Tutorials in Quantitative Methods for Psychology
2012, Vol. 8(2), p. 122-126.
DOI: 10.20982/tqmp.08.2.p122

A short introduction into Bayesian evaluation of informative hypotheses as an alternative to exploratory comparisons of multiple group means

Sébastien Béland, *Université du Québec à Montréal*
Irene Klugkist, *University of Utrecht*
Gilles Raîche, *Université du Québec à Montréal*
David Magis *Université of Liège*

This paper presents an introduction into Bayesian evaluation of informative hypotheses, that is, hypotheses representing explicit expectations about multiple group means (Hoijtink, 2011; Hoijtink, Klugkist & Boelen, 2008). The authors begin by discussing some limits of exploratory methods before presenting a non-technical overview of the Bayesian approach. References are provided for the technical details. A particular effort is made to illustrate the method with an example from psychology. References to software, more elaborate textbooks and tutorials enable researchers to apply this novel method to their own data.

[*] Comparisons among multiple groups are frequent in the context of behavioral research. For example, a researcher can be interested to know if a difference exists between three groups of students or patients that have received different treatments. Classical hypothesis testing is based on the evaluation of a null hypothesis, H_0, where all group means μ_j ($j = 1,\ldots, J$ groups) are stated to be equal:

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

This hypothesis can be conceived as a highly constrained hypothesis because it provides a lot of information about the means under investigation: the μ_j are all equal. On the other hand, the alternative hypothesis, H_a, can be written as

$$H_a : \text{not all } \mu_i \text{ equal}$$

The alternative H_a is an unconstrained hypothesis because we don't make any assumptions about the means μ_j. This hypothesis is not very informative because it states that something is happening but without any specification about what is going on.

In a situation where the researcher wants to evaluate H_0 against H_a it is common to apply a J group analysis of variance (ANOVA). The model can be written as

$$y_i = \mu_1 d_{1i} + \mu_2 d_{2i} + \cdots + \mu_J d_{Ji} + e_i \qquad (1)$$

where y_i is a dependent variable of interest, μ_j denotes the mean of group j, d_{ji} is one if person i is a member of group j, and zero otherwise, and $e_i \sim N(0, \sigma^2)$. The F-statistic, $F = \sigma^2_M / \sigma^2_e$, where σ_{M^2} is the between group means variance and σ^2_e is the within groups or residual variance, and associated F-distribution are used to test H_0 against H_a. The resulting p-value is sometimes called a measure of surprise, that is, it represents the incompatibility of data with the null hypothesis.

A significant ANOVA is usually followed by multiple pair wise comparisons. Using a traditional approach like, for instance, pair wise comparisons with a Bonferroni correction can lead to some methodological problems that will be discussed in the next section. The aim of this article is to present an alternative method based on a Bayesian approach that can evaluate specific expectations about multiple group

S. Béland and G. Raîche are at the Collectif pour le développement et les applications en mesure et évaluation, département d'éducation et de pédagogie (Cdame); I. Klugkist is at the Department of Methodology and Statistics; David Magis is at the département de mathématiques de l'Université de Liège.

means directly. Hypotheses expressing these expected patterns of means are called informative hypotheses and contain combinations of equality (=) and inequality (< or >) constraints; e.g., H_{inf1}: $\mu_1 > \mu_2 > \mu_3$ or H_{inf2}: $(\mu_1 = \mu_3) > \mu_2$.

The remainder of this paper is organized as follows. Section 1 discusses an exploratory approach for the evaluation of informative hypotheses and some of its disadvantages. Section 2 presents the Bayesian approach. Then, an example from psychology is provided in Section 3 and the paper is concluded with some reference material in Section 4.

