Available online at www.sciencedirect.com

ScienceDirect

Behavior Therapy 46 (2015) 809–823

Behavior Therapy

www.elsevier.com/locate/bt

A Tutorial on Computing Bayes Factors for Single-Subject Designs

Rivka M. de Vries
University of Groningen

Bregje M.A. Hartogs
Center for Integrative Psychiatry, Lentis Mental Health Care

Richard D. Morey
University of Groningen

When researchers are interested in the effect of certain interventions on certain individuals, single-subject studies are often performed. In their most simple form, such single-subject studies require that a subject is measured on relevant criterion variables several times before an intervention and several times during or after the intervention. Scores from the two phases are then compared in order to investigate the intervention effect. Since observed scores typically consist of a mixture of true scores and random measurement error, simply looking at the difference in scores can be misleading. Hence, de Vries & Morey (2013) developed models and hypothesis tests for single-subject data, quantifying the evidence in data for the size and presence of an intervention effect. In this paper we give a non-technical overview of the models and hypothesis tests and show how they can be applied on real data using the `BayesSingleSub` **R package, with the aid of an empirical data set.**

Keywords: single-subject; `BayesSingleSub R package`; intervention; Bayes factor

In the past decades of psychological research, many studies have been performed with an interest in whether certain interventions are effective for specific individuals. Rather than focusing on large group effects, in these single-subject studies the interest is in whether each of the individuals improve on certain criteria after the intervention (Barlow et al., 2009). In its most simple form, a single-subject study requires that subjects are measured several times before an intervention (Phase or baseline phase) and several times during or after the intervention (Phase or intervention phase). Then for each individual the scores collected during the baseline phase are compared to the scores collected during the intervention phase. The larger the difference between the scores from each of the two phases, the more support the data provide for an intervention effect.

A complicating factor that typically arises in psychological research is sampling error. Since variables of interest like depression, anxiety, and other psychological constructs cannot be observed directly, indirect measures like ratings by therapists or questionnaires are often used to measure these constructs. These indirect measures will be affected not only by the constructs they are supposed to measure, but also by other influences that may vary over time within phases. The observed scores on these constructs are thus considered a mixture of true scores and error.

The contamination of observed scores with error makes the interpretation of the observed difference in scores before and after the intervention more difficult. After all, we do not know which part of the observed difference is due to a true difference and which part is due to sampling error, if there is a true difference at all. We could translate this into two different kinds of questions. The first question is

Corresponding author, contact information: Grote Kruisstraat 2/1, 9712 TS Groningen, The Netherlands; e-mail: r.m.de.vroes@rug.nl.

0005-7894/

whether the observed data suggest that there is a true difference in scores between the two phases or not. The second question is how large the true difference is likely to be.

De Vries & Morey (2013) developed models of single-subject designs, and proposed Bayesian hypothesis tests for single-subject data that quantify the evidence in an observed data set for or against a true difference in scores between two phases. Bayesian analysis of the models can also be used to estimate the size of the true difference. In this paper we illustrate the use of the `BayesSingleSub` R package (Morey & de Vries, 2014), which allows the computation of the hypothesis tests and parameter estimates from the model. Below we will first give a short overview of the models and describe the conditions under which the `BayesSingleSub` package can be used. We also introduce an empirical example data set that will be used for illustrations of the package. We then discuss each of the models more fully and use examples to illustrate how the hypothesis tests and estimates of intervention effects can be obtained from the package, and how the output should be interpreted.

Getting Started

De Vries & Morey (2013) developed two Bayesian models for single-subject data. Inference for these models differs from typical models for single subject data, in which analysis is typically performed within the frequentist framework. The Bayesian framework, however, has important differences from the frequentist framework: for instance, within the Bayesian framework, uncertainty is described using probability distributions see, e.g., Lee, 2004). One might be uncertain about the weather tomorrow (one might believe there is a .4 probability of rain) or whether an intervention is effective (we might believe that the intervention has a .4 probability of having been effective). In contrast, the frequentist approach defines probability as a long run frequency, like the number of times a coin lands up head out of an infinite number of coin tosses. These differences lead to important differences in the kinds of answers that can be obtained from a procedure. Whereas the frequentist approach provides procedures which control the proportion of wrong conclusions about hypotheses in the long run – assuming that the null hypothesis is true – the Bayesian approach allows the quantification of evidence (that is, the change in belief justified by the data) for specific hypotheses (Good, 1985). Depending on the goal of the user, one approach can be more useful than the other. Because in science the interest is in quantifying evidence yielded by observed data for different hypotheses, the Bayesian approach is often appropriate. In 2003 the Bayesian approach was already advocated by Jones (2003) for use in meta-analysis of single-subject studies. More recently, Zucker et al. (2010), Rindskopf (2014), and Swaminathan et al. (2014) advocated the use of Bayesian statistics in several applications.

De Vries & Morey (2013) described several Bayesian models for single-subject designs, the first of which expresses an intervention effect as a true difference in means before and during or after the intervention. It is a generalization of Rouder et al.'s (2009) JZS *t* test, accounting for the dependence between time points with an autoregressive (AR) model. For this reason, de Vries & Morey (2013) denote this model the JZS+AR model. The JZS+AR model is useful when the data within each phase of the design are expected to be stable and the intervention is assumed to have a constant effect on the variable of interest at every time point after the intervention. That is, the data at baseline are stable around a certain mean, and then there may be a shift in scores right after the implementation of the intervention after which the data are again stable around a certain mean. Such a data pattern could for instance be observed when the intervention consists of the administration of drugs, where the drugs have an immediate and stable effect on the variable of interest.

The second model in de Vries & Morey (2013) is similar to the JZS+AR model, except that it expresses an intervention effect as a true difference in trends before and during or after the intervention, rather than a simple difference in means. Hence, de Vries & Morey (2013) denote this model the TAR model, where the "T" refers to the trend difference. The TAR model can be useful when the assumption of stable data within each phase does not hold and systematic changes in scores are expected during the baseline or the intervention phase. This would be the case, for instance, if the intervention effect is gradual rather than immediate. In subsequent sections each of the models will be discussed in more detail; see also de Vries & Morey (2013).

Based on these models, de Vries & Morey (2013) proposed several hypothesis tests which quantify the evidence for true differences in means or trends between the baseline and the intervention phase. The models can also be used to estimate the size of these true differences. The hypothesis tests and effect size estimations can be performed using the BayesSingleSub package, which runs under the R statistical environment (R Development Core Team, 2009). The R statistical environment can be freely downloaded from http://cran.r-project.org/.