Jouni Kuha and Patrick Sturgis

Comment on 'What to do instead of significance testing? Calculating the "number of counterfactual cases needed to disturb a finding"' by Stephen Gorard and Jonathan Gorard

Article (Accepted version) (Refereed)

Original citation: Kuha, Jouni and Sturgis, Patrick (2016) *Comment on 'What to do instead of significance testing? Calculating the "number of counterfactual cases needed to disturb a finding"' by Stephen Gorard and Jonathan Gorard.* International Journal of Social Research Methodology. 19 (4). pp. 491-495. ISSN 1364-5579

DOI: 10.1080/13645579.2015.1126495

[Comment on "What to do instead of significance testing? Calculating the 'number of counterfactual cases needed to disturb a finding'" by Stephen Gorard and Jonathan Gorard]

Jouni Kuha
Department of Statistics, London School of Economics and Political Science

Patrick Sturgis
Department of Social Statistics, University of Southampton

Introduction

Stephen Gorard has been a vociferous and vocal critic of conventional frameworks for statistical inference. His criticisms have focused more on their common mis-application and misinterpretation than on inherent flaws in the methods themselves. However, it is clear that even if correctly applied and interpreted, Gorard regards the entirety of sampling theory and classical inferential statistics to be a misleading enterprise, which should be abandoned (Gorard 2015). What has been less evident in his published writings in this area is what he believes researchers who wish to use sample data to understand population characteristics should do instead. We therefore welcome this latest contribution, in which he (with J Gorard, henceforth G&G) sets out and demonstrates a new procedure – the number needed to disturb (NNTD) – which is described by G&G as offering "a superior alternative to the failed approach of significance testing and confidence intervals, and the limitations of 'effect' sizes" (2015, px).

In this note we comment on this proposed new method. We do not discuss the statements about significance tests, effect sizes, and missing data that appear in the first section of G&G. This is not because we agree with these points – in fact we disagree with most of them – but because a short note like is not the appropriate place for repeating standard explanations of the motivations, assumptions, and logic of conventional statistical inference.

In the remainder of the space available to us, we show that G&G's suggested methods for calculating their own statistic are cumbersome and inaccurate; we give a simple exact formula instead. Next, we demonstrate that, far from being a novel calculation, the NNTD is actually a simple function of the t-statistic and the standardized effect size and therefore behaves in ways that are already well understood within conventional frameworks. We go on to show how the NNTD does depart from conventional statistical inference in its reliance on the logic of sensitivity analysis. We argue that this renders it arbitrary as a basis for making decisions about the magnitude and importance of sample values. We finish by demonstrating that under plausible conditions, application of the NNTD will lead researchers to draw incorrect conclusions about the properties of samples and of the populations from which they were drawn. It goes without saying that this is a highly undesirable property of any statistic.

Defining the NNTD

An assessment of the utility of the NNTD in the terms set out in G&G is difficult because the way the method is described and evaluated in the paper is imprecise and unclear in important ways. Terms such as 'robustness', 'security', 'trustworthiness', and 'meaningfulness' of findings are used interchangeably but without any proper definition of what they mean. Therefore, to be clear about what we are commenting on, we begin by re-stating the NNTD procedure in our own conceptual and formal terms.

We take the spirit of the NNTD to be based on the following rationale. The researcher conjectures that the sample data she has observed are, hypothetically, supplemented by additional observations with particular characteristics. She calculates the smallest number of these additional observations that would be needed to render the point estimate of a particular parameter of interest, say the difference between two means, to be of a specified null value which might be, for

instance, zero. She then uses this number of additional hypothetical cases to assess the strength of the conclusions it is reasonable to draw about the parameter, based on the observed sample data. If the number is large then she can be confident that the point estimate is of substantive interest and has not arisen due to biases or chance variability across samples.

How this general procedure is operationalized in practice depends on the measurement level of the data, the parameter of interest, and the null value specified by the researcher. It seems unlikely that extending the NNTD beyond comparisons of group means would be as straightforward as is stated in G&G. For example, it is not obvious how the additional hypothetical observations would be constructed to reduce an estimated regression coefficient in a multiple linear regression model to zero. However, we do not seek to demonstrate this lack of generality here as the two-group mean comparison, which is the only case described in detail in G&G, is sufficient for illustrating the general limitations of the approach that we wish to note.