Evaluating informative hypotheses through multiple comparisons

Many exploratory strategies are available to produce comparisons between groups to test a hypothesis. These approaches are based on pair wise comparisons between the groups under investigation. For example, a hypothesis concerning four means can be followed by a total of six pair wise tests. For each comparison, a t test can be used with a pre-specified α level; usually 0.05. This strategy does, however, not allow for family-wise error control. Here, the family-wise probability of having one or more false discoveries is $1-(1-0.05)^6 = 0.26$; that is, much larger than the pre-specified level of 0.05. A well-known strategy to control family-wise error is to use the Bonferroni correction where the α level is divided by the number of comparisons. For example, in the case of six tests, the α level per test becomes 0.05/6 = 0.0083. Unfortunately, the control of the family-wise error through a Bonferroni correction also causes a great loss of power. Although other strategies that are less conservative than the Bonferonni correction have been developed, finding a balance between proper control over the type I error and power remains problematic (Maxwell, 2004).

Another limitation of using multiple comparisons for the evaluation of informative hypotheses is that the test results can be inconclusive (i.e., the hypothesis is partially but not completely supported) or logically inconsistent (e.g., H_0: $\mu_1 = \mu_2$ and H_0: $\mu_1 = \mu_3$ are not rejected (i.e., not statistically significant), while H_0: $\mu_2 = \mu_3$ is rejected ($p<.05$)).

Finally, researchers may have multiple, competing informative hypotheses. While with significance (multiple) testing it is not possible to compare them against each other, with the Bayesian approach presented in this paper this is straightforward, irrespective of the hypotheses to be compared being nested or non-nested. A hypothetical example of two competing hypotheses, H_{inf1} and H_{inf2}, was provided in the introduction and in Section 3 we will present a real data example with three informative hypotheses. But first, in the next section, we will present the Bayesian method that was specifically designed for the evaluation of one or more informative hypotheses.

Bayesian evaluation of informative hypotheses

Hoijtink (2011) and Hoijtink, Klugkist & Boelen (2008) presented a method for the evaluation of informative hypotheses that (i) does not suffer from such multiple testing issues, (ii) has the capacity to test more informative hypotheses than the traditional H_0 and H_a, and (iii) can be used as a confirmatory method to mutually compare two or more hypotheses of interest. In the next subsection, the Bayes factor, a Bayesian model selection tool, is presented. This is followed by some remarks about the prior and posterior distributions in Section 2.2, and a presentation of the sampling based estimation of the Bayes factor in Section 2.3.

The Bayes factor

Bayesian model selection is a procedure that can be used to mutually evaluate two competing hypotheses. The Bayes factor comparing the support in the data for a hypothesis H_t relative to $H_{t'}$ is defined as:

$$B_{tt'} = \frac{m(y|H_t)}{m(y|H_{t'})}, \qquad (3)$$

that is, it is the ratio of two marginal likelihoods (Kass & Raftery, 1995). Loosely stated, a marginal likelihood is the likelihood of observed data given a specific model or hypothesis, taking both the fit and the complexity/size (also known as parsimoniousness) of the hypotheses into account. A resulting $B_{tt'} > 1$ implies more support for H_t than for $H_{t'}$ whereas a $B_{tt'} < 1$ implies the opposite. A resulting Bayes factor of, for instance, 4 can be interpreted as four times stronger support for H_t than for $H_{t'}$.

Klugkist, Laudy & Hoijtink (2005) derived a simplified formulation of the Bayes factor for the comparison of a constrained hypothesis against its unconstrained counterpart. This simplification can be made because an informative hypothesis as, for instance, H_{inf1}: $\mu_1 > \mu_2 > \mu_3$, is nested in the encompassing unconstrained hypothesis H_a: μ_1, μ_2, μ_3. The Bayes factor $B_{inf,a}$ is, in this context, defined as:

$$B_{\mathrm{inf},a} = \frac{f_{\mathrm{inf}}}{c_{\mathrm{inf}}}, \qquad (4)$$

where f_{inf} is the proportion of the posterior distribution of the unconstrained hypothesis that is in agreement with the constraints of H_{inf}, and c_{inf} is the proportion of the unconstrained prior in agreement with the constraints. The role of priors and posteriors in the Bayesian approach will be further elaborated in the following section. Note finally, that the constant c_{inf} can be interpreted as the complexity or size of the constrained hypothesis relative to the unconstrained hypothesis, and f_{inf} as the fit of H_{inf}.

