Journal of Modern Applied Statistical Methods

Volume 15 | Issue 2 Article 4

11-1-2016

Rao-Lovric and the Triwizard Point Null Hypothesis Tournament

Shlomo Sawilowsky
Wayne State University, professorshlomo@gmail.com

Journal of Modern Applied Statistical Methods
November 2016, Vol. 15, No. 2, 22-32.
doi: 10.22237/jmasm/1478001720

ISSN 1538 − 9472

Rao-Lovric and the Triwizard Point Null Hypothesis Tournament

Shlomo S. Sawilowsky
Wayne State University
Detroit, MI

The debate if the point null hypothesis is ever literally true cannot be resolved, because there are three competing statistical systems claiming ownership of the construct. The local resolution depends on personal acclimatization to a Fisherian, Frequentist, or Bayesian orientation (or an unexpected fourth champion if decision theory is allowed to compete). Implications of Rao and Lovric's proposed Hodges-Lehman paradigm are discussed in the Appendix.

Keywords: true null hypothesis, Rao-Lovric, Hodges-Lehman.

In their historical reviews of experimental design, Cochran (1977) and Frank Yates posited the first planned controlled experiment was conducted by Daniel (7^{th}–6^{th} century BCE), who employed a ten day treatment vs comparison group post-test only trial. The purpose was to demonstrate the efficacy of a Kosher diet of high protein, low fat, dried legume seeds and water on soldiering skills vs Nebuchadnezzar's army's royal comestible of non-Kosher wine and meat (Daniel 1:3-16). In Contra Celsus (1:15), Origen of Alexandria (153–253 CE) cited Hermippus (5^{th} century BCE) and Hecatæus (4^{th} century, BCE, presumably of Abdera) who opined subsequent development of analytical analyses of experimental principles by the Jews influenced, if not culminated in, Pythagoras' philosophy of mathematical sciences. Subsequently, Tana Kama (Mishna Gittin 7:1; Talmud Gittin 67b) underscored the importance of co-variables and the minimum number of repetitions for a reliable single subject study design. Shimon ben Chalafta also invoked experimental replications to test claims (e.g., Talmud Chulin 57b).

In the middle of the 2^{nd} century CE, Galen (Aelius/Claudius Galenus) mused how much credence should be given, if any, to a 50^{th} medical study if the previous 49 replications were of no significance. In the early 11^{th} century CE,

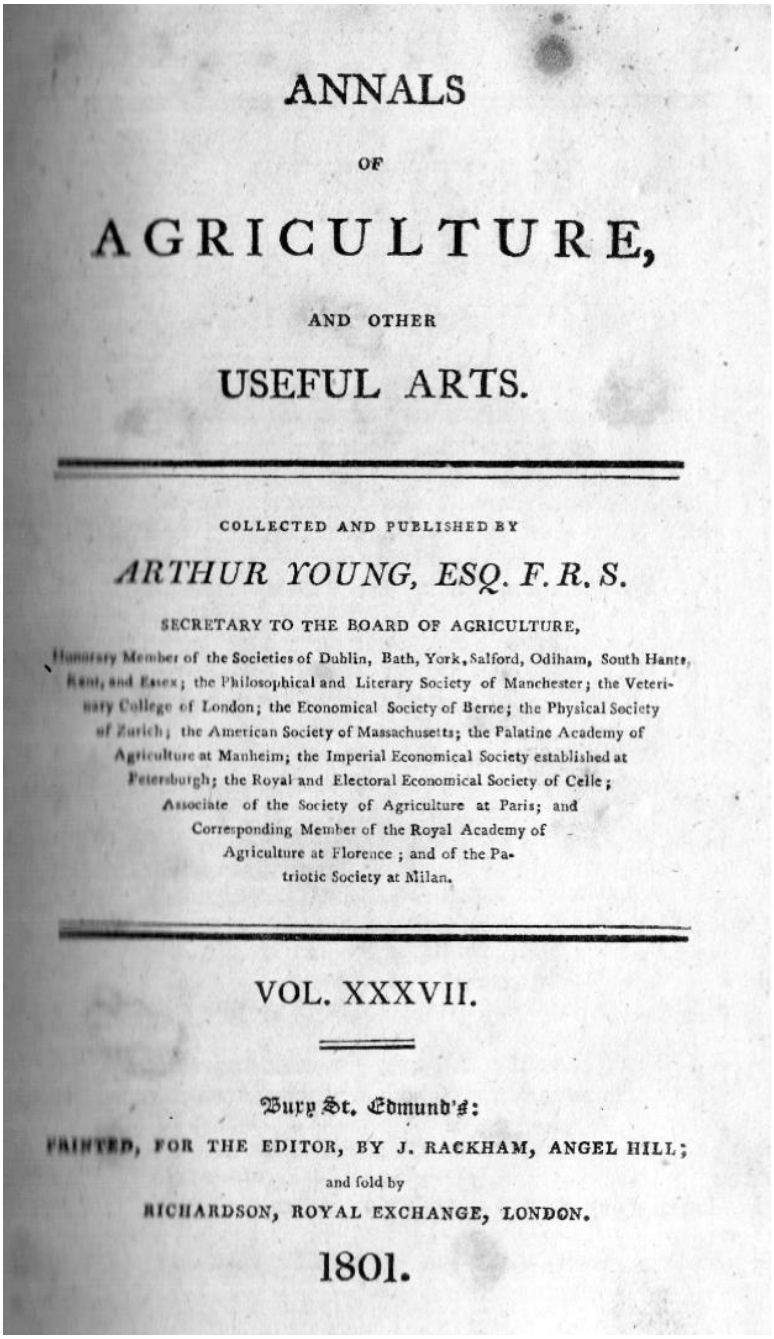
Dr. Sawilowsky is a Professor in the College of Education and the founding editor of this journal. Email him at professorshlomo@gmail.com.