Both of the real-data examples in G&G are described as 'interventions', which we take to mean randomised experiments, although the procedure need not be limited to experimental designs. Suppose that each of N_t subjects is randomly assigned to a treatment or a control condition. The value of an outcome variable *Y* (e.g. a literacy score) is then recorded for a total of $N=N_1+N_2$ of the subjects, N_1 from the control group and N_2 from the treatment group. We denote the smaller of N_1 and N_2 (or both, if they are of equal size) by fN where $0 < f \leq 1/2$. There may also be $N_m = N_t - N$ missing cases, i.e. subjects who were assigned to a condition but for whom *Y* was subsequently not recorded due to attrition or other factors. Let $\bar{Y}_1$ and $\bar{Y}_2$ denote the observed sample means of *Y* among the control and treatment groups respectively, and *s* the sample standard deviation of *Y* in the two groups combined. In G&G these quantities are denoted μ_1, μ_2 and σ respectively but we use the more conventional notation for the sample quantities, reserving the use of Greek letters for the unknown values of the corresponding parameters.

Let μ_1 denote the average of the values of *Y* for all the N_t subjects who were initially included in the experiment if, hypothetically, they were all assigned to the control condition. In other words, μ_1 is the average of the potential outcomes under the control condition among the subjects in the experiment and μ_2 is the average of the potential outcomes under the treatment condition (Imbens and Rubin, 2015). A common parameter of interest in this situation is the difference $\mu_2 - \mu_1$, the so-called average treatment effect (ATE) among the experimental subjects, compared to the control condition. The sample mean difference $\bar{Y}_2 - \bar{Y}_1$ is typically used as the point estimate of $\mu_2 - \mu_1$. The null value of special interest is taken to be $\mu_2 - \mu_1 = 0$.

Following G&G, we consider the version of the NNTD where the hypothetical additional observations are all equal to $c = \bar{Y}_L + \text{sign}(\bar{Y}_L - \bar{Y}_S)s$, where $\bar{Y}_L$ denotes the value of the sample mean ($\bar{Y}_1$ or $\bar{Y}_2$) in the group where the observed sample size (N_1 or N_2) is larger and $\bar{Y}_S$ the mean in the group where the sample size is smaller. The additional hypothetical observations are all added to the smaller group.

Comparing the NNTD to the t-test

We observe at this point that, despite G&G's trenchant critique of classical inference, their NNTD procedure has a similar aim to statistical significance testing (though not to point or interval estimation of parameters, which it does not aim to do). This is to say, the NNTD is intended to aid the researcher in coming to a judgement about whether patterns and differences observed in sample data are also likely to be evident in the population from which the sample was drawn. It is therefore illustrative to compare the NNTD to the two-group t-test as a means of clarifying exactly how it differs from the sorts of conventional methods G&G propose that it should replace.

Both the NNTD and the two-group t-test can be described as comprising three steps. Step 1 is the definition of a statistic, that is, a number which can be calculated from the observed data. For the t-test, this is the two-sided p-value $p = 2 * [1 - F_{N-2}(|t|)]$, where F_{N-2} is the cumulative distribution function of the t-distribution with *N-2* degrees of freedom, and $|t|$ is the absolute value of the test statistic $t = (\bar{Y}_2 - \bar{Y}_1)/(s\sqrt{1/N_1 + 1/N_2})$. For the NNTD, the statistic is the NNTD itself, which we

denote by *N** and which can be calculated as $N^* = fN|\bar{Y}_2 - \bar{Y}_1|/s$. G&G propose both an approximate formula and an iterative algorithm to calculate *N**, although it is not obvious why, given this simple exact formula is available.