Table 1: Prior estimation for means μ_1, μ_2 and μ_3 using the Gibbs sampler

Prior					
Iteration	**μ_1**	**μ_2**	**μ_3**	**H_i**	**H_a**
1	0.57	0.41	0.20	1	0
2	0.48	0.69	0.29	0	1
3	0.49	0.95	0.16	0	1
4	0.28	0.49	0.30	0	1
5	0.67	0.90	0.81	0	1
6	0.73	0.48	0.22	1	0
...	...	...	…	...	...
1000	0.51	0.34	0.41	0	1
Sum				163	837

Table 2: Posterior estimation for means μ_1, μ_2 and μ_3 using the Gibbs sampler

Posterior					
Iteration	**μ_1**	**μ_2**	**μ_3**	**H_i**	**H_a**
1	0.55	0.47	0.12	1	0
2	0.44	0.12	0.22	0	1
3	0.50	0.37	0.11	1	0
4	0.21	044	0.25	0	1
5	0.52	0.89	0.86	0	1
6	0.75	0.62	0.23	1	0
...	...	...	…	...	...
1000	0.47	0.35	0.25	1	0
Sum				659	341

Prior and posterior distributions

In Bayesian statistics, a prior distribution must be specified for the parameters of each model under investigation. The prior represents the knowledge about the model parameters *befor*e observing the data and can be *informative* (i.e., "knowledge" is based on experts opinions, historical data, or any acceptable assumption from theory) or *non-* or *low-informative* (specifying as little prior information about the parameters as possible). For a mean μ a low-informative prior is, for instance, a normal distribution with a very large variance, which is almost equal to stating that any value for μ is, *a priori*, equally likely.

Updating the prior distribution with observed data provides the posterior distribution. The posterior, therefore, represents the knowledge about the model parameters after seeing the data while taking the prior information into account. With low-informative priors, the posterior is dominated by the data and Bayesian parameter estimates will, in this case, provide results that are very similar to non-Bayesian estimators. Low-informative priors can, however, strongly affect Bayesian model selection results and should therefore be used with great care in those applications (e.g., Berger & Pericchi, 1996).

In the context of evaluating informative hypotheses, Klugkist, Laudy & Hoijtink (2005) formulated general criteria for the specification of the prior for the unconstrained hypothesis. Due to the nesting, priors for the constrained hypotheses follow automatically by truncation according to the constraints (see Klugkist & Hoijtink (2007) for an elaborate explanation). They showed that for hypotheses containing order constraints (i.e., < or >) it is possible to define low-informative priors that do not affect the resulting Bayes factor. However, as soon as equality constraints (i.e., =) are present this is not the case and the choice of priors becomes more complicated. How this is done in the proposed method and accompanying software is extensively described in, for instance, Hoijtink (2011) and Hoijtink, Klugkist & Boelen (2008).

Sampling based estimation of the Bayes factor

In the ANOVA context, a low-informative prior for J means μ_j and the residual variance σ_e^2 is specified. To obtain

Table 3: Bayes factor comparing the informative hypothesis against the unconstrained hypothesis for two outcome measures.

	Emotionality	Vividness
H_{inf1}: (EM = tones) > recall	0.24	0.16
H_{inf2}: EM > (tones = recall)	2.07	2.50
H_{inf3}: EM > tones > recall	1.12	1.24

an estimate for c_{inf}, that is, the proportion of the unconstrained prior in agreement with the constraints of any H_{inf} we use we Markov Chain Monte Carlo (MCMC) sampling to obtain a sample of parameter values from the unconstrained prior. In a similar vein, f_{inf} is estimated by taking an MCMC sample from the unconstrained posterior.

