represent the subjective view of the researcher, but nevertheless are useful and informative. As illustrated in Section 3, encompassing priors are one way to achieve this. In this section training data (Berger and Pericchi, 1996) or, if you like, the calibration sample, will be used to specify objective priors.

One of the interpretations given to prior knowledge is that it should reflect the current state of affairs with respect to the research questions and models at hand. It is rare (if it happens at all) that previous research can be used to specify prior distributions for the parameters of the models of interest. However, it does occur that researchers have enough data to randomly assign each person to a training and a validation set. If cross-validation is used, training (calibration) data are used to construct and estimate the parameters of multiple regression models. If the goal is to compute posterior model probabilities, training data are used to construct and summarize the information with respect to the parameters of multiple regression models. This summary is the *posterior* distribution of the model parameters for the *training data*, which will serve as the *prior* distribution of the model parameters for the *validation* data. Stated otherwise, the training data are used to summarize the knowledge with respect to the current state of affairs for each of the models under consideration, and the validation data are used to compute posterior probabilities.

For the multiple regression model

$$P(\boldsymbol{y}|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_P, \beta_0, \ldots, \beta_P, \sigma^2) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp -\frac{1}{2} \frac{(y_i - \beta_0 - \sum_{p=1}^{P} \beta_p x_{ip})^2}{\sigma^2}, \tag{9}$$

where x_{ip} denotes the response of the ith person to the pth predictor, y_i the response to the criterion variable, β_p the regression coefficient of the

pth predictor, and σ^2 the residual variance. Subsequently, the superscripts t and v will be used to denote prior distributions and data for the training and validation data set, respectively.

Different regression models are obtained using different subsets of the total set of predictors. For each model it is known which predictors are included. In the sequel M_m will be used to denote the set of predictors included for model m. More specifically, $M_1 = \{Letters\}$, $M_2 = \{Peabody\}$ and $M_3 = \{Letters, Peabody\}$. The parameters of these models will be denoted by $\boldsymbol{\theta}_m = \{\beta_{0m}, \ldots, \beta_{Pm}, \sigma_m^2\}$, where P denotes the number of predictors in the model at hand.

In the Bayesian approach the information with respect to the parameters of a model is summarized in the posterior distribution (Gelman et al. 1995, pp. 32–33):

$$\text{Post}(\boldsymbol{\theta}_m \mid \boldsymbol{y}^t, M_m^t) \propto P(\boldsymbol{y}^t \mid M_m^t, \boldsymbol{\theta}_m) g^t(\boldsymbol{\theta}_m), \tag{10}$$

which combines the information provided by the training data with the prior information with respect to the model parameters that is available *before* the *training* data are analyzed. Because it is in accordance with the goal to compute posterior probabilities that are independent of subjective choices, a non-informative prior will be used. For multiple regression models non-informative prior distributions can be chosen in various ways (Gelman et al. 1995, Chap. 8). In this paper the improper non-informative prior $g^t(\boldsymbol{\theta}_m) = 1$ will be used. This renders a proper posterior distribution that is proportional to the likelihood of the data.

The posterior distributions constructed using the training data represent the prior knowledge with respect to the current state of affairs before the *validation* data are used to compute posterior probabilities for each model in the set of models under consideration, that is, $g^v(\boldsymbol{\theta}_m) = \text{Post}(\boldsymbol{\theta}_m \mid \boldsymbol{y}^t, M_m^t)$. These, resulting from 'previous research' and thus objective prior distributions can be used to compute (6) for the multiple regression model, that is,

$$P(\boldsymbol{y}^v \mid M_m^v) = \int_{\boldsymbol{\theta}_m} P(\boldsymbol{y}^v \mid M_m^v, \boldsymbol{\theta}_m) g^v(\boldsymbol{\theta}_m)\, d\theta_m. \tag{11}$$

Combined with equal prior probabilities for each of the models, this is sufficient to compute objective posterior probabilities using Equation (2):

$$P(M_m^v \mid \boldsymbol{y}^v) = \frac{P(\boldsymbol{y}^v \mid M_m^v)}{\sum_m P(\boldsymbol{y}^v \mid M_m^v)}. \tag{12}$$

The integral in Equation (11) can be evaluated if the integral with respect to $\boldsymbol{\theta}_m$ is replaced by a summation over a sample $\boldsymbol{\theta}_m^q$, for $q = 1, \ldots, Q$ from $g^v(\boldsymbol{\theta}_m)$, that is,

$$P(\boldsymbol{y}^v \mid M_m^v) \approx \frac{1}{Q} \sum_{q=1}^{Q} P(\boldsymbol{y}^v \mid M_m^v, \boldsymbol{\theta}_m^q). \tag{13}$$

This sample can be obtained using the Gibbs-sampler. The interested reader is referred to Gelman et al. (1995, pp. 235–239, 326–329).

As can be seen in Table II, if all three models are compared the model with both predictors has by far the largest posterior probability. Since posterior probabilities implicitly penalize models with more parameters (Ockham's razor, see Kass and Raftery (1995)) the result is not just caused by the fact that an extra predictor usually leads to an increased amount of variance explained in the sample. It is caused by the fact that the extra predictor leads to an increased amount of variance explained in the population. If only both one-predictor models are compared (the last line of Table III), M_1 has a substantially higher posterior probability than M_2. Stated otherwise, after observing the data, the (conditional error) probability that M_2 is the best model is so small, that it is save to conclude that *Letters* is a better predictor than *Peabody*.