Avicenna (Abu ibn Sina) reacted to haphazard methods in the conduct and analysis of experiments and presented seven governing rules. In 1266 CE, Roger Bacon systematized observation of empirical data in controlled experiments. Arthur Young (1771, Figure 1) published a course on experimental agriculture, wherein comparative designs employing standardized methods and analyses were proposed. The analysis of the hypothesis "every year there shall be born more males than females" (1710-1712, p. 188) by John Arbuthnott (un-admittedly inspired by Sir William Petty & John Graunt) is considered the origin of the nonparametric Sign Test, although it predates more formal origins of empirical probability captured in the treatises on the doctrines of conjecture and chance by Jacob Bernoulli (1713), Abraham de Moivre (1718) and Thomas Bayes (Price, 1763, p. 370).

In the early part of the 20th century CE, Sir Ronald Fisher (influenced by Pierre-Simon Laplace, Carl Gauss, Joseph Jastrow, Sir Francis Galton, Karl Pearson, G. Udny Yule, William Gosset, and certainly others; perhaps later also with Andrey Kolmogorov & E. J. G. Pittman) defined the null hypothesis, the fundamental building block of modern hypothesis testing, as being true unless there is evidence from the sample (randomly obtained or data at hand) to the contrary. His innovations regarding blocking variables and factorial layouts were pioneering developments in the design of experiments.

Following the logic of experimentation by C. S. Peirce in late 19th century, the Frequentist lemma by Jerzy Neyman and Egon Pearson developed in the 1930s-1940s violated the Fisherian cannon with the introduction of the alternative hypothesis. It was indeed irrefragable blasphemy, because Frequentists must admit the choice and magnitude of the alternative are subjective and independent of both the null hypothesis and the sample. Other 20th century developments in experimental design included orthogonal arrays by my esteemed colleague Professor C. R. Rao, sequential experiments by Abraham Wald and later Herman Chernoff, and the quality control designs of Genichi Taguchi.

Nevertheless, the Frequentists had the advantage, because in the Fisherian system the lack of an alternative obviated the desired notion of fixed comparative statistical power, and by extension, stable effect size. These two modern approaches to statistics are antipodal. Many misunderstandings in hypothesis testing are due to their intrinsic incompatibility, starting with Sir Fisher's "lapsus linguae" (Neyman, 1941, p. 129) fiducial argument (see Sawilowsky, 2003).



