American Review of Mathematics and Statistics
June 2015, Vol. 3, No. 1, pp. 22-26
ISSN: 2374-2348 (Print), 2374-2356 (Online)

Published by American Research Institute for Policy Development
DOI: 10.15640/arms.v3n1a3
URL: http://dx.doi.org/10.15640/arms.v3n1a3

Hypothesis Testing and Statistical Confidence: An Overdue Observation on the Efficacy of a Hypothesis Test

David D. Marshall[1], Brandi N. Falley[2] & Mark S. Hamner[2]

Abstract

We introduce the novel argument that the general concept of statistical confidence applies both to an interval estimate of a parameter and to a hypothesis test. Measured degrees of statistical confidence are mathematical probabilities of accurate parameter identification established prior to drawing samples. Such probabilities serve as the foundation for a statistician's expectation and conviction that a hypothesis test will correctly identify a true hypothesis, and more familiarly, that an interval estimate will properly identify a population parameter. The incidental and potentially misleading role of the P-value is discussed in the context of statistical confidence.

Keywords: hypothesis testing, power analysis, statistical confidence, P-values

1. Introduction

The term "statistical confidence" has been defined clearly with regard to confidence interval estimates of population parameters. Confidence is the mathematical probability, established prior to drawing samples, that an interval will capture a parameter. Sources of such probabilistic statistical confidence also provide the discernible odds that a hypothesis test will identify the true hypothesis, and this is a subtle and important distinction that seems to have been hiding in plain sight; we have yet to see this suggestion in the literatures of statistical education and practice. Our purpose here is to recommend the extension of "statistical confidence" to define the efficacy of both interval estimates and hypothesis tests, and thereby identify a fundamental basis for appreciating the usefulness of hypothesis tests, and for trusting a hypothesis decision once made. So long as hypothesis testing methods continue to be taught and used, students and practitioners of inference should learn when to trust the odds that an experiment will illuminate the truth, and to appreciate the formulation of research designs that embrace high statistical confidence as herein defined, those with high probabilities that true hypotheses will not be rejected.

2. Statistical Confidence

2.1 Confidence and Ordinary Interval Estimation

Rossman and Chance (2012) describe inference processes in a way that highlights concern for the magnitudes of population parameters: "The concept of *statistical confidence* relates to how close you expect a sample statistic to come to its corresponding population value… The concept of *statistical significance* concerns how unlikely an observed sample statistic is to have occurred, assuming some conjectured value for the population parameter" (p. 320). All of inference concerns population parameters and their correct identification, whether done with an interval estimate or a hypothesis test. As we see reviewed below, confidence is the probability that a parameter will be correctly identified, despite the fact that the result of any inference process is never known to be correct. Regardless of being able to know the truth, we can trust that the methods will provide correct results, when the odds of success are strong.

[1]The Texas Woman's University, Department of Mathematics and Computer Science, Denton, Texas 76204.
Email: dmarshall@twu.edu
[2]The Texas Woman's University, Department of Mathematics and Computer Science, Denton, Texas 76204.

Carlin and Louis (2009), speaking of a conventional 95 percent interval, asserted that "…*before* any data are collected, the probability that the interval contains the [parameter] value is .95" (p. 7), and this statement captures the essence of confidence and confidence intervals. Such definitions of confidence, as a pre-sampling probability of accurate or successful parameter estimation, abound in the literature. For example, Moore, Notz & Fligner (2015) offered that "A confidence level C [such as .95]… gives the probability that the interval will capture the true parameter value in repeated samples. That is, the confidence level is the success rate for the method" (p. 353). Much earlier, Glass and Stanley (1970) stated, "When an interval estimate of a parameter is constructed so that is has a certain known probability of including the value of the parameter between its limits, the interval is called a confidence interval… the confidence coefficient is the probability that a randomly selected interval… will capture the parameter… It is understood that the probability statement refers to the sample space of all intervals that could be formed by computing one interval for each sample" (p. 260-61).Similarly, Gould and Ryan (2013) noted that "The confidence level tells us how often the estimation method is successful. Our method is to take a random sample and calculate the confidence interval… the confidence level measures the success rate of the *method*, not of any one particular interval" (p. 330). Lastly, Rossman et al. (2012) concur: "Thus the probability statement applies to what value an interval will take prior to the sample being collected (i.e., to the method), not whether or not a particular interval contains the fixed parameter value once it has been calculated. If you did have all the intervals from all possible samples… the probability that you will randomly select an interval that contains [the parameter] is .95" (p. 342). Thus far we see confidence defined as a pre-sampling probability of success in accurately capturing a parameter value within the limits of a confidence interval. The method is viewed as becoming ever more reliable or accurate as the pre-sampling probability value increases, that is, as the pre-defined, pre-sampling confidence percentage is increased.

