When power analyses based on pilot data are biased: Inaccurate effect size estimators and follow-up bias

Casper Albersa,1,2, Daniël Lakensb,2,3

a University of Groningen, The Netherlands
b Eindhoven University, The Netherlands

ARTICLE INFO

Keywords:
Effect size
Power analysis
Follow-up bias
Eta-squared
Omega-squared
Epsilon-squared

ABSTRACT

When designing a study, the planned sample size is often based on power analyses. One way to choose an effect size for power analyses is by relying on pilot data. A-priori power analyses are only accurate when the effect size estimate is accurate. In this paper we highlight two sources of bias when performing a-priori power analyses for between-subject designs based on pilot data. First, we examine how the choice of the effect size index ($\eta^2$, $\omega^2$ and $\epsilon^2$) affects the sample size and power of the main study. Based on our observations, we recommend against the use of $\eta^2$ in a-priori power analyses. Second, we examine how the maximum sample size researchers are willing to collect in a main study (e.g. due to time or financial constraints) leads to overestimated effect size estimates in the studies that are performed. Determining the required sample size exclusively based on the effect size estimates from pilot data, and following up on pilot studies only when the sample size estimate for the main study is considered feasible, creates what we term follow-up bias. We explain how follow-up bias leads to underpowered main studies.

Our simulations show that designing main studies based on effect sizes estimated from small pilot studies does not yield desired levels of power due to accuracy bias and follow-up bias, even when publication bias is not an issue. We urge researchers to consider alternative approaches to determining the sample size of their studies, and discuss several options.

1. Introduction

It is common practice in psychological and behavioral research to express the results of a quantitative study in at least two numbers: One expressing the probability or likelihood of data under specified statistical models, usually through a $p$-value or Bayes factor, and one expressing the magnitude of the effect, often through a (standardized) effect size (ES). Reporting effect size estimates serves various purposes, one of which is facilitating cumulative science by allowing other researchers to use the effect size estimate in a-priori power analyses (Cohen, 1988). Power analyses can be used to design studies that have a desired probability of observing a statistically significant effect, assuming there is a true effect of a specified size. However, a-priori power analyses are only accurate when the effect size estimate is accurate. It has been pointed out that effect sizes reported in the literature are known to be inflated due to publication bias, and this widespread bias in reported effect sizes is a challenge when performing a-priori power analyses based on published research.

In this manuscript, we focus on two other sources of bias in power analyses that play an important role in power analysis even when publication bias and researchers' degrees of freedom do not influence effect size estimates (e.g., when researchers perform their own pilot study). These sources of bias point out clear limitations of the common practice to use the effect size from a pilot study to determine the sample size of a follow-up study through an a-priori power analysis. First, we will discuss the relatively straightforward matter of the impact of a biased effect size estimator ($\eta^2$), compared to less biased effect size estimators ($\epsilon^2$ and $\omega^2$) on the sample size estimate in power analyses. Second, we examine a source of bias which we refer to as follow-up bias. Effect size estimates vary around the true effect size. Even without publication bias, researchers are more likely to follow-up on initial studies that yielded higher effect size estimates than initial studies that...
yielded lower effect size estimates (cf. Greenwald, 1975), simply because these studies require less resources to observe a statistically significant result in the expected direction. We examine how this understandable behavior leads to an overestimation of the true effect size, on average, when performing a-priori power analyses, and thus leads to follow-up studies that are underpowered. Based on these observations, we argue against recent recommendations (Sakaluk, 2016) to use small pilot studies to explore effects. In the discussion, we offer some general recommendations to design well-powered studies.

2. Eta-squared, Epsilon-squared, and Omega-squared

In experimental psychology, it is extremely common to perform studies where participants are randomly assigned to different conditions, and analyze the results using analysis of variance (ANOVA) or (unpaired) t-tests (where a t-test is mathematically identical to a one-way ANOVA with two groups). We will illustrate our main points using ANOVA and the related effect sizes, but our conclusions generalize to other effect sizes and statistical tests. In one-way ANOVA, all analyses are based on the following decomposition of the variance. The total variance of all measurements together, $\sigma^2$, is split into a part that can be attributed to group-membership ($\sigma_g^2$) and a part that can not ($\sigma_w^2$):

$$\sigma^2 = \sigma_g^2 + \sigma_w^2.$$

The subscripts W, B, and T indicate 'within' samples, 'between' samples, and 'total'. Equivalently, one can decompose the so-called sums of squares:

