Journal of Neuroscience Methods 179 (2009) 173–178

Contents lists available at ScienceDirect

Journal of Neuroscience Methods

journal homepage: www.elsevier.com/locate/jneumeth

Comparing two small samples with an unstable, treatment-independent baseline

Skirmantas Janušonis*

Department of Psychology, University of California, Santa Barbara, CA 93106-9660, USA

ARTICLE INFO

Article history:
Received 10 November 2008
Received in revised form 22 January 2009
Accepted 22 January 2009

Keywords:
Small samples
Normalization
ANCOVA
Exact tests
Non-parametric tests
Wilcoxon rank-sum test
Mann–Whitney test
Developmental neurobiology

ABSTRACT

Due to time and resource constraints, small samples ($N = 3$–7 cases per group) are often used in neurobiological studies that employ multiple techniques. In a simulation study, five statistical tests were used to compare two small samples (treated and control) with an unstable, additive baseline. These five tests differed in the way that they used the baseline variable (B) to adjust or normalize the variable affected by the treatment (Y). We conclude that, if $N = 3$ or 4, the independent t-test on $Y–B$ tends to have the highest power; if $N \geq 7$, ANCOVA on Y with B as the covariate tends to have the highest power; and both tests have comparably high power if $N = 5$ or 6. The Wilcoxon rank-sum test (or, equivalently, the Mann–Whitney test) has precisely zero power if one group has 3 cases and the other has 3 or 4 cases. Some other problems of small-sample analysis are considered.

1. Introduction

High-profile neuroscience journals tend to favor manuscripts in which authors demonstrate the existence of a phenomenon by using a number of different (genetic, anatomical, pharmacological, etc.) experimental techniques. An unintended consequence of this approach is that the results of each of the experiments are often based on small samples. This approach is becoming standard in some subfields of neuroscience; for example, extremely small samples (3–5) are routinely used in developmental neurobiology papers published in prestigious journals (e.g., Gulacsi and Anderson, 2008; Naka et al., 2008; Pascual et al., 2008). The editorial preference for many interlocking pieces of evidence over the solidity of each of the individual pieces appears to rest on the assumption that the self-consistency itself provides good enough proof. At best, this type of reasoning makes research purely qualitative, mathematical modeling difficult, and puts neurobiology on a path long abandoned by exact sciences. At worst, it may lead to grossly incorrect conclusions, as noted even by scholars in humanities (Eco, 1990). The "soft" science of psychology began to seriously address these and other related questions (including the dubious value of null-hypothesis significance testing) several decades ago (Meehl, 1967; Cohen, 1994; Cohen et al., 2003). In this respect, some of the "harder" neurobiology continues to fall behind. Serious problems with statistical analysis in biology have been recently addressed by Nakagawa and Cuthill (2007).

The seriousness of these issues notwithstanding, small samples will continue to be used because a single measurement in neurobiology often costs hundreds of dollars. An important problem therefore is to know how to use such samples in the most optimal way. Specifically, in null-hypothesis significance testing, one should be likely to arrive at a non-significant result if two small samples are not different and a significant result if they are different. These probabilities are $1 - \alpha$ and $1 - \beta$, respectively, where α is the Type I error, β is the Type II error, and $1 - \beta$ is the power of the test. Unfortunately, the Type II error is rarely controlled for in neurobiological research. A typical inferential error is to assume that a P value greater than .05 indicates that the two samples are statistically equal. For the sake of argument, let us assume that the theoretical mean of a treated sample ($N = 3$) is 15, the theoretical mean of a control sample ($N = 3$) is 10, the theoretical standard deviations in both samples equal 3, both samples are drawn from normal distributions, and the two-tailed independent t-test is used to compare them. In this case, the $P \geq .05$ result should be expected in more than 65% of experiments (i.e., $\beta > .65$) despite the 50% greater theoretical mean in the treated sample. In other words, the $P \geq .05$ result *is* the expected result before the sampling even began and, as such, proves virtually nothing about the equality of the samples. Based on these considerations, it is obvious that the test with the highest power should always be preferred over other tests irrespective of the expected result ($P < .05$ or $P \geq .05$).

The main focus of this paper is to investigate the power of several statistical tests that can be used to compare two small samples

* Tel.: +1 805 893 6032; fax: +1 805 893 4303.
E-mail address: janusonis@psych.ucsb.edu.

0165-0270/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.jneumeth.2009.01.017

when the experimental baseline fluctuates (as it always does in practice). Since in neurobiology treated and control samples are often obtained from uncorrelated or weakly correlated sources (different animals, cell cultures, etc.), here they are assumed to be statistically independent ("unpaired").

An unstable experimental baseline is often dealt with by using the "normalization by division" procedure, in which the variable of interest is divided by a "baseline variable" that is immune to the experimental treatment but is sensitive to uncontrolled fluctuations of the experimental baseline. In immunohistochemistry, such a "normalized" variable may be the proportion of labeled cells with respect to another cell population that is not affected by the experimental treatment. In Western blotting, it may be the relative optic density of a protein band with respect to the band of a "housekeeping" (e.g., actin) protein in the same sample. Since such "normalization" and the normal distribution have nothing in common, to avoid confusion "normalized" variables can be referred to as "ratio variables".

Most statistical tests have not been designed for small samples of ratio variables. They typically use normal approximations that are valid only when samples are not small (e.g., the standard implementations of the Mann–Whitney and Wilcoxon rank-sum tests), or assume normality of the populations from which the samples are drawn (e.g., the *t*-test). However, the ratio of two normally distributed variables is not normally distributed. If two normally distributed variables are independent and have zero means, their ratio has the Cauchy distribution which is "unusual" in that it has no theoretical mean. A closed form of the distribution of the ratio of two normal variables with arbitrary means and standard deviations has been discovered only recently (Pham-Gia et al., 2006). Interestingly, this distribution can be asymmetric and/or bimodal (Pham-Gia et al., 2006). This finding has important consequences even for large samples; for example, a recent study has suggested that the distribution of serotonin levels in blood platelets (calculated as the amount of serotonin per platelet) may be bimodal in individuals diagnosed with pervasive developmental disorders (Mulder et al., 2004). Vickers (2001) has suggested that "normalization by division" should be avoided since it tends to reduce rather than increase the power of the *t*-test.

