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

Table 1
The numerical values of the parameters used in the simulations.

Variable	Values used
N (per group)	3; 5; 7
μ_0	100
μ_1	0–100 (in increments of 10)
μ_b	100; 500
$\sigma_y = \sigma_b$ and σ_c	30 and 10; 20 and 20; 10 and 30

with B as the covariate. It should be noted that in this model σ_c^2 is simply the covariance between Y and B and that their correlation coefficient is $\rho = \sigma_c^2 \left[(\sigma_y^2 + \sigma_c^2)(\sigma_b^2 + \sigma_c^2)\right]^{-1/2}$. Even though ANCOVA is an attractive alternative to the other tests, especially considering its power in larger samples (Vickers, 2001; Van Breukelen, 2006; Senn, 2006), it may not be the best test for small samples (Vickers, 2001). However, little is known about what sample sizes should be considered "small" in this regard.

Numerical simulation was used to investigate the power of the five statistical tests as the numerical values of the parameters were varied over an experimentally relevant range (Table 1). The sample sizes of the two groups (treated and control) were 3, 5 and 7 cases per group. For each combination of the parameters, experiments were repeated 5000 times and the proportion of the correct decisions made by each of the five statistical tests was plotted as a function of the theoretical mean difference between the treated and control groups. When $\mu_1 > 0$, a decision was considered correct if the test yielded $P < .05$; the proportion of correct decisions closely approximated the statistical power of the test ($1 - \beta$). When $\mu_1 = 0$, a decision was considered correct if the test yielded $P \geq .05$; the proportion of correct decisions closely approximated $1 - \alpha$, where α is the Type I error. Note that in this simulation the critical P value (.05) and the actual α may differ if the assumptions of a test are violated (as they are in the case of the t-test on Y/B), or if the discrete nature of the test-statistic distribution does not allow setting the actual α precisely at .05 (as is the case in the Wilcoxon rank-sum test). All calculations were carried out in a program written in Mathematica 6.0.3 (Wolfram Research, Inc.). Experiments with one or more negative values of Y or B were rerun until all values were non-negative. A published Mathematica algorithm (Weiss, 2005) was used to calculate the exact P values of the Wilcoxon rank-sum test. The source code of the program (Supplementary data) can be easily modified to calculate the power of the five tests with any other values of the parameters.

3. Results

When $\mu_1 = 0$, the Type I error of the t-test on Y/B did not exceed .05 (Figs. 1 and 2). This level of the Type I error was automatically guaranteed for the other tests. When $\mu_1 > 0$, the following results were obtained (Figs. 1 and 2):

1. The Wilcoxon rank-sum test never had the highest power compared to the other tests and had zero power when $N = 3$.
2. $\rho = .10$: the *t-test on Y* (with B ignored) had the highest power (Figs. 1 and 2A, D, G). Since in this case $\sigma_b > \sigma_c$, the t-test on $Y–B$ had less power, as theoretically expected. When the theoretical mean of the baseline variable was large ($\mu_b = 500$), the t-test on Y/B performed equally well but not better (Fig. 2A, D, G).
3. $N = 3$ and $\rho \geq .50$: the *t-test on Y–B* tended to have the highest power (Figs. 1B, C; 2C). Since in this case $\sigma_b \leq \sigma_c$, the t-test on Y had less (when $\sigma_b < \sigma_c$) or equal (when $\sigma_b = \sigma_c$, $\rho = .50$) power, as theoretically expected. More power was gained in the special case when all of the following conditions were satisfied: the theoretical mean of the baseline variable was large ($\mu_b = 500$), the correlation between Y and B was moderate ($\rho = .50$), and the t-test was performed on Y/B (Fig. 2B).
4. $N = 5$ and $\rho \geq .50$: the *t-test on Y–B* or *ANCOVA* tended to have the highest power with little difference between the two (Figs. 1 and 2E, F). When $\sigma_b = \sigma_c$ ($\rho = .50$), the t-tests on Y and $Y–B$ had equal power, as theoretically expected.
5. $N = 7$ and $\rho \geq .50$: the *ANCOVA* tended to have the highest power (Figs. 1 and 2H, I).

Some other findings that may be helpful in interpreting already published results will be mentioned:

1. $\rho = .50$: with the exception of the Wilcoxon rank-sum test at $N = 3$, the tests on Y/B tended to have statistical power comparable to that of the other tests when $\mu_b > \mu_0$ (Fig. 2B, E, H). However, the tests on Y/B had much less power than the other tests when $\mu_b = \mu_0$ (Fig. 1B, E, H).
2. $\rho = .90$: overall, the tests on Y/B tended to have relatively low statistical power. They had less power when $\mu_b > \mu_0$ (Fig. 2C, F, I) than when $\mu_b = \mu_0$ (Fig. 1C, F, I).
3. $N \geq 5$; $\rho = .50$; $\mu_b = 500$: all five tests performed almost equally well (Fig. 2E, H).