In this section it was illustrated that prior distributions can be specified using training data. The interested reader is referred to Berger and Pericchi (1996), who show that this procedure can be refined in various ways. One way is to split a data file in a training and validation sample not once, but many times and combine the results. The other way is to adjust the former such that the training data contain as few persons as possible. In this section it was shown that training data based posterior probabilities can be used in addition to cross-validation, to select the best of a number of non-nested regression models. As far is known to the authors, there are no onsets in the literature to do the same using hypothesis testing.

5. Summary and Remaining Issues

The goal of this paper is to show the potential of posterior probabilities and model selection as an alternative for the use of p-values and hypotheses testing. As was illustrated using a simple model for the evaluation of a population mean, a one-way analysis of variance followed by pairwise comparisons of means, and multiple regression, posterior probabilities have a number of advantages over p-values:

1. Posterior probabilities are computed for each of the models in the set of models under consideration. This avoids (possibly incorrect) indirect conclusions of the kind 'not H_0' implies 'H_1'.
2. Posterior probabilities can be used to compare non-nested models, and in doing so incorporate the complexity (number of parameters) of a model (Ockham's razor).

3. Posterior probabilities do not need arbitrary criteria like 0.05 in order to be evaluated.
4. Posterior probabilities can be interpreted as conditional error probabilities, that is, the error probabilities *after* observing the data.
5. Posterior probabilities can be used to evaluate models that are *relevantly* different from each other.

A difficulty using posterior probabilities is that for each model the prior distribution of the parameters of that model has to be specified. This paper illustrated two methods that can be used to derive priors that are easy to apply.

Unlike hypotheses testing and p-values, posterior probabilities have not yet found their way into the toolkit of psychological researchers. Only in the last decade (see for example Carlin and Chib 1995; Kass and Raftery 1995; Berger and Pericchi 1996) computational developments have enabled the calculation and use of posterior probabilities. Research with respect to application of posterior probabilities is still ongoing, and the standard software packages used by psychological researchers have not yet included modules and models for which posterior probabilities can be computed. Readers interested in using the methods proposed, can write an e-mail to the authors describing their data and research questions.

References

Bayarri, M. J. and Berger, J. O. (2000). *P*-values for composite null models. *Journal of the American Statistical Association* 95: 1127–1142.

Berger, J. O. and Perricchi, L. (1996). The intrinsic bayes factor for model selection prediction. *Journal of the American Statistical Association* 9: 109–122.

Berger, J. O. and Sellke, T. (1987). Testing a point null hypothesis: the irreconcilability of p-values and evidence. *Journal of the American Statistical Association* 82: 112–122.

Camstra, A. and Boomsma, A. (1992). Cross-validation in regression and covariance structure analysis. *Sociological Methods and Research* 21: 89–95.

Carlin, B. P. and Chib, S. (1995). Bayesian model choice via Markov Chain Monte Carlo methods. *Journal of the Royal Statistical Society, B*, 57: 473–484.

Congdon, P. (2001). *Bayesian Statistical Modelling*. New York: John Wiley and Sons.

Cohen, J. (1994). The earth is round ($p < 0.05$). *American Psychologist* 12: 997–1003.

Dayton, C. M. (2003). Information criteria for pairwise comparisons. *Psychological Methods* 8: 61–71.

Frick, R. W. (1996). The appropriate use of null hypothesis testing. *Psychological Methods* 1: 379–390.

Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (1995). *Bayesian Data Analysis*. London: Chapmann and Hall.

Howson, C. (2002). Bayesianism in statistics. In: R. Swinburne (ed.), *Bayes Theorem*, Oxford: Oxford University Press, pp. 39–69.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association* 90: 773–795.

Klugkist, I., Kato B. and Hoijtink, H. (2005). Bayesian model selection using encompassing priors. *Statistica Neerlandica* 59: 57–69.

Newton, M. A. and Raftery, A. E. (1994). Approximate Bayesian inference by the weighted likelihood bootstrap. *Journal of the Royal Statistical Society B* 56: 3–48.

Ramsey, P. H. (2002). Comparison of closed testing procedures for pairwise testing of means. *Psychological Methods* 7: 504–523.

Robertson, T., Wright, F. T. and Dykstra, R. L. (1988). *Order Restricted Statistical Inference*. New York: John Wiley and Sons.

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

Sober, E. (2002). Bayesianism–its scope and limits. In: R. Swinburne (ed.). *Bayes Theorem*, pp. 21–38. Oxford: Oxford University Press.

Stevens, J. (1992). *Applied Multivariate Statistics for the Social Sciences*. London: Lawrence Erlbaum.

Tabachnick, B. G. and Fidell, L. S. (2001). *Using Multivariate Statistics*. London: Allyn and Bacon.

Toothaker, L. E. (1993). *Multiple Comparison Procedures.* London: SAGE.

Wainer, H. (1999). One cheer for null hypothesis significance testing. *Psychological Methods* 4: 212–213.