On the failures of the null-hypothesis test

Nicholas T. Longford*
SNTL and Universitat Pompeu Fabra, Barcelona, Spain

Abstract

This report comprises four reactions to the recent policy statement in *Basic and Applied Social Psychology* that announced a ban on null-hypothesis testing in that journal. A personal perspective is presented which agrees with the editors that null-hypothesis testing has become dysfunctional, but proposes a solution different from the editors'. In particular, application of formal statistical methods is defended, but they have to be tailored much more closely to the specifics of the problem that is addressed.

Keywords: decision theory; expected loss; hypothesis testing; plausible value; research agenda.

MSC classification: 62F-03, 62G-10, 62H-15.

*N.T.Longford, SNTL and Universitat Pompeu Fabra, c/ R. Trias Fargas 25–27, 08005 Barcelona, Spain. Email: sntlnick@sntl.co.uk

To the quantitative psychologist

In a recent policy statement, the editors of a journal in psychology announced that they would no longer permit the results of any null-hypothesis tests to appear on its pages (Trafimow and Marks, 2015). This is not an abrupt decision. It follows a one-year stay of execution which allowed for ample consultation and reflection by the editors and time for prospective autors to change how they conduct and report their research. The editorial uses some forceful language to condemn the null-hypothesis as 'invalid' and 'an obstacle to creative thinking'. The editors emphasise descriptive statistics as an alternative to formal statistical inference.

This note outlines a perspective which wholeheartedly supports this 'ban' but defends the role of simple but formal statistical methods in psychology and all related sciences. The catch is that these methods are not in the mainstream of statistical practice and require some work and hard thinking by the psychologist (the expert).

First, we offer a simple statement that dismisses any application of a hypothesis test. In this statement, we assume that the purpose of a hypothesis test is to decide how to proceed in a research, business or some other agenda. One course of action, denoted by A, would be appropriate if the hypothesis were valid, and the other, B, if the alternative were valid. Being aware of the imperfection of the outcome of a hypothesis test, we should carefully assess the consequences (ramifications) of the two kinds of error. All procedures that disregard these consequences, and they include the hypothesis test, are deeply flawed.

Our proposal repairs this flaw, but requires the expert to quantify the consequences. This is an activity that, unfortunately, has a very weak tradition in psychology and other social sciences. Its straightforward application would yield a trivial result that might at first appear as totally illogical. If a rational person had to choose between zero and non-zero for an unknown quantity of scientific interest, such as the difference of the means of a variable between men and women in a particular population, which could attain any one of a continuum of values, the obvious choice would be 'non-zero', because it offers infinitely more possibilities than the solitary zero. Just a reminder of the context: 10^{-1000} is not zero.

This suggests that the null-hypothesis misrepresents the expert's position. The hypothesis and alternative, or the two contemplated courses of action (the available options), correspond not to zero and non-zero, but to small and large. This is corroborated by every instance of nonsignificance that is interpreted as evidence that 'the effect is small'. This interpretation is incorrect because failure to reject the null-hypothesis offers no support for the null, nor for any hypothesis formulated ad hoc. Hypothesis

test is akin to a bet, such as on the winner of a horse race. The bet has to be placed prior to the race (data collection), and the terms of the bet cannot be altered at any time after its placement.

For choosing between small and large, we should first specify what these two terms mean, or where the borderline between them lies. For example, small may correspond to the interval $(-\Delta, \Delta)$ and large to its complement. The value of Δ should be set by the expert (or in consultation with him or her, referred to as *elicitation*), as it crucially depends on the two contemplated courses of action. This is not a trivial exercise. It entails a deliberation of which action, A or B, would be taken if the value of the studied effect were known. Some leeway in setting Δ is permitted, but such laxity does not come free because it opens up the possibility of an impasse, as we discuss below.

Suppose for the moment that the value of Δ has been set. The test of the hypothesis that the effect is in the range $(-\Delta, \Delta)$, against the alternative of its complement, might now be appropriate, except for the earlier dismissal owing to the disregard for the consequences of the two types of error. Quantifying these consequences is another task for the expert in consultation with the analyst. Suppose they conclude that choosing action A inappropriately, when the effect is large, is R times more damaging than choosing action B inappropriately, when the effect is small.