4. Discussion

Generally, the obtained results are consistent with findings obtained in larger samples (Vickers, 2001), but they provide more detailed information regarding small samples. Specifically, the following recommendations can be given if an experiment has a "good" baseline variable (i.e., when $\rho \geq .50$):

1. If $N = 3$ or $N = 4$, the independent t-test on $Y–B$ is recommended.
2. If $N = 5$ or $N = 6$, the independent t-test on $Y–B$ or, alternatively, ANCOVA on Y (with B as the covariate) are recommended.
3. If $N \geq 7$, ANCOVA (with B as the covariate) is recommended.
4. If $N = 3$, the Wilcoxon rank-sum test (or the Mann–Whitney test) should never be used.

It should be emphasized that these recommendations are based on the assumption that the model used in this analysis correctly reflects one's experimental situation (e.g., the variables are normally distributed; the baseline fluctuation is additive, etc.). The model assumes that the regression slopes in both groups are equal; violation of this assumption may lead to incorrect ANCOVA results (Engqvist, 2005). The assumption of homogeneity of the regression slopes can be tested by determining the significance of the interaction between G and B. If the interaction is not significant, one can proceed with ANCOVA after the interaction term has been removed (Engqvist, 2005). This analysis can be easily performed in most statistical programs; for SPSS users, Field (2005) is recommended. An analogous assumption is made by the t-test on $Y–B$, since the test assumes that the variances of the two populations are equal. If the variances of Y and B do not differ in the two groups, the variances of $Y–B$ in the two groups can be equal only if the covariances (and, consequently, the regression slopes) between Y and B in the two groups are equal. This similarity between ANCOVA and the t-test on $Y–B$ is not surprising, since the t-test on $Y–B$ can be considered to be a special case of ANCOVA when the ANCOVA slope is constrained to 1 (Van Breukelen, 2006).

The model also assumes that the theoretical means of the baseline variable are equal in both groups. The assumption is important for ANCOVA (Van Breukelen, 2006; Blance et al., 2007), although it can be violated in some experimental designs (Van Breukelen, 2006; Senn, 2006). It is also important for the t-test on $Y–B$ (Senn, 2006).

There may be cases when a "good" baseline variable is measured in units other than the units of *Y*, in which case *Y*–*B* cannot be calculated. In these cases, ANCOVA can be considered even if $N = 3$ or $N = 4$. It also should be noted that ANOVA and ANCOVA are special cases of linear regression (Cohen et al., 2003). If an experiment has more than two treatment conditions that are measurable on a continuous scale, linear regression (where the conditions can be considered to be a continuous variable) may have higher power than the corresponding ANOVA or ANCOVA (where the factors are categorical) (Owen and Froman, 2005; Lazic, 2008).

As demonstrated in Fig. 2B, there may be situations when the relatively high power of the *t*-test on *Y*–*B* can be surpassed by an even higher power of the *t*-test on *Y*/*B*. However, the *t*-test on *Y*/*B*