ANNALS

OF

AGRICULTURE,

AND OTHER

USEFUL ARTS.

COLLECTED AND PUBLISHED BY

ARTHUR YOUNG, ESQ. F. R. S.

SECRETARY TO THE BOARD OF AGRICULTURE,

Honorary Member of the Societies of Dublin, Bath, York, Salford, Odiham, South Hants, Kent, and Essex; the Philosophical and Literary Society of Manchester; the Veterinary College of London; the Economical Society of Berne; the Physical Society of Zurich; the American Society of Massachusetts; the Palatine Academy of Agriculture at Manheim; the Imperial Economical Society established at Petersburgh; the Royal and Electoral Economical Society of Celle; Associate of the Society of Agriculture at Paris; and Corresponding Member of the Royal Academy of Agriculture at Florence; and of the Patriotic Society at Milan.

VOL. XXXVII.

Bury St. Edmund's:

PRINTED, FOR THE EDITOR, BY J. RACKHAM, ANGEL HILL;

and sold by

RICHARDSON, ROYAL EXCHANGE, LONDON.

1801.

Figure 1. Arthur Young (1801), *Annals of Agriculture and other Useful Arts*, Vol 37. London: Rackham & Hill. (From the *JMASM* Archives.)

This struggle provided the segue for a Bayesian resurrection from Fisher's epithet, "From a purely historical standpoint it is worth noting that the ideas and nomenclature for which I am responsible were developed only after I had inured myself to the absolute rejection of the postulate of Inverse Probability" (1937a, p. 151; see also 1937b, 1939). Although also receiving a boost from C. S. Peirce's logic, Bayesian analysis during Sir Fisher's reign was conducted without benefit of his development of degrees of freedom. The initial inability to replicate Fisherian/Frequentist numerical results was a serious setback to the modern Bayesian paradigm (Sawilowsky, 2002, 2003). Although they have since recovered and inverse probability is currently quite popular, unless there are documented informative prior probabilities available, such as baseball batting averages, Fisher's inurement prevails.

Now comes the debate on certifying the literal truth of the null hypothesis. Original Fisherians needs no proof, because postulation of the putative null was the pivotal theoretic spanning well over two millennia in the science of discovering truth. Frequentists, however, can never accept any proof. The most that can be said is based on the current sample there is no evidence to support the alternative. (This should not be considered an open invitation to collecting potentially endless (a) random samples, known as the quest for a Type I error and its attendant rewards of publishing and tenure or (b) data sets at hand, known as non-representative findings never interpreted with caution to support situational truths with its attendant rewards of political fodder, ill-begotten relief from the court, financial returns based on false advertising, etc.) *Moreover, it wouldn't matter even if the null hypothesis is always literally false, because it must be false to an a priori specified magnitude to be rejected.*

The Frequentist nomenclature, failure to reject the null hypothesis, was just the ticket in the social and behavioral sciences, where politically correct thinking of the 1960s had begun to take control of those in charge of the keys to situational truths. At best, near-null, near-nil, and the like, were approved substitutes. Philosophically, the yellow submarine is a closed system, so at some decimal of the mantissa there must be a non-Zero value.

The various Frequentist counterproofs were flawed attempts to make something out of nothing by incorrectly preserving the post hoc effect size even when the statistical test was not significant. For example, in the two sample layout, the *t* statistic is a test of difference between two means. If the *p* value is above the a priori selected nominal α level, it means the observed difference is not real and should be read as zero. Based on the sample, assumed to be random for generalization purposes, there is no evidence that the populations from which they

were drawn differed in terms of location. Just as the observed difference in means can be safely ignored, the effect size was not statistically significantly different from zero, and can be safely ignored.

This means regardless of the magnitude of the obtained value (e.g., Cohen's *d*, 1962, 1969, 1977, 1988) in the two sample layout [from very small (0.01; Sawilowksy, 2009) to small (0.2; Cohen, 1988) to moderate (0.5; Cohen, 1988) to large (0.8; Cohen, 1988) to very large (1.2; Sawilowsky, 2009) to huge (2.0; Sawilowsky, 2009)], it should be read and interpreted as zero. Hence, the point null hypothesis, to the Fisherian, is indeed considered to be literally true regardless of the magnitude of Cohen's *d* when the *p* value is greater than nominal α.

