Journal of Data Science,17(2). P. 383 - 390,2019
DOI:10.6339/JDS.201904_17(2).0008

WHY NOT AN INTERVAL NULL HYPOTHESIS?

Michael P. Cohen*

Survey and Data Sciences, American Institutes for Research, Washington DC 20007 U.S.A.

Abstract

Although hypothesis testing has been misused and abused, we argue that it remains an important method of inference. Requiring preregistration of the details of the inferences planned for a study is a major step to preventing abuse. But when doing hypothesis testing, in practice the null hypothesis is almost always taken to be a "point null", that is, a hypothesis that a parameter is equal to a constant. One reason for this is that it makes the required computations easier, but with modern computer power this is no longer a compelling justification. In this note we explore the interval null hypothesis that the parameter lies in a fixed interval. We consider a specific example in detail.

Keywords: Alpha, interval null hypothesis, point null hypothesis, power, type I error.

* Corresponding author:
Email:mpcohen@juno.com, mcohen@air.org

1. Introduction

The main focus of this article is on the choice of the null hypothesis in significance testing. Because of controversy about the need for hypothesis testing as a method of inference, we begin with a brief defense of its use. We then discuss the choice of an interval null hypothesis versus a point null hypothesis, relying largely on a simple example. We end with some further discussion.

2. Do We Need Significance Testing at All?

2.1 Progress in Science Relies in Part on Testing Hypotheses

"Progress in science relies in part on generating hypotheses with existing observations and testing hypotheses with new observations." (Nosek, et al., 2018) Other statistical techniques, including confidence intervals and graphical displays of data, are important supplements but not replacements for significance testing. For a very readable and thorough account of how science progresses, Mayo (1996) is recommended.

2.2 Preregistration of Analysis Plans

Despite the value of significance testing, there is much cotroversy surrounding its use (Wasserstein and Lazar, 2016). In particular, the method has been abused by researchers altering their analyses or changing what they choose to publish in response to the new data being analyzed. Doing so violates the principles on which significance tests are based. A step forward in treating this abuse is preregistration in which the researcher specifies the analysis plan in advance of data collection. The recent article of Nosek et al. (2018) thoroughly addresses preregistration. They note that: "The World Health Organization maintains a list of registries by nation or region (www.who.int/ictrp/network/primary/en/), such as the largest existing registry, https://clinicaltrials.gov/." (p. 2605) They mention other registries as well. We take this opportunity to mention a new registry planned to be active in late 2018 and not covered by Nosek et al. (2018). This is the Society for Research on Educational Effectiveness (SREE) Registry of Efficacy and Effectiveness Studies (REES) https://www.sree.org/pages/registry.php dedictaed to causal inference studies in education and related areas of social science.

Nosek et al. (2018) also discusses analyses not fitting the standard pattern, such as the case of a researcher wanting to do a fresh analysis of an existing dataset.

Registries can be of benefit to a researcher doing a meta-analysis, that is, a study that combines all studies on a particular topic into one all-encompassing analysis. A correct meta-analysis needs to incorporate negative as well as positive results. The diligent meta-analyst can search registries for studies that

were

planned but not published and contact the proposed data analyst to find out what happened.

It should be emphasized that preregistration does not preclude discussing in research reports unanticipated findings in the data or findings that do not quite reach statistical significance. Such findings can be mentioned in an exploratory fashion as deserving further research.

2.3 Significance Testing When Used as Intended

Many have criticized significance testing even when used as intended. It is useful to first state what significance testing is supposed to do. Mayo and Cox (2006, p. 81) write:

The immediate objective is to test the conformity of the particular data under analysis with H0 in some respect to be specied. To do this we find a function $t = t(y)$ of the data, to be called the test statistic, such that

- the larger the value of t the more inconsistent are the data with H0;
- the corresponding random variable T = t(Y) has a (numerically) known probability distribution when H0 is true.

The probability that $T \geq t$ given that the null hypothesis is true becomes the criterion on which the conformity is judged.

A common confusion is to think that significance testing is designed to test the probability that the null hypothesis is true, but that is not its purpose. Here it differs from a Bayesian hypothesis test, which does measure the probability that the hull hypothesis is true, assuming a specific prior distribution. It is therefore not correct to consider a Bayesian hypothesis test as a substitute for a significance test, or vice versa.

Like all statistical procedures (even nonparametric ones), significance testing depends on underlying assumptions. The data may be assumed, for example, to be independent and identically distributed, and perhaps normally distributed. If these assumptions fail, significance testing can give erroneous results.

3. Point Null and Interval Null Hypotheses

If it is accepted that significance testing is worthwhile, there remains the choice of the null hypothesis. Most commonly, a point null hypothesis is used. By a point null hypothesis we mean one of the form $H_0 : \theta = c$ where c is a constant. In most situations that arise in practice, if the sample size is large

enough, the point null hypothesis will be rejected. Practitioners will often say the hypothesis was rejected because the sample size was "too large." But this is an anathema to a statistician where the guiding principle is the more data the better if the data are properly used. The problem arises because of the form of the null hypothesis. We do not usually care if the unknown parameter θ is exactly equal to c provided that it is close. It therefore makes sense to consider the null hypothesis that the parameter lies in a small interval around c.