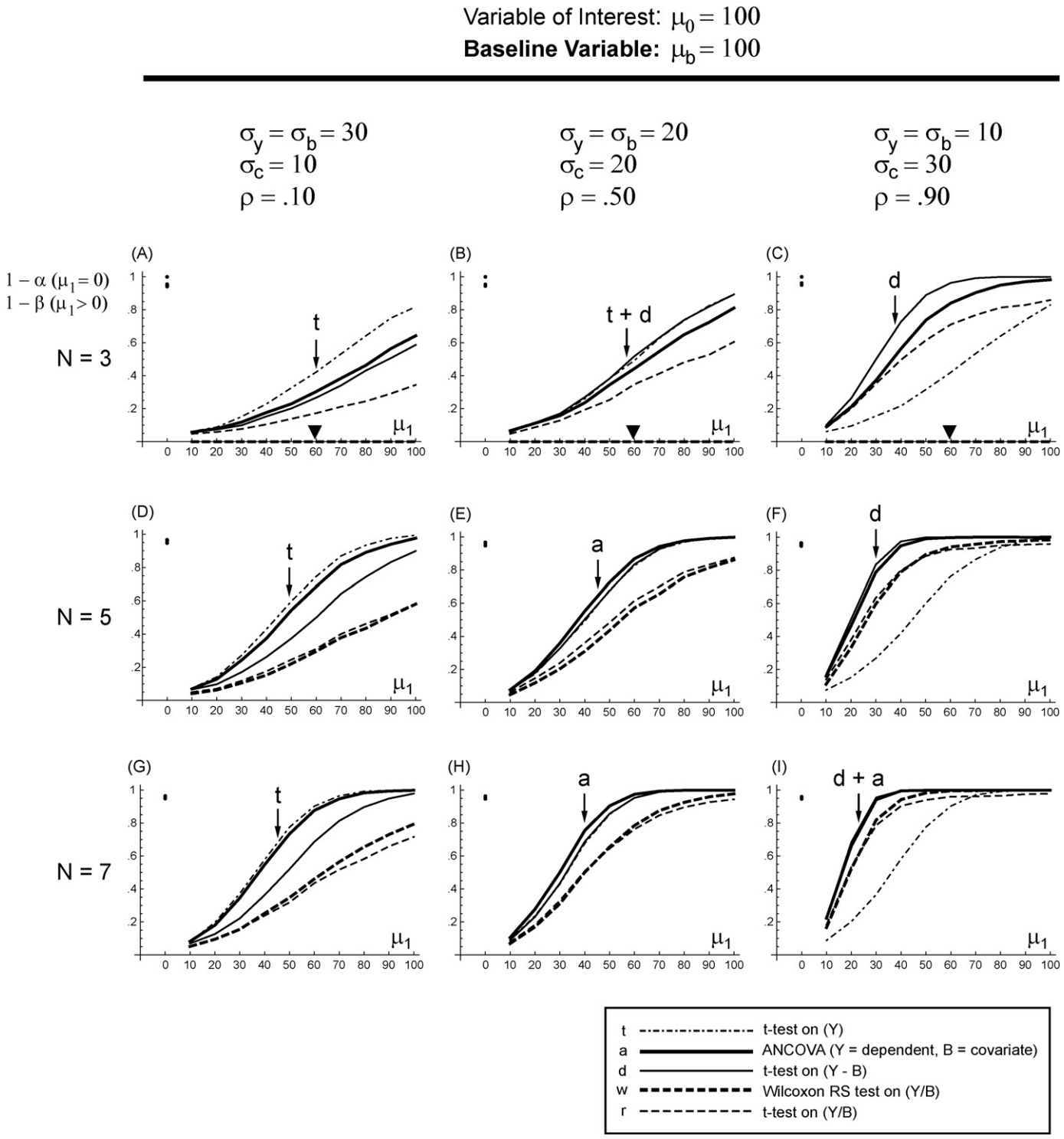

Fig. 1. The proportion of correct decisions made by each of the five tests when the treated and control samples had 3 (A–C), 5 (D–F), and 7 (G–I) cases and when the theoretical mean of the baseline variable (*B*) was equal to the theoretical mean of the variable of interest (*Y*) in the control group ($\mu_0 = \mu_b = 100$). The X-axis represents the difference between the theoretical means of *Y* in the treated and control groups (μ_1). The standard deviations were $\sigma_y = \sigma_b = 30$ and $\sigma_c = 10$ in (A), (D), (G); $\sigma_y = \sigma_b = 20$ and $\sigma_c = 20$ in (B), (E), (H); and $\sigma_y = \sigma_b = 10$ and $\sigma_c = 30$ in (C), (F), (I). The tests with the highest power are marked with arrows. Note that the Wilcoxon rank-sum test has zero power when $N = 3$ (arrowheads). The single letters in the legend are the abbreviations for the tests.