In the antecedent article, colleagues C. R. Rao and M. Lovric (http://digitalcommons.wayne.edu/jmasm/vol15/iss2/2), cited Cohen (1990) who wrote the null hypothesis can only be true "in the bowels of a computer processor running a Monte Carlo study (and even then a stray electron may make it false)" (p. 1, 308). Based on my letters with him, documented elsewhere, Cohen's statement was not surprising.

Subsequently, this was discussed conceptually in Knapp and Sawilowsky (2001, p. 71-74; for expanded commentary relative to the debate see Harlow, et al., 1997; Imbens & Rubin, 2015). I included Meehl's (1990) recapitulation that he initially referred only to quasi-experiments and surveys (Meehl, 1978), but later admitted the null hypothesis can be literally true in an "experimental study" (Meehl, 1990, p. 204). (Carol H. Ammons, the co-Editor of *Psychological Reports* where it was published, sent me a reprint of Meehl (1990) soon after its publication. In our subsequent conversation, I was supportive of Meehl's recapitulation, and I remain so today.) Similarly, in Knapp and Sawilowksy (2001) I also included Hagen's (1997, p. 20) imputed recapitulation of Cohen (1994).

A simple demonstration of the algorithm I presented in Knapp and Sawilowsky (2001) is coded in R in Figure 2. When executed, it creates two groups, *x* and *y*, and populates them with scores randomly selected from the standard normal curve. Although a Monte Carlo is unnecessary when underlying assumptions are met, it is employed to facilitate the demonstration. The two independent samples pooled variance *t* test is conducted on the data, and if the *p* value is less than nominal $\alpha = 0.05$, a counter is incremented. The process is repeated 100,000 times. The final value of the counter is divided by the number of repetitions to produce the Type I error rate.

The code will produce the same result on any computer platform and operating system, because the seed number is set for the pseudo-random number generator. That result is 0.04919. Rejections occurred across the 100,000 repetitions, but they were known false positives. The point null hypothesis was indeed literally true, because it was programmed to be so. The collection of false positives that give rise to the notion the point null is never literally true were simply the constituent figments of imagination that sum to the Type I error rate.

```
set.seed (123457)
to5 <- NULL
rep <- 100000
rejt05 <- numeric(length=rep)
ss <- 30
for (i in 1:rep) {
   x1 <- rnorm(ss)
   x <- x1+0.0
   y <- rnorm(ss)
   tp <- t.test(x,y,var=TRUE)[["p.value"]]
   rejt05[i] <- ifelse (tp < 0.05,1,0)
 }
t05 <- sum(rejt05)/rep
```

Figure 2. Monte Carlo *t* Test in R Code

The rejection rate obtained from the code will approach 0.05 as (a) the sample size, set to 30 per group in this example, increases, (b) the number of repetitions of the experiment increases, or (c) possibly even with the current study parameters if a different initial seed number is selected (Hill & Sawilowsky, 2011). For example, if the number of repetitions is increased to 1,000,000, the Type I error improves to 0.049858.

A non-null condition can be created by replacing the 0.0 with a non-zero number (positive or negative) in the line `x <- x1 + 0.0`. For example, to model a very small effect size of 0.01 (Sawilowsky, 2009), replace the 0.0 in this code segment with a constant $c = 0.01$ (representing $0.01*\sigma$; where σ refers to the standard deviation of the normal curve = 1). The constant c is added to each member of the x group and shifts its location by that magnitude, while leaving the

scale unaffected. The resulting rejection rate is known as statistical power (not Type I error rate). With 100,000 repetitions it amounts to 0.04923, a nuanced but detectable difference of 0.00004 above nominal α for this sample size and data pseudo-randomly sampled from the standard normal curve.

If the effect size is increased to 0.05 the power yield increases to .05342, and for an effect size of 0.1 the power increases to 0.06542. For Cohen's (1988) small effect size of 0.2, the power increases further to 0.11611. As the effect size approaches infinity (and depending on the distribution and sample size, the effect size may not need to increase past a small fraction or multiple of its σ) the power approaches 1.