This constant R, called the penalty ratio, is an essential element of the analysis. Suppose R is extremely large. Then the study should not be conducted and action B should be taken unconditionally. In this case, taking action A is too risky in comparison with action B. If R is extremely small the study is a waste of resources because action A should be taken irrespective of its outcome. In a typical setting, the expert's position is somewhere between these two extremes, and it should be declared by setting the penalty ratio R. There is no need to pinpoint its value; it is more valuable to maintain integrity by admitting being uncertain and specifying a plausible range for R, such as $(R_{\mathrm{L}}, R_{\mathrm{U}})$. On the one hand, a narrower range is preferred because it reduces the chances of an impasse. On the other hand, all the parties involved have to agree that any value of R outside the range can be ruled out.

We outline the solution for a special case. For details, background, extensions and some applications, see Longford (2013), and for more motivation Lindley (1985). For an effect θ that is of interest, and its estimator $\hat{\theta}$, consider the trivial model

$$\hat{\theta} \,=\, \theta + \varepsilon\,,$$

where ε is the estimation error, assumed to be normally distributed with zero mean and known variance τ^2, which usually depends on the sample size or, more generally,

on the design of the study. Here $\hat{\theta}$ and ε are random and θ is fixed. After concluding the study, the value of $\hat{\theta}$ is established, so its status is changed from random to fixed. We also change the status of θ, from fixed to random. Such operations are more natural in a Bayesian framework, but having to formulate prior distribution(s) is a distraction in the present context.

Owing to the symmetry of the normal distribution, θ is now normally distributed, with mean $\hat{\theta}$ and variance τ^2. Denote by $\phi(x;\mu,\sigma^2)$ the density of the normal distribution with mean μ and variance σ^2, and by $\Phi(x;\mu,\sigma^2)$ its distribution function. When $\mu = 0$ and $\sigma^2 = 1$, these arguments are dropped. Define the unit of loss as the damage (cost, harm, or similar) caused by one instance of inappropriate choice of action B. Such a unit could be called a *lossile*. The loss caused by inappropriate choice of A is R lossiles. An appropriate choice results in no loss.

The expectation of the loss if we choose action A is

$$\begin{aligned} Q_{\mathrm{A}} &= R\left\{\int_{-\infty}^{-\Delta}\phi\big(x;\hat{\theta},\tau^2\big)\,\mathrm{d}x + \int_{\Delta}^{+\infty}\phi\big(x;\hat{\theta},\tau^2\big)\,\mathrm{d}x\right\} \\ &= R\left\{2 - \Phi\big(\Delta;-\hat{\theta},\tau^2\big) - \Phi\big(\Delta;\hat{\theta},\tau^2\big)\right\}. \end{aligned}$$

The expected loss if action B is chosen is

$$Q_{\mathrm{B}} = \int_{-\Delta}^{\Delta}\phi\big(x;\hat{\theta},\tau^2\big)\,\mathrm{d}x = \Phi\big(\Delta;\hat{\theta},\tau^2\big) + \Phi\big(\Delta;-\hat{\theta},\tau^2\big) - 1\,.$$

We want to lose least, so, in the presence of uncertainty, we choose the option with the smaller expected loss. Thus, action A is chosen when $Q_{\mathrm{A}} < Q_{\mathrm{B}}$, that is, when

$$(R+1)\left\{\Phi\big(\Delta;\hat{\theta},\tau^2\big) + \Phi\big(\Delta;-\hat{\theta},\tau^2\big)\right\} > 2R+1\,,$$

or equivalently, when

$$\Phi\left(\frac{\Delta-\hat{\theta}}{\tau}\right) + \Phi\left(\frac{\Delta+\hat{\theta}}{\tau}\right) > \frac{2R+1}{R+1}\,.$$

As a function of $\hat{\theta}$, the left-hand side increases for negative $\hat{\theta}$, attains its maximum for $\hat{\theta} = 0$, when it is equal to $2\Phi(\Delta/\tau) > 1$, and then decreases. Its limits as $\hat{\theta} \to \pm\infty$ are zero. When even $2\Phi(\Delta/\tau)$ is smaller than $(2R+1)/(R+1)$, action B is associated with smaller expected loss for all possible outcomes, so the study is not worthwhile; the value of R is very high. (This is not to suggest that the value should be set lower.) As Δ approaches zero, the left-hand side converges to unity for all values of $\hat{\theta}$. Since the (constant) right-hand side is greater than 1.0, for sufficiently small Δ, action B is preferred for all values of $\hat{\theta}$. For $\hat{\theta}$ fixed, $Q_{\mathrm{A}} < Q_{\mathrm{B}}$ for sufficiently large Δ (a liberal definition of 'small'), and then action A is appropriate.