To be useful as a means of detecting differences between groups (whether or not one considers the magnitude of the difference to be substantively important) a statistic should have the property of being an increasing (or decreasing) function of all of $|\bar{Y}_2 - \bar{Y}_1|$, $1/s$ and *N*. Both *N** and *p* (as well as |*t*|) satisfy this condition. It is worth pointing out at this first step that $|t| = \sqrt{f(1-f)N}\,|d|$, $N^* = fN\,|d|$ and $N^* = \sqrt{fN/(1-f)}\,|t|$, where $|d| = |\bar{Y}_2 - \bar{Y}_1|/s$ is a commonly used sample measure of the absolute value of 'effect size'. Thus, rather than being a substantially new or different kind of statistic, the NNTD for the two-group mean comparison is in fact a simple function of the effect size and the t statistic.

Step 2 in both procedures is to make an interpretation of the value of the calculated statistic in terms of a hypothetical scenario. For NNTD, the 'what if' is that the sample mean difference would have been zero if the sample had, in addition to the cases actually observed, included *N** observations with the value *c* in the group with the smaller number of actually observed cases. For the t-test, the hypothetical is that the data were drawn from a population where $\mu_2 - \mu_1 = 0$, and *p* is the probability that a population with this characteristic would yield a value of |*t*| which is equal to or larger than the value of |*t*| for the observed sample data.

Steps 1 and 2 on their own are correct but essentially inconsequential mathematical statements about *p* and *N**. Step 3 is the crucial one because this is where a researcher makes a claim about how the hypothetical of step 2 relates to the real process that produced the observed data and uses this evidence to draw conclusions about the value of the population parameter. Step 3 may be done, broadly speaking, with a stronger or a weaker level of commitment. The stronger one is to assume that the hypothetical scenario corresponds in some sense to the real data-generating process. The weaker is the logic of *sensitivity analysis*, where the hypothetical is treated as a scenario that *could* have been real, followed by an assessment of how different choices for this scenario would affect conclusions about the parameters of interest (see for example Molenberghs and Kenward, 2007 and Rosenbaum, 2010).

The NNTD as a Sensitivity Analysis

Significance testing is normally done under the stronger level of commitment, that is supposing that the assumptions under which the *p*-value is calculated are an adequate representation of the process that actually generated the observed data. Conclusions about population parameters can then be drawn under the familiar logic of statistical inference – provided the assumptions of the test actually do represent the data-generating process sufficiently well.

For a two-group randomised experiment, the two key assumptions are that the subjects actually received the treatments assigned to them, and that any missing data are missing at random (MAR). The MAR assumption is that the probability that the outcome *Y* was not observed for a subject depends only on the group to which the subject was assigned (see Little and Rubin (2002) for more on the conventional frameworks and terminology for missing data). The adequacy of such assumptions cannot of course be guaranteed in general, neither here nor in any other types of experimental or observational research designs. An inevitable implication of the stronger level of commitment to inference, and a key responsibility for the empirical researcher, is thus concern about the validity of the assumptions of inference in each individual analysis. It is for good reason that a very large part of research methodology is devoted to this concern; to clarifying assumptions, devising diagnostic methods for assessing their adequacy, and developing research designs and methods of analysis which require minimal assumptions and/or are robust to their violation.

The NNTD is not meant to be employed with the strong logic of inference (this would require an assumption that if we had actually observed a further *N** observations, these *would* all have been

equal to *c*). Instead, it is a sensitivity analysis and thus shares the inherent strengths and weaknesses of this approach. Its strengths are that it is always feasible and typically straightforward to carry out. It also cannot be wrong, in the sense that it is not required to represent the true data-generating processes. For these same reasons, however, sensitivity analysis provides only an arbitrary and inconclusive basis for inference. Many different plausible scenarios can generally be considered, their degree of plausibility will always be at least partially subjective, and they will generally imply different conclusions. For the NNTD procedure, only one particular value for the hypothetical additional observations was proposed in G&G, but other plausible ones could be used instead and would yield different values of *N**.

