Journal of Modern Applied Statistical Methods

Volume 15 | Issue 2 Article 5

11-2016

Some Remarks on Rao and Lovric's 'Testing Point Null Hypothesis of a Normal Mean and the Truth: 21st Century Perspective'

Bruno D. Zumbo
University of British Columbia, bruno.zumbo@ubc.ca

Edward Kroc
University of British Columbia

Journal of Modern Applied Statistical Methods
November 2016, Vol. 15, No. 2, 33-40.
doi: 10.22237/jmasm/1478001780

ISSN 1538 − 9472

Some Remarks on Rao and Lovric's 'Testing Point Null Hypothesis of a Normal Mean and the Truth: 21st Century Perspective'

Bruno D. Zumbo
University of British Columbia
Vancouver, BC, Canada

Edward Kroc
University of British Columbia
Vancouver, BC, Canada

Although we have much to agree with in Rao and Lovric's important discussion of the test of point null hypotheses, it stirred us to provide a way out of their apparent Zero probability paradox and cast the Hodges-Lehmann paradigm from a Serlin-Lapsley approach. We close our remarks with an eye toward a broad perspective.

Keywords: Hypothesis testing, point null, statistical practice

Statistical methods and the testing of hypotheses play a pivotal role in day-to-day practical science, but not always an enlightened one. There are several well-known criticisms of testing a point null hypothesis in the statistical literature that go back at least to Berkson (1938, 1942) and Hodges and Lehmann (1954). Debates about the role of statistical hypothesis testing, its uses, misinterpretations, and abuses as well as adjacent discussions of interpretations and abuses of confidence intervals, effect sizes, and statistical power continue unabated in the methodological, statistical and substantive literatures. Through all of this, however, conventional significance tests, point null hypotheses, and p-values continue to be used in nearly all experimental publications in the social, behavioral, natural, and health sciences to dichotomize claims from statistical hypotheses in to significant versus nonsignificant findings. The use of significance tests of point null hypotheses, as a kind of ritualistic cultural behaviour, continues unabated because these statistical techniques appear (at least to practicing scientists) to be objective and exact, they are easily and readily available in statistical software packages and on web applets, students are taught to use them, and journal reviewers and editors demand them.

Dr. Zumbo is the Paragon UBC Professor of Psychometrics and Measurement and an editor of this journal. Email him at bruno.zumbo@ubc.ca. Dr. Kroc is a post-doctoral Research Fellow in the Measurement, Evaluation, and Research Methodology Program.

Rao and Lovric's (2016, this issue) recent paper rests in this backdrop of our discipline's longstanding ineffective critical obsession to challenge and repurpose our most sacred of empirical methodological cows: the testing of point null hypotheses via significance testing. Rao and Lovric are to be most warmly thanked for bringing this fundamentally important issue to the attention of readers of the *Journal of Modern Applied Statistical Methods* and initiating an important conversation. Their recent contribution to the literature gives us much to agree with, but also stirs us to critically reflect on some of their claims and observations. This is distinctly a sign of good scholarship.

We have arranged our remarks in to three categories. In what follows we: (i) reserve the majority of our remarks for Rao and Lovric's point null Zero probability paradox and the matter of events of vanishingly small probability ultimately happening, a key point in the on-going controversy surrounding testing point null hypotheses, (ii) bring what we call the Serlin-Lapsley perspective to the Hodges-Lehmann paradigm briefly attending to its strengths and limitations, and (iii) close with some remarks that aim to move us to a broader perspective.