Uncertainty about Δ and R is addressed by defining plausible ranges for them. Relying on certain monotonicity properties, the problem is then solved for the four combinations of borderline plausible values of Δ and R. If the same action is found appropriate in all four cases, then we have an unequivocal preference for action A or B. Otherwise an impasse arises, which can be resolved only by resuming the elicitation process and reducing the plausible range of Δ or R, or both. However, elicitation is always conducted preferably at the design stage because insights gained by inspecting and analysing the data may subvert it.

Conclusion

The approach outlined in this section has three important features that the established methods based on hypothesis testing lack:

- symmetry — the two contemplated options are treated symmetrically, unlike a hypothesis (the default) and its alternative;
- continuity — the analysis is informed by the context, the expert's perspective and agenda, which are ignored in a hypothesis test;
- economy — the aim is to minimize the (quantified) consequences of the error, not its hypothetical probabilities.

The transparent subjectivity of this approach might be regarded as a drawback. If the expert's perspective, priorities and value judgements, as quantified by the values of Δ and R, change, then so may the outcome of the analysis. But this is due to the features of the method that enable us to tailor the analysis closely to a particular expert's perspective. The established process of hypothesis testing also caters for a perspective, but we have never cared to formulate it explicitly, nor to inquire whether it is suitable. As a result, it caters for everybody and nobody at the same time. That is how, in the opinion of Trafimow and Marks (2015), it has been turned into a ritual that has derailed the research psychologists' scientific priorities.

For the quantitative educational researcher

Hypothesis testing is a ubiquitous activity in statistical applications in educational research. Among the multitude of its applications, probably the simplest is to compare two groups defined by a treatment, intervention, a dichotomous background variable or another factor. Model selection can be regarded as a sequence of hypothesis tests and the nature of many model diagnostic procedures, e.g., for outlier detection, is not much different. As a statistical workhorse, hypothesis testing has had no equal for the last few decades. There is some dissent from this trend, but the objections are mainly to some of its unprincipled applications and abuses.

It comes therefore as a surprise that a journal in psychology has recently banned null-hypothesis from its pages (Trafimow and Marks, 2015). The decision is not abrupt. It was announced in early 2014 (Trafimow, 2014), with a year's stay of execution to allow for an adjustment by prospective authors. The recent editorial (Trafimow and Marks, 2015) is quite scathing about the null-hypothesis, regarding it as 'an obstacle to creative thinking' from which research in psychology should be liberated. As social sciences, psychology and education have a lot in common. Both study mental processes, and grapple with measurement of human characteristics by imperfect instruments (questionnaires or tests) and have to deal with various human inconsistencies and idiosyncracies. Many statistical methods, such as multilevel analysis, factor analysis and structural equations, as well as their adaptations for categorical outcomes, are about equally prominent in applications in these two sciences.

We should therefore closely examine why the editors of the *Basic and Applied Social Psychology* took such an drastic step and reflect whether some of the reasons for it might be applicable also to other journals and to quantitatively oriented educational research in general. This section presents a personal examination of this matter.

Taken in isolation, a single hypothesis test is a form of a statistical bet on the outcome of a study. Like in betting on a horse in a race, it is a contract drawn up prior to the study, and its terms cannot be altered at any time thereafter. There are only two possible outcomes. Whether 'my' horse comes close second or dead last, the bet is lost. Whether the horse wins by a whisker or by a furlong, the bet is won. The bookmaker's principal asset is integrity — unconditional adherence to the contract. It leaves very little scope for interpretation. The statistical version of the bet is contracted prior to inspecting the data, or when planning a study that will generate the data, and every possible outcome of the study (every possible value of a specified test statistic) is classified as either 'significance' or 'no significance'. Significance means that we have evidence of an effect; for instance, that a particular parameter is not equal to zero.

That we routinely drop the 'evidence of', and continue as if the effect was present, is a minor misdemeanor; 5%, the conventional level of significance, is thought to be next to nothing. These minor offences accumulate when we conduct many hypothesis tests, when we report them selectively, and when they are formulated ad hoc; when we place some bets after the race is over, and pretend that we have collected the winnings when none were due, or take on the roles of the bettor and bookmaker simultaneously.