We are, of course, far from the first to express concern about point null hypotheses. Berkson (1938, 1942) wrote on this extensively. Hodges and Lehmann (1954) studied in detail some specific problems involving non-point null hypotheses. Serlin and Lapsley (1985) supported the use of non-point null hypotheses with an emphasis on applications in psychology and other "soft" sciences. Anderson, Burnham, and Thompson (2000) investigated an information theoretic alternative to point null hypothesis testing. Tryon (2001) wrote: "Null hypothesis statistical testing (NHST) has been debated extensively but always successfully defended." He advocated using "inferential" confidence intervals to test hypotheses in a way that ameliorates their misuse. Very recently, Rao and Lovric (2016) and Zumbo and Kroc (2016) addressed point null hypothesis testing. This is by no means a complete list of studies treating point null statistical hypotheses.

4. An Example of the Problem

We illustrate the problem with point null statistical hypotheses with a specific example. Suppose an expert has asserted that the average salary θ in a particular occupation is \$68, 000 a year. To check this, a simple random sample of size n is drawn. We assume the response rate is 100% and the data are exactly normally distributed with a known standard deviation of \$4, 000. (These assumptions are unrealistic, but they simplify the presentation without affecting the basic point we are making.) We test $H_0 : \theta = 68,000$ versus $H_A : \theta \neq 68,000$. Suppose the true value of θ is 68, 100. Table 1 shows the probability of rejecting the null hypothesis H_0 as a function of the sample size n when the Type I error α is set to .05.

We see that as the sample size increases, the probability of rejecting H_0 increases, eventually becoming almost 1. In one sense, this is as it should be, in that $\theta \neq 68,000$. But it is very possible that the expert meant that $67,500 \leq \theta \leq 68,500$ since annual salaries are often rounded to the nearest thousand. So why not make the null hypothesis $H_0^* : 67,500 \leq \theta \leq 68,500$?

Table 1: Probability of rejecting point-null $H_0 : \theta = 68,000$ when $\theta = 68,100$ for sample size n.

n	Probability of rejecting
10	.051
50	.054
100	.057
500	.086
1,000	.124
5,000	.424
10,000	.705
50,000	1.000

NOTE: Probabilities are rounded to three decimal places.

5. The Example Continued with an Interval Null Hypothesis

Let's now consider the interval null hypothesis $H_0^* : 67,500 \leq \theta \leq 68,500$. Letting I be the interval [67, 500, 68, 500], we can write this as $H_0^* : \theta \in I$. What is the type I error; that is, the probability of rejecting H_0^* if $\theta \in I$? Clearly if θ is near the midpoint of the interval, the probability of rejecting H_0^* is less than if it were at or near one of the endpoints. Let $\alpha(\theta)$ be the probability of rejecting H_0^* for $\theta \in I$. Let α_{MAX} be the maximum value of $\alpha(\theta)$, $\theta \in I$. To be conservative, we shall seek a rejection region such that $\alpha_{MAX} = .05$. The choice of .05 is conventional in many fields but, of course, other values could be used.

Table 2: Probability of rejecting interval-null $H_0^* : \theta \in I$ when $\theta = 68,100$ for sample size n.

n	Probability of rejecting
10	.036
50	.013
100	.005
500	.000
1,000	.000
5,000	.000
10,000	.000
50,000	.000

NOTE: Probabilities are rounded to three decimal places.

In Table 2, we display the probability of rejecting H_0^* when $\theta = 68,000$ and $\alpha_{MAX} = .05$. In problems where 68,000 and 68,100 are "practically equal," the behavior of H_0^* in Table 2 is preferable to the behavior of H_0 in Table 1 in terms of the probability of rejection.

6. Power

If the true value of θ is such that the null hypothesis does not hold then the power $P(\theta)$ is the probability of rejecting the null hypothesis. In Table 3 we compare the power of H_0 and H_0^* for various values of θ and sample size n. Because of symmetry about 68, 000 in the example, $P\,(66,500) = P\,(69,500), P\,(66,000) = P\,(70,000)$, etc., so we only display power for θ greater than 68, 000.

The interval null hypothesis does have somewhat less power than the point null hypothesis.

7. Discussion

The purpose here is to encourage the use of null hypotheses that accurately reflect what one seeks to reject or not, statistically, depending on the data. We are not addressing the issue of subject-matter significance that is typically handled by effect sizes. Judging effect sizes is a vitally important part of significance testing requiring sophisticated subject-matter knowledge, and we prefer to keep it as a separate step. It is worth noting, however, that there is some interesting recent work (Blume, 2017) seeking to combine the determination of statistical and subject-matter significance.