2.2 Confidence and Hypothesis Testing

Gene V. Glass and Julian C. Stanley are venerated names in statistical education and methodology. Their classic statistics text (Glass et al., 1970) includes a statement about hypothesis testing which unmistakably identifies a test as a sampling context that can be made rich in the odds that a correct decision will be made no matter which hypothesis is true: "From any sample it can never be concluded with certainty that *H* is true or false; *the best one can do is to make a decision that has a high probability of being true*" [emphasis added] (p. 281). The decision in question is a choice to reject or not reject the null hypothesis (H_0). For the hypothesis decision to have a high probability of being true or correct there must be a high probability that H_0 will not be rejected when it is true, and a correspondingly high probability that H_0 will be rejected when it is false. We say that such probabilities define statistical confidence for a hypothesis test; the higher each of the probabilities, the more likely that the correct decision about H_0 will be made. Prudent arrangements of sampling experiments are done by selecting samples of sufficient size so that null hypotheses that are false to predetermined degrees of magnitude will be rejected, say, 80 to 90 percent of the time or more often in practice. Such experiments that couple high probabilities that true H_0 will not be rejected with high probabilities that false H_0 will be rejected are said to harness high statistical confidence as their foundations.

Probabilities established prior to the drawing of samples govern the rate of success with interval estimation. By comparison, adequately powered hypothesis tests are based in defining the magnitudes of competing parameters representing both a true H_0 and a true alternative hypothesis (H_A), and establishing desirably high probabilities that either of them will be correctly detected, that is, that the false hypothesis will be rejected. Such chances or probabilities of correct hypothesis support can be made, say, 95 percent or higher in the ideal. Following standards set by Cohen's (1969, 1977, 1988, 1990, 1992) frequentist-oriented framework for hypothesis testing based in statistical power analysis, modern practice encourages the researcher to define H_A in terms of an "effect size" or speculation about differences and relationships of varying magnitudes in contrast to the no difference or no relationship H_0. Sample sizes are chosen which provide for high probabilities that H_0 will be rejected, but only if the alternative H_A is the true hypothesis. Establishing an effect size as a hypothesized difference between H_0 and H_A amounts to expressing expectations about differences and relationships that are predicted from theory, or are of substantive clinical or practical relevance (Cohen, 1969, 1990; Kirk 1996; Moore et al., 2015; Utts & Heckard, 2014), as compared with an expectation defined in H_0.

Establishment of a standardized effect size in H_A that competes with the value of a parameter in H_0 clearly concerns parameter magnitudes, as we model all possible sampling outcomes in either of two adjacent and overlapping populations with parameters "known" by assertion. Particular probabilities that each parameter will be correctly identified, that is, that a particular true hypothesis will not be rejected, are established prior to the taking of samples, and these probabilities constitute the comprehensive statistical confidence basis, and potential effectiveness of the hypothesis test. The odds are 19 to one that a 95 percent confidence interval will capture a parameter, and the odds are 19 to one that the true parameter will be identified by hypothesis test (that the false hypothesis will be rejected) when $\alpha = .05$ and statistical power equals 95 percent. Such is statistical confidence as here conceived. High values of statistical confidence provide for correct decisions more often than not, because they define sampling contexts that only rarely produce extreme instances of data which are misleading to the statistician, at least "on paper" as we plan a sampling experiment. As a consequence, we are free to subjectively trust or believe in an inference method that is designed to produce truthful outcomes more often than not, to trust the odds that attend such processes, and to trust our hypothesis decisions once made if, as Glass and Stanley contend, the endpoint of a hypothesis test will be making a decision that will have had a high probability of being true or correct before any data were drawn.