Rao and Lovric's point null Zero probability paradox, and on the event of vanishingly small probability ultimately happening

As Rao and Lovric remind us, Hand (2014, p. 6) stated: "extremely improbable events are commonplace. It's a consequence of more fundamental laws, which all tie together to lead inevitably and inexorably to the occurrence of such extraordinarily unlikely events". We are in agreement, per Kolmogorov (1956), that the probability of an event A being zero does not imply that the event A is impossible. Indeed, it is the *support* of a probability measure that separates the possible from the impossible, not the value of the measure on its support. It is true, for example, that the probability of observing the event $\{X = 1\}$ is zero when X is an exponential random variable, but that the event $\{X = 1\}$ should not be considered impossible, since the measure is well-defined and nonzero on any open set containing this event. However, the event $\{X = -1\}$ is indeed ontologically impossible when X is an exponential random variable; this event is simply not in the support of X.

We must disagree though with Hand's claim that "events of vanishingly small probability will ultimately happen." This is not true in general, at least, not if an event of "vanishingly small probability" is to be interpreted as an event that is almost surely null; i.e., an event whose probability is equal to zero. Broadly

speaking, most random variables of practical interest fall into one of two categories: they are defined by measures that are either (1) absolutely continuous with respect to Lebesgue measure, usually defined on the real line or half line; or (2) absolutely continuous with respect to counting measure on some countable set, usually the positive or nonnegative integers. We remind the reader that a measure μ is absolutely continuous with respect to another measure λ if every λ-null set is also a μ-null set.

We note that absolute continuity implies the existence of a probability density function, case (1), or a probability mass function, case (2), by the classical Radon-Nikodym Theorem; indeed, this is precisely how these objects are formally defined. We also note that this definition presupposes the specification of a legitimate σ-algebra, and it suffices to take the Borel sets on the real line, or the power set on the integers respectively. With this terminology in mind, we will see that Hand's statement is false when our probability measure is absolutely continuous with respect to Lebesgue measure, and unnecessary when our measure is absolutely continuous with respect to counting measure.

Let X be distributed according to a probability measure, Pr, that is absolutely continuous with respect to Lebesgue measure. Choose any real number a. What is the probability that we eventually sample $\{a\}$? Formally, if we let X_i denote the i^{th} (independent) sampling, we wish to calculate the probability of the union of events $\{X_i = a\}$ over all $i > 0$. Apply countable subadditivity of the measure (a defining property of measures) to bound this probability by the sum of $\Pr(\{X_i = a\})$. Each of these is identically zero (by absolute continuity), therefore the probability of their union is as well. Thus, we can sample infinitely often and we will in fact *never* sample the singleton $\{a\}$, almost surely. This argument immediately generalizes to any countable set, which is automatically a Pr-null set by absolute continuity. So, for example, the probability that we ever observe *any* rational number in infinitely many samples of X is zero. The argument can be fully extended to apply to any Lebesgue-null event, including those containing certain uncountable sets of reals, such as Cantor or various other fractal sets.

Two key points are noteworthy about the above argument. First, we take as definition that any sampling scheme must consist of a countable number of steps. That is, we do not allow the possibility of drawing uncountably many samples. Theoretically, this kind of uncountable sampling scheme is not impossible, but it would be completely meaningless in practice: any mechanistic process requires countability of its steps.

The second point to note is that the theoretical argument above relies on the infinite precision of our sample, and this is where the crux of the matter lies. A

careful reading of the above argument will reveal an apparent paradox: the probability of ever observing *any* rational number under a probability measure, absolutely continuous with respect to Lebesgue measure, is identically zero; yet, in practice, every singleton sample that we draw from such a distribution will be a rational number. This is simply another, equivalent instantiation of what Rao and Lovric term the Zero probability paradox. Any practical measuring device will demand that a sampled point is drawn to only a finite level of precision; i.e., we can only observe real numbers with finite decimal expansion in practice. The way out of this apparent paradox is to realize that all probability measures, *in practice*, are only supported on a finite set. The size of this set is dictated by the precision of our measurements, but we know that this precision must always be of finite detail. Consequently, if we choose any real number a in the support of our practical probability measure, Pr*, we have $\Pr^*(X = a) > 0$; this follows since any finite set of real numbers, under the classic topology, is nowhere dense. Revisiting our generic sampling scheme from before, we now calculate:

$$\begin{aligned}\Pr^*\left(\bigcup_{i=1}^{\infty}\{X_i = a\}\right) &= 1 - \Pr^*\left(\bigcap_{i=1}^{\infty}\{X_i \neq a\}\right) \\ &= 1 - \prod_{i=1}^{\infty}\Pr^*\left(X_i \neq a\right) \\ &= 1 - \lim_{N\to\infty}\left[1 - \Pr^*\left(X_i = a\right)\right]^N\end{aligned}$$