For both prior and posterior, a Gibbs sampler (see also, for instance, Casella & George, 1992; Jackman, 2009) consisting of the following steps is applied:

step 1: specification of initial values for the parameters μ_1, …, μ_J, σ_e^2.

step 2: sample a new value for each parameter μ_1, …, μ_J, σ_e^2 conditional on the current values of all other parameters (and in the case of the posterior also conditional on the data) in a fixed order.

step 3: repeat step 2 many times.

With the Gibbs sampler, many values for each of the parameters are obtained and together provide the marginal posterior distribution of each parameter and subsequently of all posterior estimates of interest. Note, however, that some important issues always need to be considered, that is, burn-in and convergence (initial values of the sampler need to be discarded because the chain is still 'starting up') and sufficient total number of iterations. Several diagnostics and plots are available to monitor this; see for a general overview, for instance, Cowles & Carlin (1996).

To illustrate how the sampled values provide an estimate for $B_{inf,a}$ we present, for a 3 means ANOVA, (part of) a sample of 1000 iterations from an unconstrained prior and posterior in Table 1 and 2, respectively. On each row in both tables, it is determined if the sampled values for μ_1, μ_2, μ_3 are in agreement with the hypothesis of interest: H_{inf1}: $\mu_1 > \mu_2 > \mu_3$. Consider, for instance the first row of Table 1 with means $\mu_1 = 0.57$, $\mu_2 = 0.41$, $\mu_3 = 0.20$. These means are indeed increasing and therefore in the last column "1 hit" is recorded. Likewise, on the next row the last column lists a "0" because the means violate the order stated in H_{inf1}. Overall, Table 1 shows that the estimate for the complexity c_{inf1} is 163 / 1000 = 0.163 and Table 2 shows an estimate of 659 / 1000 = 0.659 for the fit f_{inf}.

In this hypothetical illustration, the resulting Bayes factor is $B_{inf1,a}$ = 0.659 / 0.163 = 4.04, that is, the constrained hypothesis is supported by the data. In the next section an analysis based on a real data set from a psychological clinical trial is discussed.

Analysis of a real data set

A recent study by Van den Hout et al. (2012) aimed to investigate the efficacy of the use of tones as the source of bilateral stimulation in the treatment of posttraumatic stress disorder (PTSD). Interventions using Eye Movement and Desensitization and Reprocessing (EMDR), and not tones, have repeatedly proven to be effective. In the clinical practice, however, eye movements have been replaced by tones; apparently it is assumed that tones are equally or more effective.

In a clinical study, 12 patients were asked to recall their most upsetting memory while making eye movements (EM condition), or while hearing beep tones (tones condition), or without bilateral stimulation (recall only condition). The researchers formulated three competing hypotheses:

$$H_{inf1} : (EM = tones) > recall$$
$$H_{inf2} : EM > (tones = recall)$$
$$H_{inf3} : EM > tones > recall$$

For the theoretical motivation of the informative hypothesis and all details of the study we refer to the original publication and references therein. Here, we will summarize how the Bayesian approach provides direct answers to the question about the effectiveness of using tones: if tones are equally effective as the (evidence-based) EM intervention, H_{inf1} should receive the strongest support. H_{inf2} represents the possibility that tones are not effective at all, and support for the third hypothesis would suggest some effect of tones but inferiority to EM.

In Table 3, the results for two outcome measures are provided; reduction in emotionality and reduction in vividness of trauma memories. For both measures the findings clearly show that there is no support for H_{inf1} (Bs < 1), that is, there is reason to believe that tones are *not* as effective as EM. Furthermore, the Bayes factors for the second and third informative hypothesis indicate stronger evidence for no effect at all (2.07 and 2.50) than for a smaller effect of tones compared to EM (1.12 and 1.24). Note that,

the goal of this illustration is to demonstrate the usefulness of the Bayesian approach for researchers with explicit expectations about multiple group means. Readers that have an interest in the particular study are referred to Van den Hout *et al.* (2012) for an elaborate discussion of the interpretation (much more careful than presented here) of the results.