Several studies have shown that analysis of covariance (ANCOVA) has high power in randomized studies with an unstable baseline if several important assumptions are met (Vickers, 2001; Senn, 2006; Van Breukelen, 2006). Also, "normalization by subtraction" (when the baseline variable is subtracted from the variable of interest) has been shown to improve the power of the *t*-test when the correlation between the variable of interest and the baseline variable is large (Vickers, 2001). However, most published studies have focused on relatively large samples and small effect sizes (Cohen, 1992), which is a typical situation in psychology or epidemiology. In neurobiology, often small or extremely small samples are used to detect large effect sizes. Therefore, in the present study the power of five different statistical methods was assessed when the treated and control samples had as few as 3–7 cases.

2. Materials and methods

In a typical situation, one has to compare two samples, one of which represents the "treated" condition and the other one is a "control". In each individual case, we measure the variable of interest (Y) and a "baseline" variable (B) that is immune to the experimental treatment but sensitive to baseline fluctuations. Specifically, we consider the following model:

$$Y_{Gi} = \mu_0 + G\mu_1 + ey_{Gi} + ec_{Gi},$$

$$B_{Gi} = \mu_b + eb_{Gi} + ec_{Gi},$$

where Y_{Gi} and B_{Gi} are the values of Y and B in the ith case located in group G (where $G = 0$ if the group is the control group and 1 if it is the treated group); μ_0 and $\mu_0 + \mu_1$ are the theoretical means (expected values) of Y in the control and treated group, respectively; μ_b is the theoretical mean of B (equal in both groups); and ey_{Gi}, eb_{Gi}, and ec_{Gi} are statistically independent and normally distributed error terms with theoretical zero means and standard deviations σ_y, σ_b, and σ_c, respectively. It should be emphasized that, for given a pair of Y_{Gi} and B_{Gi}, ey_{Gi} and eb_{Gi} are generally different, whereas the same ec_{Gi} ("baseline fluctuation") is added to both Y_{Gi} and B_{Gi}. In other words, ey_{Gi} and eb_{Gi} vary within units of analysis (i's, or "cases"), whereas ec_{Gi} varies only between units, but not within.

We compare two very small samples ($N = 3$, 5, 7 per group) and numerically estimate the power of five statistical tests: (i) the independent, two-tailed *t*-test on Y (with B disregarded); (ii) the independent, two-tailed *t*-test on Y divided by B (i.e., Y/B); (iii) the two-tailed Wilcoxon rank-sum test on Y/B; (iv) the independent, two-tailed *t*-test on the difference between Y and B ($Y - B$); and (v) ANCOVA with Y as the dependent variable and B as the covariate. It should be noted that the Wilcoxon rank-sum test does not assume normality and is equivalent to the Mann–Whitney test. Next, we consider some advantages and drawbacks of each of the tests.

(i) The independent *t*-test on Y (with B disregarded) is appropriate considering the normality of the variables. However, the baseline-fluctuation term (ec_{Gi}) increases the variance of Y from σ_y^2 to $\sigma_y^2 + \sigma_c^2$, which reduces the apparent effect size of the treatment. Therefore, the treatment effect is less likely to be detected than if baseline fluctuations were taken into consideration.

(ii) The normalization Y/B takes baseline fluctuations into consideration but creates a variable that is not normally distributed (Pham-Gia et al., 2006). This violates the normality assumption of the *t*-test. The consequences of this violation are poorly understood when samples are very small.

(iii) In order to avoid the normality violation in (ii), the Y/B variables can be compared using the Wilcoxon rank-sum test or (equivalently) the Mann–Whitney test which do not assume normality. However, exact P values have to be calculated in these tests when samples are small (in standard software implementations, normal approximations are used for the statistics of these tests). Unless one is well familiar with the underlying mathematics, obtaining exact P values can be costly (currently, the SPSS *Exact tests* module is priced at $400). More importantly, the power of these tests (which are actually one test) is rarely considered when samples are very small.

(iv) The normalization $Y–B$ takes baseline fluctuations into consideration (mathematically, it eliminates the ec_{Gi} term) and creates a variable that is normally distributed. However, it also changes the variance of the tested variable from $\sigma_y^2 + \sigma_c^2$ to $\sigma_y^2 + \sigma_b^2$. Therefore, compared to the *t*-test on Y, the *t*-test on $Y - B$ will perform better if $\sigma_b < \sigma_c$ but worse if $\sigma_b > \sigma_c$.

(v) ANCOVA can naturally take into account baseline fluctuations if the baseline variable is considered to be a covariate. In our model, the relationship between Y_{Gi} and B_{Gi} can be written in the linear regression form:

$$Y_{Gi} = [\mu_0 + G\mu_1] + \left[\frac{\sigma_c^2}{\sigma_b^2 + \sigma_c^2}\right](B_{Gi} - \mu_b) + er_{Gi},$$

where the error term er_{Gi} is normally distributed with mean zero and variance $[\sigma_y^2 + \sigma_c^2] - [\sigma_c^4/(\sigma_b^2 + \sigma_c^2)]$ (Shiryaev, 1995). Since the regression weight at the centered baseline variable is the same in both groups (i.e., the regression slopes are independent of G), the model is equivalent to a standard ANCOVA model