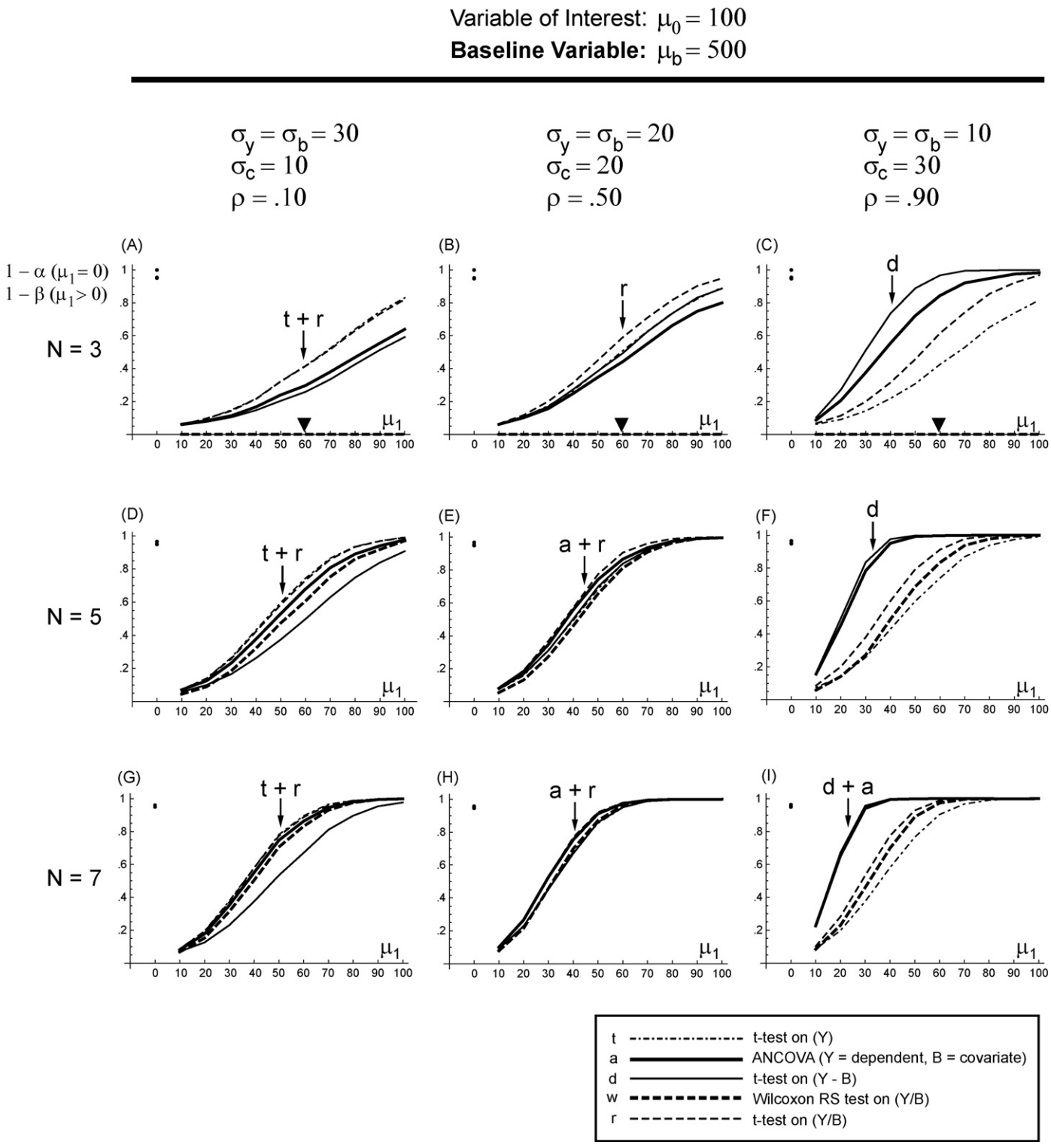

Fig. 2. The proportion of correct decisions made by each of the five tests when the treated and control samples had 3 (A–C), 5 (D–F), and 7 (G–I) cases and when the theoretical mean of the baseline variable (B) was much larger than the theoretical mean of the variable of interest (Y) in the control group ($\mu_0 < \mu_b = 500$). As expected, only the tests on $Y|B$ are affected by the μ_b value (see Fig. 1 for comparison). The X-axis represents the difference between the theoretical means of Y in the treated and control groups (μ_1). The standard deviations were $\sigma_y = \sigma_b = 30$ and $\sigma_c = 10$ in (A), (D), (G); $\sigma_y = \sigma_b = 20$ and $\sigma_c = 20$ in (B), (E), (H); and $\sigma_y = \sigma_b = 10$ and $\sigma_c = 30$ in (C), (F), (I). The tests with the highest power are marked with arrows. Note that the Wilcoxon rank-sum test has zero-power when $N = 3$ (arrowheads). The single letters in the legend are the abbreviations for the tests.

may have much less power than the t-test on $Y–B$ if the values of the parameters are not optimal (Fig. 1B). Since in a typical situation no accurate estimates of the parameters (e.g., μ_b, ρ) are available, this potential improvement is probably best avoided. If, however, one uses a well-established experimental set-up, plugging the known values into the program used in this study (Supplementary data) and obtaining power plots for the five tests may be a good strategy.