Conclusion

In this paper we shortly introduced the Bayesian model selection approach to the evaluation of informative hypotheses. With this approach a powerful tool is provided that can directly evaluate specific expectations about the outcomes of a study. As was shown in the psychological illustration, it is also possible to mutually compare different competing hypotheses on the same data set.

In Section 2, the methodology was introduced for a simple (between subjects) ANOVA. In this context, a FORTRAN program called *confirmatoryANOVA* (Kuiper, Klugkist and Hoijtink, 2010) can be used. The illustration in Section 3 concerned a within subjects design and therefore requires other software. Mulder *et al*. (2009, 2010) developed the FORTRAN program BIEMS that can evaluate informative hypotheses for (multivariate) normal models (AN(C)OVA, MAN(C)OVA, (multivariate) multiple regression, and repeated measures analysis). The interested reader can (freely) download both software packages from: http://vkc.library.uu.nl/vkc/ms/research/ProjectsWiki/Software.aspx.

References

Berger, J. & Pericchi, L. (1996). The intrinsic Bayes factor for model selection and prediction. *Journal of American Statistical Association, 91,* 109-122.

Casella, G. & George, E. (1992). Explaining the Gibbs Sampler. *American Statistician, 46*, 167-174.

Cowles, M. K. & Carlin, B. P. (1996). Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review, J*ournal of the American Statistical Association, 91*, 883-904.

Hoijtink, H. (2011). *Informative Hypotheses: Theory and Practice for Behavioral and Social Scientists*. London, UK: Chapman & Hall/CRC.

Hoijtink, H., Klugkist, I. & Boelen, P. A. (2008). *Bayesian evaluation of informative hypotheses.* New York, NJ: Springer.

Jackman, S. (2009). *Bayesian analysis for the social sciences.* Chichester, UK: Wiley.

Kass, R. E. & Raftery, A. E. (1995). Bayes Factor. Journal of the American Statistical Association, 90, 773-795.

Klugkist, I., Laudy, O. & Hoijtink, H. (2005). Inequality Constrained Analysis Of Variance: A Bayesian Approach. *Psychological Methods, 10*, 477-493.

Klugkist, I. & Hoijtink, H. (2007). The Bayes Factor for Inequality and About Equality Constrained Models. *Computational Statistics and Data Analysis,* 51, 6367-6379.

Kuiper, R. M., Klugkist, I. & Hoijtink, H. (2010). A Fortran 90 Program for Confirmatory Analysis of Variance. *Journal of Statistical Software, 34*, 1-31.

Kuiper, R. M. & Hoijtink, H. (2010). Comparisons of Means Using Exploratory and Confirmatory Approaches. *Psychological Methods, 15*, 69-86.

Lee, P. M. (2004). *Bayesian statistics: an introduction (*3td edition). London, UK: Hodder Arnold.

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods, 9,* 147-163.

Mulder, J., Klugkist, I., van de Schoot, R., Meeus, W., Selfhout, M. & Hoijtink, H. (2009). Bayesian model selection of informative hypotheses for repeated measurements. *Journal of Mathematical Psychology, 53,* 530-546.

Mulder, J., Hoijtink, H. & Klugkist, I. (2010). Equality and inequality constrained multivariate linear models: objective model selection using constrained posterior priors. *Journal of Statistical Planning and Inference, 140,* 887-906.

Van den Hout, M. A., Rijkeboer, M. M., Engelhard, I. M., Klugkist, I., Hornsveld, H., Toffolo, M. & Cath, D. C. (2012). Tones inferior to eye movements in the EMDR treatment of PTSD. *Behaviour Research and Therapy, 50*, 275-279.

Article received 28 May 2012
Article accepted 29 May 2012