2.3 Illustrating Confidence for a Hypothesis Test

Table 1 provides a demonstration of high statistical confidence for a common hypothesis test concerning a population correlation when an effect of "medium" magnitude is to be detected if it exists, in contrast to H_0. Sufficiently large samples maximize the likelihood that a false H_0 will be rejected, thus that the true H_A will not be rejected. With adequate sample size, thecomputed power to reject a false H_0 in favor of a true H_A is made equal to the probability that true H_0 will be supported. The researcher can expect that the test will result in detection of the hypothesis that is true; he or she can be subjectively confident in the test results. Here the odds of correct parameter detection are 19 to one regardless of whether H_0 or H_A is true.

Table 1: Confidence Basis for a Powerful Sampling Experiment, H_0: $\rho = 0.00$, H_A: $\rho = 0.30$, n = 115, Medium Effect

	H_0 is true	H_A is true
Reject	$\alpha = .05$	$\beta = .05$
Do not Reject	1.0 - α = .95	1.0 - $\beta = .95$

Note: To reject a true H_A is to not reject a false H_0.

Employing samples of inadequate size results in unacceptably large chances that false H_0 will not be rejected, that is, that true H_A will be rejected unwittingly, as seen below in Table 2 and Figure 1. If sample data lead the researcher to support or not reject H_0 in this situation, personal faith in the decision will be minimized because

Table 2: Compromised confidence for a sampling experiment, H_0: $\rho = 0.00$, H_A: $\rho = 0.30$, n = 15, medium effect

	H_0 is true	H_A is true
Reject	$\alpha = .05$	$\beta = .70$
Do not Reject	1.0 - α = .95	1.0 - $\beta = .30$

Note: H_0 is not likely to be rejected regardless of its truth status

Of the sheer chance that H_0 will not be rejected no matter whether it is true or false. In this case, the mechanism of statistical confidence is high, and the correct hypothesis is not likely to be rejected, if and only if H_0 is true. Odds of correct detection of H_A are not more than about four in 10. Figure 1 presents output adapted from the popular G*powerprogram (Erdfelder, Faul, & Lang, 2009), and gives graphic expression to such compromised confidence. Seventy percent of the samples when $H_A$is true fall within the region of the H_0 model that mandates that H_0 not be rejected.

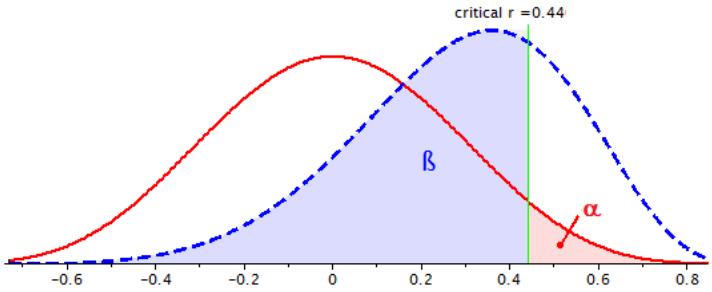

Figure 1: H_0: ρ = 0.00, H_A: ρ = 0.30, n = 15, α = .05, β = .70

In this scenario, P-values will tend to be very large, larger than a conventionally stated α-level by far, regardless of whether H_0 is true, rendering the size of the P-value useless as an indicator of weak, moderate, or strong evidence for H_0. When power is not known and large P-values are to be taken as evidence in support of H_0, the user will be uncertain about her or his decision, which can often be incorrect when the β probability overwhelms a sampling process and H_A is the true hypothesis. Incomplete definition of decision probabilities leaves the user in doubt about retaining (not rejecting) H_0.

