Chapter 1

Introduction

For some people, life is an exciting, interesting path from birth to death, full of challenges and learning opportunities. For other people, life can be tough from time to time, resulting in all kinds of mental issues like depression, anxiety, etcetera. Over the years, many different interventions have been developed which are supposed to cure people from their mental issues. But do these interventions really have a curing effect on people’s mental issues? In order to answer this question, clients undergoing an intervention are often measured on relevant constructs, like their level of depression, several times before, during, and/or after the intervention. The scores obtained with these measurements should give insight into how a client has changed due to the intervention.

A complicating factor in the evaluation of psychological interventions is that the observed scores, which are supposed to represent specific aspects of a client’s mental state, are usually not completely reliable. For example, the scores on a questionnaire that is supposed to measure depression, will be partly affected by the client’s level of depression, but also partly by random influences called measurement error (Buonaccorsi, 2010). Thus, scores on psychological measurement scales typically consist of a mixture of true scores and random measurement error. The more reliable the scale, the more the observed scores reflect true scores and the smaller the influence of random measurement error.

Because observed scores are partly affected by measurement error, simply looking at how the observed scores change over time can lead the researcher into seeing certain change patterns while in fact the observed change in scores is caused by measurement error. Hence, several hypothesis tests have been developed with the aim to test whether there is a true intervention effect or not. For example, the reliable change index (RCI; Jacobson and Truax, 1991) divides the difference between an observation before the intervention and an observation after the intervention by the standard error of the difference. If the RCI exceeds a certain threshold, the difference is considered reliable, that is, not due to random measurement error.

The tests that have been developed for the evaluation of interventions, like
the RCI, are all based on a similar principle: the probability of finding an intervention effect at least as extreme as the observed intervention effect, given that the true intervention effect is zero, is computed, and if this probability is small, it is concluded that there is a true intervention effect. This is the classical approach to hypothesis testing, and these tests are also known as null hypothesis significance tests. The probability of finding an intervention effect at least as extreme as the observed intervention effect, given that the true intervention effect is zero, is called the $p$ value. Typically, a $p$ value smaller than .05 is considered enough evidence against the hypothesis of a zero intervention effect in order to conclude that there probably is a true intervention effect.

Unfortunately, the $p$ values returned by null hypothesis significance tests cannot be interpreted as measures of evidence against the hypothesis of zero change, even though it is common practice to do so. Using $p$ values as such will provide the researcher with a measure of evidence with strange properties and lead the researcher into drawing ill-founded conclusions (e.g., Johnson, 2013; Rouder et al., 2009, tted; Wagenmakers, 2007). Although null hypothesis significance tests can be useful for other purposes, like when the interest is in decision making and the proportion of wrong decisions must be kept below a certain level in the long run, they are not useful for computing the evidence in data for specific hypotheses.

In this thesis we develop several hypothesis tests for intervention effects which are based on a different statistical approach and hence allow different kinds of conclusions than null hypothesis significance tests. The statistical approach within which the hypothesis tests are built is the so-called Bayesian approach. Within this approach, hypothesis are tested by computing the relative evidence in observed data for two competing hypotheses, with evidence expressed as the change in relative belief about the two hypotheses justified by the data (Good, 1985). Specifically, for each of the competing hypotheses it is computed how well the hypothesis can predict the observed data, that is, how likely the data are to have occurred under the hypotheses. The ratio of these two likelihoods is called the Bayes factor and indicates how much the data support one hypothesis as compared with the other hypothesis. When the data are equally likely to have occurred under each of the hypothesis, the Bayes factor is equal to 1. When the data are more likely to have occurred under one hypothesis than under the other hypotheses, the Bayes factor deviates from 1. The stronger the Bayes factor deviates from 1, the stronger the evidence in the data for one hypothesis over the other hypothesis.

In Chapter 2 of this thesis, we give a more extensive introduction into Bayes factors and develop new Bayes factor tests for single-subject designs, where the interest is in mean or trend differences before and during or after an intervention for a single-subject. We also provide an R (R Development Core Team, 2009) package, the BayesSingleSub package, for easy application of the tests. The Bayes factor for mean difference is a time series extension of Rouder et al.’s (2009) Jeffreys-Zellner-Siow (JZS) Bayes factor for mean difference. The Bayes factor for trend difference extends the time series Bayes factor for the mean difference with slope parameters.
CHAPTER 1. INTRODUCTION

In Chapter 3 of the thesis, we informally discuss the Bayes factors proposed in Chapter 2 and extensively discuss how the BayesSingleSub R package, which allows computation of the Bayes factors, can be applied in practice. With the aid of an empirical data set, we guide the reader step by step through all the steps necessary in order to obtain the Bayes factors and explain how the output should be interpreted.

In Chapter 4 we extend the Bayes factors for single-subject designs proposed in Chapter 2, to Bayes factors for group designs, where a group of subjects coming from the same population is measured several times before an intervention and several times during or after an intervention. Bayes factors are proposed for each subject as well as over subjects. R functions for computation of the Bayes factors are contained in the BayesSingleSub R package.

Finally, in Chapter 5 of the thesis, we apply part of the Bayes factors proposed in Chapter 4 on an empirical data set and explain how conclusions based on null hypothesis significance testing differ from conclusions based on Bayes factors. Inferential tests like the Bayes factors proposed in this thesis can be a useful aid in the process of data analysis.