Random numbers represent a literally true null condition. This R code proves that when the point null is literally true, the *t* test (if all conditions are met, i.e., normality, homoscedasticity, independence) will retain the null hypothesis to the nominal α level. Hence, in real world applications of a true randomized experimental design, if there is no difference between $\bar{x}$ and $\bar{y}$ (the two sample means) the *t* test will testify to that fact.

Execution of the R code demonstrates increasing the sample size and/or number of repetitions of the experiment to ∞ will not lead to a rejection rate of the null hypothesis different from nominal α, which is the answer to Cohen's speculation of what might happen in the bowels of a Monte Carlo study. Moreover, despite the current fascination with big data (and hopefully its ardent fans are able to recognize and deprecate its often hidden or embedded stepwise methods), Gosset noted many in applied disciplines we are forced to work with small samples. This was aptly captured in Sir Fisher's revelation to Samuel Stouffer regarding the inspiration for deriving a certain postulate: something had to be done when rabbits got into the garden and ate a lot of the degrees of freedom.

To the Fisherian, QED. To the Frequentist, the discussion is much ado about (something that can never be literally) nothing. To the Bayesian, add non-informative priors to the perils of non-normality, heteroscedasticity, and non-independence; and then choose sides.

References

Arbuthnott, J. (1710-1712). An argument for divine providence, taken from constant regularity observ'd in the births of both sexes. *Philosophical transactions* (1683-1775), *27* (1710-1712), 186-190.

Bernoulli, J. (1713). *Ars Conjectandi Opus Posthumum*. Basileæ: Thurnisiorum.

Cochran, W. G. (1977). *Sampling techniques*. 3rd ed. NY: Wiley.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal Social Psychology*, *65*(3), 145-153. doi: 10.1037/h0045186

Cohen, J. (1969). *Statistical power analysis for the behavioral sciences*. San Diego, CA: Academic Press.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences*, Rev. Ed. San Diego, CA: Academic Press.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*, 2nd ed. Hillsdale, NJ: Erlbaum.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, *45*(12), 1304-1312. doi: 10.1037/0003-066x.45.12.1304

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, *49*(12), 997-1003. doi: 10.1037/0003-066x.49.12.997

De Moivre, A. (1718). *The doctrine of chances: Or, a method for calculating the probabilities of events in play*. London: W. Pearson.

Fisher, R. A. (1937a). Editorial note. *Annals of Eugenics*, 7, 146-151.

Fisher, R. A. (1937b). On a point raised by M. S. Bartlett on fiducial probability. *Annals of Eugenics*, *7*(4), 370-375. doi: 10.1111/j.1469-1809.1937.tb02154.x

Fisher, R. A. (1939). A note on fiducial inference. *The Annals of Mathematical Statistics*, *10*(4), 383-388. doi: 10.1214/aoms/1177732151

Hagen, R. L. (1997). In praise of the null hypothesis statistical test. *American Psychologist*, *52*(1), 15-24. doi: 10.1037/0003-066x.52.1.15

Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (Eds.). (1997). *What if there were no significance tests?* Mahwah, NJ: Erlbaum.

Hodges, J. L., & Lehmann, E. L. (1963). Estimates of location based on rank tests. *Annals of Mathematical Statistics*, *34*(2), 598-611. doi: 10.1214/aoms/1177704172

Hill, J. C., & Sawilowsky, S. S. (2011). Bias in Monte Carlo simulations due to pseudo-random number generator initial seed selection. *Journal of Modern Applied Statistical Methods*, *10*(1), 29-50.

Imbens, G. W., & Rubin, D. B. (2015). Fisher's exact P-values for completely randomized experiments. In G. W. Imbens & D. B. Rubin, Eds.