$$SS_T = SS_B + SS_W.$$

One of the most common effect size indices in one-way ANOVA is eta-squared ($\eta^2$), which describes the proportion of variance that is explained by group membership. It dates back to Pearson (1911), who introduced it in a regression context, and to Fisher (1928), who used it in the ANOVA context. In statistical packages such as SPSS, eta-squared is the default effect size measure. Eta-squared is an upwardly biased estimate of the true population effect size, and two alternative effect size indices have been suggested that are less biased, namely epsilon-squared ($\epsilon^2$, Kelley, 1935) and omega-squared ($\omega^2$, Hays, 1963). For background reading on these (and other) indices, we refer to Levine and Hullett (2002), Okada (2013) and McGrath and Meyer (2006), and the references therein. Eta-squared, epsilon-squared, and omega-squared are defined as follows:

$$\eta^2 = \frac{SS_B}{SS_T},$$

$$\epsilon^2 = \frac{SS_B - df_g \times MSW}{SS_T},$$

$$\omega^2 = \frac{SS_B - df_g \times MSW}{SS_T + MSW},$$

where, using standard ANOVA-notation, $SS, MS$ and $df$ denote the sum-of-squares, mean sum-of-squares, and degrees of freedom. From these effect size estimates, the well-known Cohen’s $d$ and Cohen’s $f$ can be estimated (Cohen, 1988). For population effect sizes Cohen (1988, p. 276) states that $d = Z/\sqrt{N}$ with $Z = \eta^2/(1 - \eta^2)$. An unbiased estimate of Cohen’s $d$ is called Hedges’ $g$ (see Lakens, 2013), and recommendations in this article concerning the use of $\omega^2$ and $\epsilon^2$ instead of $\eta^2$ extend to the use of Hedges’ $g$ instead of Cohen’s $d$.

Alternative formulas for these effect sizes, where the computation is based only on the F-value and the degrees of freedom, are given in Appendix A.

These indices are estimators of the unknown true population effect size and, as such, contain possible bias and variability. It is well-known that $\eta^2$ has more bias than the other two indices, but the other two indices have more variability (cf. Albers, 2015; Lakens, 2015; Okada, 2013). The amount of bias and variability of these indices depends on the size of the sample and the true population effect size. When looking at performance measures that take both bias and variability into account, such as the (root) mean squared error, none of the three indices is uniformly optimal and very little is known on in which situations one method outperforms another. The first goal of the current manuscript is to provide practical guidelines on how to deal with these different effect size estimates when used in a-priori power analysis based on the effect size estimate in a previous study.

3. Bias in power analyses

The sampling distributions of $\eta^2$, $\omega^2$ and $\epsilon^2$ are considerably skewed (shown in Fig. 1 for $\eta^2$). Furthermore, the smaller the sample size, the more variable the effect size estimate is. Statisticians have warned against using effect size estimates from small samples in power analyses (Leon, Davis, & Kraemer, 2011). The two main reasons researchers should be careful when using effect sizes from the published literature in power analyses is that effect size estimates from small studies are inaccurate, and that publication bias inflates effect sizes. At the same time, many applied statistics texts recommend using effect sizes from related studies reported in the literature to perform a power analysis (e.g., Fritz, Morris, & Richler, 2012; Polit & Beck, 2004; Sawyer & Ball, 1981). In many cases, this is the only information researchers have about the possible size of the effect they are interested in. For example, the Reproducibility Project (Open Science Collaboration, 2015) relied on the effect sizes observed in the original studies to perform power analyses for replication studies.

The statistical power of a test depends on the true effect size, the sample size, and the alpha level that is used. The goal of a power analysis is to control Type II error rates, or to limit the probability of observing a non-significant result, assuming there is an effect of a specific size. In the presence of bias, researchers might unknowingly increase the Type II error rate of their studies. Alternatives to a-priori power analysis exists, such as deciding upon a smallest effect size of interest and using this to determine the required sample size in a power analysis (e.g. Lakens & Evers, 2014, Lang & Sesc, 2006, denoted the ‘minimal clinically important difference’ in medical research, Jäschke, Singer, & Guyatt, 1989). Other researchers have suggested to perform conservative power analyses (Perugini, Gallucci, & Costantini, 2014), or to model and correct for bias (Taylor & Muller, 1996).

Nevertheless, researchers might believe that building on effect size
دانلود مقاله

http://daneshyari.com/article/5045599

امکان دانلود نسخه تمام متن مقالات انگلیسی
امکان دانلود نسخه ترجمه شده مقالات
پذیرش سفارش ترجمه تخصصی
امکان جستجو در آرشیو جامعی از صدها موضوع و هزاران مقاله
امکان پرداخت اینترنتی با کلیه کارت های عضو شتاب
دانلود فوری مقاله پس از پرداخت آنلاین
پشتیبانی کامل خرید با بهره مندی از سیستم هوشمند رهگیری سفارشات