The limit goes to zero since $\Pr^*(X_i = a) > 0$ for all i, so the probability that we eventually observe the singleton $\{a\}$ is exactly 1, almost surely. The same reasoning applies to any subset of the practical probability space.

This is the distinction between probability in practice, the ultimate subject of statistics, and the platonic structure of the mathematical objects that we use to conveniently describe that practice. These descriptions are nearly always approximations: we simplify our practical probability spaces by smudging them into theoretical ones. This has undeniably proven to be an extremely fruitful tactic, but it has also given rise to several conundrums and apparent paradoxes like the ones discussed here. Point null hypotheses may be almost surely false in the platonic sense, but this is only a reflection of the disconnect between the literal structure of the objects we study and the approximations, like the various scaled-Lebesgue measures, that we use to conveniently describe them mathematically. It is meaningful effects that we truly care about, relative to the precision of our

measurements and the object of our research, and in this regard we are very much in agreement with Rao and Lovric.

Finally, we note that this entire discussion is unnecessary when considering random variables that are absolutely continuous with respect to counting measure (all the standard "discrete" distributions, for example). By definition, such a corresponding probability space contains *no* nontrivial null sets in the support of the measure; thus, there are no events of "vanishingly small probability" to speak of. Just as in the resolution of our apparent paradox on a practical probability space above, every event will eventually happen almost surely.

Hodges-Lehmann Paradigm with a Serlin-Lapsley Twist

We would agree with Rao and Lovric that the Hodges-Lehmann method is not a magical alternative to the traditional point null testing but that it may provide a useful paradigm for the practicing empirical scientist. However, we would contend that in its day-to-day use among empirical researchers the Hodges-Lehmann paradigm still suffers from some of the same issues as the point null. In particular, the magic choice of "delta", in the Hodges-Lehmann or Serlin-Lapsley senses, remains arbitrary – or necessarily defined subjectively by the researcher, contingent on precision, etc., as before. Furthermore, a key to the widespread adoption of the Hodges-Lehmann paradigm is what we will refer to as the Serlin-Lapsley approach to statistical science that incorporates a ‘good enough’ principle and embodies Imre Lakatos’s view of science. Our message is the same as Rao and Lovric’s but from a different framework.

Efforts to facilitate testing what may be called ‘range nulls’, which require assumptions about the distribution of a statistic when the null is false, have been made by Serlin and Lapsley (1985, 1993). In short, this approach involves incorporating an external criterion or statistic, such as an effect size, into the hypothesis test via a range-null hypothesis approach. The kernel of the ‘range-null hypothesis approach’ specifies a range of values under the null hypothesis for which a rejection implies a meaningful result. One tests against a negligibly small or trivial effect. If one rejects the null range hypothesis this implies not only that, for example, the mean of the experimental group is different than control group, but the difference is of large enough magnitude to be meaningful. As Serlin and Lapsley (1985) note, minimum effects testing can test more realistic hypotheses, rather than the “straw man” zero effect (p. 74). The important difference in terms of scientific practice is that Serlin and Lapsley’s (1985) framework focuses on testing one’s own theory as the null, along with using what they call a “good-

enough belt" around a "complex null hypothesis" (p. 79). Central to this paradigm is the principle that this "good-enough belt" be defined subject to the analyst's particular research question, knowledge from the previous literature, and the precision of measurement. One machinery for applying Serlin and Lapsley's framework is the Hodges-Lehmann paradigm described in Rao and Lovric or one could, as Serlin and Lapsley (1985) do, construct a test criterion by directly computing the percentile of the noncentral distribution involved in the test of the hypothesis to set the critical value – for example, the noncentral F distribution in a situation similar to the one described in Rao and Lovric.