The Wilcoxon rank-sum test had zero power when $N = 3$. This is not surprising: there are only $6!/(3!3!) = 20$ ways to distribute 6 cases into 2 groups, which sets the lower theoretical limit for the exact two-tailed P at the non-significant $(1/20) \times 2 = .10$. It can be shown theoretically that the lower P limit is still non-significant (at the .05 significance level) if one group has 3 and the other 4 cases; the first time the two-tailed P can become less than .05 is when the total number of cases is 8 and neither of the two groups has fewer than 3 cases. Unwary authors and reviewers can be easily convinced that two small ($N = 3$) samples are statistically equal, since a non-parametric test was performed on a non-normally distributed variable (Y/B) and no significant difference was found between the groups ($P > .05$). In reality, any experiment with this many cases will unfailingly produce this result.

It is worth to briefly consider already published experimental studies that have performed t-tests on variables normalized by division (Y/B). Assuming that a "typical" correlation coefficient (ρ) between the variable of interest and a "good" baseline variable is closer to .50 than to .10 or .90, studies that have used a relatively large baseline variable ($\mu_b > \mu_0$) might be more reliable than those that have not ($\mu_b = \mu_0$) (compare Fig. 1B, E, H with Fig. 2B, E, H). However, the present study did not systematically investigate these suboptimal tests, the power of which can be very sensitive to the exact values of parameters (Vickers, 2001).

Acknowledgements

This study was supported, in part, by the Santa Barbara Cottage Hospital-UCSB Special Research Award and UCSB Academic Senate Faculty Research grants. I thank the anonymous reviewers for their insightful comments and advice.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jneumeth.2009.01.017.

References

Blance A, Tu Y-K, Baelum V, Gilthorpe MS. Statistical issues on the analysis of change in follow-up studies in dental research. Community Dent Oral Epidemiol 2007;35:412–20.
Cohen J. A power primer. Psychol Bull 1992;112:155–9.
Cohen J. The Earth is round ($p < .05$). Am Psychol 1994;49:997–1003.
Cohen J, Cohen P, West SG, Aiken LS. Applied multiple regression/correlation analysis for the behavioral sciences, third ed. Mahwah, New Jersey: Lawrence Erlbaum Associates; 2003. p. 5.
Eco U. Foucault's pendulum. New York: Ballantine Books; 1990. p. 3–533.
Engqvist L. The mistreatment of covariate interaction terms in linear model analyses of behavioural and evolutionary ecology studies. Anim Behav 2005;70:967–71.
Field A. Discovering statistics using SPSS. second ed. SAGE Publications Ltd.; 2005. p. 382–383.
Gulacsi AA, Anderson SA. β-Catenin-mediated Wnt signaling regulates neurogenesis in the ventral telencephalon. Nat Neurosci 2008;11:1383–91.
Lazic SE. Why we should use simpler models if the data allow this: relevance for ANOVA designs in experimental biology. BMC Physiol 2008;8:16.
Meehl PE. Theory-testing in psychology and physics: a methodological paradox. Philos Sci 1967;34:103–15.
Mulder EJ, Anderson GM, Kema IP, de Bildt A, van Lang ND, den Boer JA, Minderaa RB. Platelet serotonin levels in pervasive developmental disorders and mental retardation: diagnostic group differences, within-group distribution, and behavioural correlates. J Am Acad Child Adolesc Psychiatry 2004;43:491–9.
Naka H, Nakamura S, Shimazaki T, Okano H. Requirement for COUP-TFI and II in the temporal specification of neural stem cells in CNS development. Nat Neurosci 2008;11:1014–23.
Nakagawa S, Cuthill IC. Effect size, confidence interval and statistical significance: a practical guide for biologists. Biol Rev 2007;82:591–605.
Owen SV, Froman RD. Why carve up your continuous data? Res Nurs Health 2005;28:496–503.
Pascual A, Hidalgo-Figueroa M, Piruat JI, Pintado CO, Gomez-Diaz R, Lopez-Barneo J. Absolute requirement of GDNF for adult catecholaminergic neuron survival. Nat Neurosci 2008;11:755–61.
Pham-Gia T, Turkkan N, Marchand E. Density of the ratio of two normal random variables and applications. Commun Stat Theory Method 2006;35:1569–91.
Senn S. Change from baseline and analysis of covariance revisited. Stat Med 2006;25:4334–44.
Shiryaev AN. Probability. second ed. Springer; 1995. p. 43.
Van Breukelen GJP. ANCOVA versus change from baseline had more power in randomized studies and more bias in nonrandomized studies. J Clin Epidemiol 2006;59:920–5.
Vickers AJ. The use of percentage change from baseline as an outcome in a controlled trial is statistically inefficient: a simulation study. BMC Med Res Methodol 2001;1:6.
Weiss P. Applications of generating functions in nonparametric tests. Mathematica J 2005;9:803–23, http://www.mathematica-journal.com.