Table 3: Power $P_P\,(\theta)$ for point-null $H_0: \theta = 68,000$ and power $P_I\,(\theta)$ for interval- null $H_0^*: \theta \in I$ for three values of θ and sample size n.

n	$P_P\,(69,500)$	$P_I\,(69,500)$	$P_P\,(70,000)$	$P_I\,(70,000)$	$P_P\,(70,500)$	$P_I\,(70,500)$
10	.220	.180	.353	.301	.507	.449
50	.755	.548	.942	.842	.993	.970
100	.963	.804	.999	.982	1.000	1.000
500	1.000	1.000	1.000	1.000	1.000	1.000
1,000	1.000	1.000	1.000	1.000	1.000	1.000
5,000	1.000	1.000	1.000	1.000	1.000	1.000
10,000	1.000	1.000	1.000	1.000	1.000	1.000
50,000	1.000	1.000	1.000	1.000	1.000	1.000

NOTE: Power is rounded to three decimal places.

Another approach to dealing with a point null hypothesis and a large sample size is to let the Type I error level α decline as the sample size increases. This is a very artificial way of treating the problem of having to reject the point null hypothesis when the true θ is very close to the point null hypothesis value and evades acknowledging that the point null hypothesis is, in fact, false.

The computations involved with an interval null hypothesis are typically more difficult than those for a point null hypothesis. We were able to do the

calculations directly in the example presented here, but this may not be possible in other problems. As Rao and Lovric (2016) noted, with modern computing power these problems are tractable, by simulation if necessary.

This article has been written as if a single hypotheis were being tested, but it is more typical that multiple hypotheses are tested from the same experiment or obsevational study. Adjustments to control the familywise error rate (e.g., Tukey, 1949, or Dunnett, 1955) or the false discovery rate (Benjamini and Hochberg,1995) are needed. These adjustments are independent of the choice of the null hypotheses.

The parameter θ has been one dimensional in our treatment but it could be a vector as well. In that case, the "interval" would be a multidimensional box or ellipsoid whose size and shape were prespecified.

References

[1] Anderson, D. R., Burnham, K. P., and Thompson, W. L. (2000), "Null Hypothesis Testing: Problems, Prevalence, and an Alternative," *Journal of Wildlife Management*, **64**, 4, 912–923.

[2] Benjamini, Y., and Hochberg, Y. (1995), "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society*, Series B, **57**, 1, 289–300.

[3] Berkson, J. (1938), "Some Difficulties of Interpretation Encountered in the Application of the Chi-Square Test," *Journal of the American Statistical Association*, **33**, 526–542.

[4] Berkson, J. (1942), "Tests of Significance Considered as Evidence," *Journal of the American Statistical Association*, **37**, 325–335.

[5] Blume, J. D. (2017), "Evidential Metrics and Second-Generation p-values," Presentation at the Symposium on Statistical Inference, Alexandria, VA: *American Statistical Association*. Available at *https://ww2.amstat.org/meetings/ssi/2017/onlineprogram/Program.cfm.*

[6] Dunnett, C. W. (1955), "A Multiple Comparison Procedure for Comparing Several Treatments with a Control," *Journal of the American Statistical Association*, **50**, 1096–1121.

[7] Hodges, J. L., and Lehmann, E. L. (1954), "Testing the Approximate Validity of Statistical Hypotheses," *Journal of the Royal Statistical Society*, Series B, **16**, 2, 261–268.

[8] Mayo, D. G. (1996), Error and the Growth of Experimental Knowledge, Chicago: University of Chicago Press.

[9] Mayo, D. G., and Cox, D. R. (2006), "Frequentist Statistics as a Theory of Inductive Inference," in Optimality: The Second Erich L. Lehmann Symposium, ed. J. Rojo, Lecture Notes–Monograph Series, Institute of Mathematical Statistics, **49**, 77–97.

[10] Nosek, B. A., Ebersole, C. R., DeHaven, A. C., and Mellor, D. T. (2018), "The Preregistration Revolution," *Proceedings of the National Academy of Sciences*, **115**, 11, 2600–2606. Available at *http://www.pnas.org/content/pnas/115/11/2600.full.pdf.*

[11] Rao, C. R., and 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.

[12] Serlin, R. C., and Lapsley, D. K. (1985), "Rationality in Psychological Research: The Good-Enough Principle," *American Psychologist*, **40**, 1, 73–83.

[13] Tryon, W. W. (2001), "Evaluating Statistical Difference, Equivalence, and Inde- terminacy Using Inferential Confidence Intervals: An Integrated Alterna- tive Method of Conducting Null Hypothesis Statistical Tests," *Psychological Methods*, **6**, 4, 371–386.

[14] Tukey, J. W. (1949), "Comparing Individual Means in the Anaysis of Variance,"*Biometrics*, **5**, 99–114.

[15] Wasserstein, R. L., and Lazar, N. A. (2016), "The ASA's Statement on p-Values: Context, Process, and Purpose," *The American Statistician*, **70**, 2, 129–133.

[16] Zumbo, B. D., and Kroc, E. (2016), "Some Remarks on Rao and Lovric's 'Testing Point Null Hypothesis of a Normal Mean and the Truth: 21st Century Perspective,' " *Journal of Modern Applied Statistical Methods*, **15**, 2, 33–40.