No significance means that we have no evidence of an effect. It is a bet lost. Without some verbal contortions, the language we use is not rich enough to express clinically what this means. It offers us nothing to say about the null-hypothesis. If we conclude with no significance, it is not even a finding, and claims of evidence that the null applies or that the effect is small are gross abuses of both logic and statistics. Equally so is a denial that we have placed a bet, or a decision not to place a bet since it is obvious that we would lose, having seen the race, or some of it.

In brief, no significance has only one interpretation: a state of ignorance. This is an unsustainable state of affairs. The considerable investment to plan and conduct a study is rewarded by the risk of learning absolutely nothing. The commonly adopted solution is to continue with research, business or other affairs as if the studied effect were absent — as if the null-hypothesis were valid. That is an abuse that is unfortunately so widespread as to be protected institutionally. For example, 'no significance' in model selection is a right to delete a covariate from the model (setting its beta coefficient to zero).

These flaws are not in the theory of hypothesis testing, but in its poor match to the problems we face. In a typical setting, we would like to find out whether an effect is zero or not. This formulation is inaccurate, because if we knew that the effect is extremely small we would regard it as if it were zero. So, a more accurate description of our intent is to decide whether a particular effect is small or large. The first step in this is to declare, or define, what 'small' means. Suppose that in a particular context it means that the absolute value of the effect is smaller than 0.4. Identifying this value is not a matter for data analysis but for an assessment of what the researcher intended to do following the test of the null-hypothesis. How big an effect is still to be regarded as 'zero'?

This may suggest testing the hypothesis that the effect is in the range $(-0.4, 0.4)$. But the power of such a test would be very low for the effect of 0.40001, so we would end up widening the interval we set earlier. The culprit in this is the *asymmetry* of the hypothesis and the alternative. Choosing between two courses of action, one corresponding to the effect being small and the other to being large, is a symmetric

problem.

Hypothesis testing is poorly suited for this task (Lindley, 1998), because it is oblivious to the consequences of the two kinds of error that can be committed. These consequences are specific to each study. If they are not important, then the study is inconsequential, and both its conduct and analysis are a waste of resources. This is not an argument for banning hypothesis testing from any forum, but an indication of its weaknesses that have to be attended to with some earnestness.

A reference to weaknesses does not imply any disrespect for the hypothesis test as an invention of key importance that has done a great service to statistics in the past. But times have moved on, and it has for some time not been fit for the demands we've been placing on it. Hypothesis tests are applied in the same form and with the same conventions as were proposed, not always with a firm resolve, several decades ago, in a different social, economic and scientific context.

A framework alternative to hypothesis testing is outlined in a very readable form by Lindley (1985). It is focused on decisions and is concerned with the loss (harm, damage, setback, or the like) arising from making erroneous choices. It is implemented for the most common problems by Longford (2013). It requires a definition of the categories 'small' and 'large' and a quantification of the consequences (losses) of the two kinds of error. They bring more of the context of the study into the analysis. A hypothesis test is not informed by such context. Hypothesis testing operates with *hypothetical* probabilities of the errors. The framework we propose, centered on decision-making, is concerned with minimising the expected loss associated with the choices between the available options. Expected loss is much easier to handle because expectation is additive; losses, and their expectations are simply added up. The calculus of probabilities, especially when they are for dependent events, is much more complex.

To the statistics lecturer

Most basic statistics textbooks on hypothesis testing spell out clearly the relevant assumptions and warn against misinterpretation of its results and other misuse. But the focus of the instruction in undergraduate courses is invariably on the mechanics of applying a test, starting with the one-way analysis of variance. As aspiring authors of research reports and journal articles, the graduates then focus on the wording of the conclusions based on such tests, and studying the policies of journal editors becomes more relevant, urgent, and fruitful, than adhering to the basics and logic of the general method. In practice, hypothesis tests are used to to make decisions — to choose

between two courses of action, one appropriate under the hypothesis and the other under the alternative. Hypothesis testing is poorly suited for this task (Lindley, 1998).

If we subscribed wholeheartedly to the strictures of the method, such as transparently formulating (or declaring) every hypothesis prior to data inspection, not following up the verdict of 'failure to reject' a null-hypothesis by assuming that the null is valid, and the like, we would find most applications of hypothesis testing in journals of applied statistics as flawed. I would disqualify all applications of hypothesis testing on the grounds that it is oblivious to the consequences of the two kinds of error that may be committed. The only good reason for not exploring the consequences is that they are not important, but then the application is inconsequential, and should also be disqualified.