Causal inference in statistics, social, and biomedical sciences. NY: Cambridge University Press. doi: 10.1017/CBO9781139025751.006

Knapp, T. R., & Sawilowsky, S. S. (2001). Constructive criticisms of methodological and editorial practices. *The Journal of Experimental Education*, *70*(1), 65-79. doi: 10.1080/00220970109599498

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, *46*(4), 806-834. doi: 10.1037/0022-006x.46.4.806

Meehl, P. E. (1990). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it. *Psychological Reports*, *1*(2), 108-141. doi: 10.1207/s15327965pli0102_1

Neyman, J. (1941). Fiducial argument and the theory of confidence intervals. *Biometrika*, *32*(2), 128-150. doi: 10.1093/biomet/32.2.128

Price. (1763). An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, F.R. S., communicated by Mr. Price, in a letter to John Canton, A.M. F.R.S. *Philosophical Transactions*, *53*, 370-418.

Sawilowsky, S. S. (2002). Fermat, Schubert, Einstein, and Behrens-Fisher: The probable difference between two means when $\sigma_1^2 \neq \sigma_2^2$. *Journal of Modern Applied Statistical Methods*, *1*(2), 461-472.

Sawilowsky, S. S. (2003). Deconstructing the arguments from the case against hypothesis testing. *Journal of Modern Applied Statistical Methods*, *2*(2), 467-474.

Sawilowsky, S. S. (2009). New effect size rules of thumb. *Journal of Modern Applied Statistical Methods*, *8*(2), 597 – 599.

Young, A. (1771). *A course on experimental agriculture; containing an exact register of all the business transacted during five years on near three hundred acres of various soils; including a variety of experiments on the cultivation of all sorts of grain and pulse, both in the old and new methods*. Dublin: Exshaw, Saunders, Chamberlaine, Sleater, Potts, Hoey, Williams, Porter, Moncrieffe, and Walter.

Appendix

In Knapp and Sawilowsky (2001), I presented rebuttals to "the following propositions:

- The null hypothesis is always false.
- A sufficiently large enough sample guarantees rejection of the null hypothesis.
- Statistical tests are of no use because the results do not address practical importance.
- Testing a near-nil null hypothesis is better than testing a null hypothesis.
- Hypothesis testing does not lead to scientific discoveries.
- Confidence intervals are superior to hypothesis testing.
- Effect sizes should be reported regardless of the outcome of hypothesis testing." (p. 71).

The subjectivity of defining a near-nil null hypothesis will also have a deleterious effect on equivalence testing, and could be added to the above list.

With regard to testing a near-nil null instead of a null hypothesis, Rao and Lovric, in the antecedent article, proposed a paradigm shift to testing the negligible null hypothesis:

$H_0 : |\theta - \theta_0| \leq \delta$ (Effect size is negligible) against
$H_1 : |\theta - \theta_0| > \delta$ (Effect size is practically meaningful).

They aptly named it the "Hodges-Lehmann paradigm," a nomenclature well known in other contexts. In R-measures of location, for example, the inversion of signed ranks can lead to the Hodges-Lehmann estimator, a robust (median unbiased) pseudo-θ point estimator of symmetry (Hodges & Lehmann, 1963). In bracketed (see Sawilowsky, 2003, p. 128) intervals, the Hodge-Lehmann treatment alternative is modeled by a systematic progression from pseudo-θ, although no expertise is called on to determine negligible or practical meaningfulness.

Regarding near-nill null hypotheses within the context of hypothesis testing, I've opined (Knapp & Sawilowsky, 2001),

This remedy's attendant difficulties are obvious considering the chaos that would arise from the infinite number of near-nils that might be chosen. (Eventually, we speculate, some common near-nils would emerge and evolve into a universally accepted traditional near-nil, completing the circle.) Moreover, the near-nil weakens the Fisherian logic regarding the null hypothesis, which is indirect proof by contradiction. If the probability associated with sample data obtained from a designed study is so remote, the null hypothesis or the model that generated it is contradicted. Rejecting a null hypothesis should be more compelling than rejecting an arbitrarily chosen near-nil hypothesis. Also, in the social and behavioral sciences for cases in which treatment effects or naturally occurring differences are often tiny, using the near-nil hypothesis when investigating interventions with potentially subtle differences may hide a treatment effect. Similarly, as the magnitude of the near-nil increases, the sample size necessary to detect a false near-nil null hypothesis increases in the treatment versus control group and related designs, which would be highly undesirable. (p. 73).