In the best case, sensitivity analysis can suggest that qualitative conclusions about parameters of interest (e.g. the direction of an average treatment effect) remain unchanged under all plausible scenarios for the data-generating process, but this is rare. Perhaps more commonly, this type of analysis may reveal that *all* conclusions are possible under *some* plausible assumptions. Significance tests could also be interpreted as a special case of sensitivity analysis, that is, as an assessment of what we would conclude from the test *if* its assumptions were satisfied. This is not often done in practice, although it is quite standard for inference to be combined with elements of sensitivity analysis with regard to auxiliary assumptions. For example, it is recommended for examining the consequences of deviations from the MAR assumption for inference in the presence of missing data (i.e. `missingness not at random'; see e.g. Molenberghs and Kenward, 2007).

However, the NNTD as proposed in G&G is problematic even as a sensitivity analysis and, if implemented as they suggest, risks leading researchers to draw incorrect conclusions. This is because they consider only one value of *c* for the hypothetical additional observations. To illustrate the problem, consider a two-group randomised experiment where *N* is large and the true treatment effect is close to zero. Suppose further that there is missing data in the treatment group, so that $\bar{Y}_2$ is based on a subset of the cases that were originally randomised to the treatment condition, and that cases with lower values of *Y* in the treatment group are more likely to drop out of the study. The sample mean difference $\bar{Y}_2 - \bar{Y}_1$ is now expected to be positive, and so is *N**. If *N** is much larger than the number of missing cases, which it could be, applying the NNTD would lead the researcher to conclude that the observed difference between groups is 'robust' and should be 'taken seriously'. However, the difference between groups is in fact primarily due to the bias arising from the missing-data mechanism and, the bigger this bias, the larger the NNTD will be.

A final important limitation of the NNTD is the lack of any explicit guidance for how researchers should interpret different values of the statistic, a point acknowledged by G&G themselves in the concluding section of their article. For the two examples in G&G, the p-value for the t-test are *p=0.034* for Table 2 and *p=0.259* for Table 3. Of course, *p*-values closer to zero indicate stronger evidence against the null hypothesis of no average treatment effect and discrete decisions, should they be required, could be made based on the conventional levels of significance. The corresponding integer values of the NNTD statistic are *N*=38* for Table 2 (not 36, as stated in G&G) and *N*=21* for Table 3 (before subtracting the number of missing cases in the smaller group). Here, larger values of N* indicate stronger evidence in favour of a 'robust' or 'meaningful' treatment effect. However, G&G give no explicit suggestions about how different values of N* should be interpreted relative to one another. Does a value of 38 indicate an 'important' difference and a value of 21 an unimportant one? Can these numbers be interpreted in the same way irrespective of the underlying sample size? On these key matters, the reader is left completely in the dark.

Conclusion

G&G argue that conventional methods of statistical inference are widely mis-used and misinterpreted by researchers, a point which has also been made in a number of previous publications by the first author. Rather than advocating the correct use of conventional statistics, G&G propose what they claim is a new and superior approach to assessing the 'importance' of

research findings, the NNTD. However, we have shown that the NNTD procedure is actually a form of sensitivity analysis, based on a simple function of the effect size and the t statistic. It is our view that the logic of sensitivity analysis does not provide a general replacement for statistical inference; it is simply too weak and arbitrary for that purpose. Sensitivity analysis does have an important supporting role to play in data analysis, particularly in assessing how substantive conclusions might be affected by empirically un-verifiable assumptions, but this has long been well understood. The NNTD procedure is one means of implementing such an analysis. However, we are not tempted to recommend its use it in this context either, not least because there is currently no explicit guidance on how different values of the statistic should be interpreted. In summary, our assessment is that the NNTD as proposed by G&G is fundamentally flawed as a general replacement for existing methods of statistical inference and is problematic for more limited application as a variant of sensitivity analysis.

Bibliography

Gorard, S. (2015) Rethinking 'quantitative' methods and the development of new researchers, *Review of Education*, 3(1), 72-96.

Imbens, G. W. and Rubin, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. New York: Cambridge University Press.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*, 2nd ed. Hoboken, NJ: Wiley.

Molenberghs, G. and Kenward, M. G. (2007). *Missing Data in Clinical Studies*. Chichester: Wiley.

Rosenbaum, P. R. (2010). *Design of Observational Studies*. New York: Springer.