Hypothesis testing is an important invention in statistics, but its practice has largely ossified (and its application became ritualised), not keeping up with or adapting to the revolutionary changes in the scientific environment in the last few decades. I believe that the rightful place of hypothesis testing in its present form, in statistics as a science and profession that aspires to be relevant throughout the society, is more in history books than in textbooks and other publications documenting contemporary research or practice.

Of hypothesis tests and steam engines

I heard the other day that a leading journal in social psychology has banned hypothesis testing from its articles (Trafimow and Marks, 2015). The editorial policy statement rather audaciously claims that hypothesis testing is 'invalid' and entails a stultifying structure that hinders creative thinking. I mentally leaf through the statistics journals I occasionally read and count the papers that do not use any hypothesis testing. After a while, I find one, but before finding the second I give up and turn my thoughts to how should I, a self-respecting statistician with some academic pretensions, respond to this assault on the bread-and-butter of my profession. This is the diplomatic equivalent of country B declaring war on the United Kingdom because the 'Brits' drive on the left and speak deplorable English. Let's check first on the B's alliances and, if need be, shore up any small divisions which we might have with our most potent allies, the U.S.A. in particular.

This is an attack on the foundations of statistics, an affront to the heritage of Fisher, Neyman, the Pearsons, Lehmann and their illustrious peers. Would anybody in their right mind suggest that Michael Faraday, James Watt, George Stevenson or Thomas

A. Edison were second-rate tinkers and charlatans? Imagine banning electricity for just a day. And here I come uncomfortably close to being treasonous. Steam engines are not exactly banned, but became obsolete soon after the inventions of the combustion engine, the electric locomotive, and the like. The technology implemented in much of the lighting used today is only distantly related to the original light bulb. Yet, we use hypothesis testing very much like Sir Ronald Fisher demonstrated how it might be applied, asking the reader to attend to the minutiae. I think that this we have collectively failed to do. Hypothesis testing has become the tradition, a convention not to be questioned, a stationary point in the statistical toolkit, while the rest of the science, technology, industry and the society at large has moved on. The modern computer has helped us to do more tests, using a wider variety of test statistics, but in some ways to abuse the original idea more effectively and more frequently. Hypothesis testing is a product of the pre-computing age that survives, a bit like the iguana and the denizens of the Galapagos Islands.

Still, this is outrageous. How does a specialist in psychology dare say that a statistical method is flawed? Such a statement, or verdict, should originate from the 'expert', that is, a statistician, and preferably a theoretical one. But today's theoretical statistics is concerned with matters much more important than some hypothesis testing which has for all purposes been sorted out a long time ago. How should we respond to this 'insult' from a psychologist? Ignore it and wait for developments? Mobilise our forces to defend the Realm? Take up an aggressive stance and prepare for a campaign to 'correct' the enemy's views by peaceful means, and if that turns out to be infeasible, then annihilate it? In other words, build new logic and spread science throughout the world? We have the staunch support of the army of basic statistics textbooks, lecturers of Stats 101 courses, statisticians employed in pharmaceutical research and development, all those promoting evidence-based policy and lots more. A glossy journal published jointly by the two leading statistical societies in the World bears the name of the principal keyword associated with hypothesis testing — *Significance*. We can count on its editorial board and readers. Organised force against it is comparable to a 'lunatic fringe' and one or two misguided and insignificant pacifists.

Without labelling myself openly, I propose to speculate, but preferably analyse, the possible precursors to this conflict. Basic statistics textbooks on hypothesis testing contain all manner of caution against the misuse of hypothesis testing and misinterpretation of its results. But by the time a graduate comes to the responsible task of applying a test, its mechanics become the main preoccupation, together with the (politically) correct wording of the conclusion and a discussion that is intended to smooth

the way to acceptance of the report or manuscript by the boss or an editor. The caution is difficult to exercise. By way of an example, suppose the test of a null-hypothesis concludes with 'failure to reject' the hypothesis. The conclusion that we have learned nothing in this case is singularly disagreeable and obstructionist. But at the same time, it is correct if you agree with the following description of testing a null-hypothesis.

We start by presuming the null-hypothesis. Presumption is not a belief, but the cooperating response to the request to imagine and explore what would happen if ... So, if the null-hypothesis were valid, then certain kinds of outcomes of a study would be expected (would be unexceptional) and others would be exceptional. An exceptional outcome constitutes a statistical (or probabilistic) contradiction with the null, or 'evidence against' it. Unexceptional outcomes should be interpreted as: we tried to find a contradiction, but failed. The premise may be correct, but does not have to be.