Closing Remarks

In closing, Rao and Lovric's paper highlights for us the continuing need for dialogue on conceptual and foundational matters in statistics. The statistical significance test is the most widely known point in empirical science wherein probabilities and probability models enter the scientific process as either a platonic structure of the mathematical objects or a practical mathematical model. In this light, the nature, use and misuse of significance tests have been widely discussed in both statistical and non-statistical circles. Clearly for significance tests to be of much use to empirical researchers they must focus on sensible and interesting null hypotheses. As is widely discussed in the methodological literature and growing in importance in the statistical literature, there are clear distinctions between statistical significance and the more important notions of practical, clinical or biological significance. Likewise, we need to move beyond the conventional language of Type I and Type II error rates and also consider errors that are directly related to day-to-day statistical practice such as Type S (sign) and Type M (magnitude) errors, which Gelman and Tuerlinckx (2000) describe as relating to the probability that claims with confidence have the wrong sign or are far in magnitude from underlying effect sizes. These errors speak more directly to quantifying our subjective assumptions about what matters and what does not.

Highest among our concerns is that there is a misunderstanding among some experimental researchers that statistical theories of hypothesis testing (be they of the Fisherian, Neyman-Pearson, or some blended approach of the two frameworks) are intended to give an automated and (naively) objective support to an empirical claim. This misunderstanding reflects a lack of alignment of statistical and scientific reasoning. Cox (1982) stated the matter best:

"Failure to achieve an interesting level of significance in a study does not mean that the topic should be abandoned. Significance tests are not intended to inhibit the free judgment of investigators. Rather they may on the one hand warn that the data alone do not establish an effect, and hence guard against over interpretation and unwarranted claims, and on the other hand show that an effect is reasonably firmly proved." (Cox, 1982, p. 327).

References

Berkson, J. (1938). Some Difficulties of Interpretation Encountered in the Application of the Chi-Square Test. *Journal of the American Statistical Association, 33*(203), 526-542. doi: 10.2307/2279690

Berkson, J. (1942). Tests of Significance Considered as Evidence. *Journal of the American Statistical Association, 37*(219), 325-335. doi: 10.1080/01621459.1942.10501760

Cox, D. R. (1982). Statistical Significance Tests. *British Journal of Clinical Pharmacology, 14*(3), 325-331. doi: 10.1111/j.1365-2125.1982.tb01987.x

Gelman, A., & Tuerlinckx, F. (2000). Type S error rates for classical and Bayesian single and multiple comparison procedures. *Computational Statistics, 15*, 373–390. doi: 10.1007/s001800000040

Hand, D. (2014). *The improbability principle: why coincidences, miracles, and rare events happen every day*. NY: Scientific American / Farrar, Straus and Giroux.

Hodges, J. L. & Lehmann, E. L. (1954). Testing the approximate validity of statistical hypotheses. *Journal of the Royal Statistical Society. Series B (Methodological), 16*(2), 261–268.

Kolmogorov, A. N. (1956). *Foundations of the theory of probability* (Second English edition). NY: Chelsea.

Rao, C. R., & Lovric, M. M. (2016). Testing Point Null Hypothesis of a Normal Mean and the Truth: 21st Century Perspective. *Journal of Modern Applied Statistical Methods, 15*(2), 2-21. doi: 10.22237/jmasm/1478001660

Serlin, R. C., & Lapsley, D. K. (1985). Rationality in psychological research: The good-enough principle. *American Psychologist, 40*(1), 73-83. doi: 10.1037//0003-066x.40.1.73

Serlin, R. C., & Lapsley, D. K. (1993). Rational appraisal of psychological research and the good-enough principle. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 199-228). Hillsdale, NJ: Erlbaum.