We admit that our conclusion may be incorrect, because an exceptional outcome would be possible even if the null were valid, but the chances of that are probabilistically controlled. The null-hypothesis cannot be confirmed — we all know that. But how often do we follow up 'failure to reject' by acting as if the null were valid? In model selection based on a test, such as the likelihood ratio, or an information criterion, that is exactly what we do. In model diagnostic procedures, we presume that the contemplated model is valid, and if we fail to find any contradiction with it, then we proceed as if the model were valid. 'Seen nothing' becomes 'There is nothing'. This step is deeply flawed. In the analysis of an experiment, the statement that the effect is small is tolerated by many journals. The statement 'No effect was found.' is quite common, inviting the faulty interpretation that there is no effect. So, maybe the method is correct, but the analyst is at fault and the language used to communicate the results is ambiguous.

This would be too early to rest my case. My argument against testing any hypothesis rests on the following statement, formulated as a war-cry of the B's government mouthpiece:

> In modern statistics, hypothesis test has absolutely no role because it is oblivious to the consequences of the two kinds of error that can be committed.

Hypothesis testing is appropriate if we are not interested in such consequences (ramifications), but then the whole analysis is inconsequential, and its conduct can hardly be regarded as a professional activity; it has no scientific value. It could still be published, but only in JIM (*Journal of Irrelevant Matter*). The consequences may be difficult to

elicit, but that is mostly a 'cultural' problem. The client, from whom they should be elicited, is often a distant impersonal entity that expects to obtain an answer in exchange for data and its background and details. The perspective of the client is wrongly regarded as the client's own business. Yet, it is an important input to a (statistical) analysis that is tailored to the client's needs, perspectives and priorities (and every analysis should be); Longford (2013). Without such elicitation, a hypothesis test caters for a certain perspective, but we do not care to assess whether it is realistic. That is a good case for a ban.

As part of the build-up to war, we should ask the military intelligence to explore the background to its declaration. As a referee and a (co-)author of manuscripts submitted to journals like the offending one, I have experienced the polarisation of the relevant community to those who swear by the p-value and those who regard it as worthless — as a menace to the scientific process. The selective attention to the various caveats and warnings has created a very divisive atmosphere among applied statisticians in psychology and related fields. Opinions range from the norm of attaching a p-value to every single data-based statement to acknowledging that a p-value or the verdict of a hypothesis test is a worthless ritual with value only for one's professional existence. These closet-dissindents are waiting for a version of the 1989 velvet revolution or the 2011 Arab spring. As a result of trying to negotiate this divide in the publication process and in other professional discourse, the research psychologists' clinical priorities have been derailed. A ban on hypothesis testing may well be a measure taken to bring the subject back to a mix of good reason and basics. In some journals, the referees' names are disclosed, and authors sometimes lodge complaints with the editor who then plays the role of the arbitrator. In a few disputes I have been involved, all parties, including myself, were abjectly incorrect, in obvious contradiction with one or another point prominently made in any quality elementary statistics textbook. Simply, logic and integrity would not be constructive, as no party would stoop down to discuss the elementary statistics background.

As the focus of the statistical Commonwealth, of which quantitative psychologists are subscribing members, we should look on their difficulties as if they were also ours. In the simile I have used earlier, hypothesis test is a steam engine, a great invention that has done valiant service for the development of our academic subject and profession in the past, just like Newton had done for physics and ultimately for the industrial revolution. Later, Newton was proven flawed, most notably by Einstein. In no way does it diminish his achievements, or those of James Watt, even though steam-engines are nowadays marginalised to museums and clubs indulging in industrial nostalgia. We

should treat hypothesis testing and its inventors with similar respect and assist country B in its current difficulties by mutually agreeable and peaceful means that amend or replace the practice that brought about a constitutional crisis with a response that is decisive but questionable. A critical review of the practices in un-applied statistics would be a bonus.

References

Lindley, D.V. (1985). *Making Decisions.* Wiley, Chichester, UK.

Lindley, D. V. (1998). Decision analysis and bioequivalence trials. *Statistical Science* **13**, 136–141.

Longford, N.T. (2013). *Statistical Decision Theory.* Springer-Verlag, Heidelberg, Germany.

Trafimow, D. (2014). Editorial. *Basic and Applied Social Psychology* **36**, 1–2.

Trafimow, D., and Marks, M. (2015). Editorial. *Basic and Applied Social Psychology* **37**, 1–2.