3. P-Values and Statistical Confidence

3.1 Using P-values in Making Hypothesis Decisions

Take caution when data mandate that you not reject a null hypothesis, while at the same time power is known to be minimal, or when power is unknown. However, it is reasonable to feel confident in a decision to support H_A (to doubt and reject H_0) based solely on a small P-value, regardless of whether power has been defined and statistical confidence is known to be high overall. We expect to support a true H_0 given 19 to one odds in its favor when α = .05, and the finding of a small P-value taken as evidence to the contrary is "surprising," as Starnes, Yates, & Moore (2011) relate: "A P-value tells us how surprising the observed outcome is. Very surprising outcomes (small P-values) are good evidence that the null hypothesis is not true" (p. 465). Mathematically speaking, P-values are not "evidence" for or against any hypothesis, yet suitably small P-values herald data, modeled as sampling errors when H_0 is true, that sit at great distances away from the H_0 parameter, making it very difficult to believe in H_0; hypothesis tests modify our beliefs without providing proofs. We do caution that large P-values cannot be taken at face value as good evidence in favor of H_0 unless one also knows the chance that H_0 will not be rejected when it is false in comparison to an alternate true effect of relevant magnitude. Hypothesis testing aims to impact beliefs; and whether measured statistical confidence is known can occasionally impact the strength of our convictions.

3.2 P-values are not Substitutes for Statistical Confidence

Note well that the vaunted P-value is not in the probability set that defines Glass and Stanley's ideal of a "high probability" for decision truth, which includes, and only includes 1.0 - α and 1.0 – β. Effective hypothesis testing relies on established pre-sampling statistical confidence, no component of which is defined by the post-sampling P-value. We stress that the "high probability" in Glass and Stanley's formulation is not found in the P-value nor in its complement, 1.0 - P. Well-known and debunked P-value fallacies such as the Odds-Against-Chance and Reliability/Replicability fantasies (Carver, 1978; Kirk, 1996), were once pressed into service to provide an illusion of support for a decision to reject or not reject H_0, but the only legitimate and comprehensive source of the odds that a true hypothesis will not be rejected is the full and generous definition of statistical confidence as identified in this paper. We deem that the terminal point of a hypothesis test is a statement of belief about differences or relationships given circumstantial evidence, without benefit of incontrovertible proof. Our trust in the positions we take about H_0 must be based on the generalized power of our study designs that enable us to believe that we will "get it right" in the end, even though we can never know the truth of any hypothesis. Establishment of comfortable levels of confidence prior to sampling, where a correct decision on a true H_0 is almost always attended by 1.0 – α = .95, and probability that a true H_A will be supported (that a false H_0 will be rejected) is at least, say, 1.0 – β = .80 in practice, provides both mathematical and subjective expectations upon which to base trust in the efficacy of our research designs and in the value of eventual test results.

3.3 Post Script

Our presentation could be taken as no more than a reminder that powered hypothesis tests provide a theoretical edge favoring the detection of a true hypothesis. Our deeper message is that we base our trust in the accuracy of confidence intervals by virtue of mathematically defined statistical confidence as shown herein, and we can similarly trust in the accuracy of hypothesis tests and our eventual hypothesis decisions for the very same reason, and as a more reliable and telling basis than using the post-sampling P-value as our sole index of certitude. We suggest adding this concept to hypothesis testing instruction and education at all levels.

4. Summary and Conclusions

The classic ideals of Jacob Cohen, and of Glass and Stanley, reflect the more technically arcane early work of Jerzy Neyman and Egon Pearson (Neyman and Pearson 1928a, 1928b) in portraying a hypothesis test as an opportunity to probe for population values in a manner that will lead to a correct hypothesis decision most of the time, at least according to the mathematical models used for hypothetical populations and hypothesis tests. Hypothesis tests can be supported by high mathematical probabilities, established prior to the taking of samples, that true hypotheses will not be rejected. Such attention to power and establishment of ideal statistical confidence for a hypothesis test requires specification of a meaningful parameter value for H_A, which places science directly into the hypothesis test, with H_A stated so as to reflect the researcher's original intent for conducting a study, which is often to detect differences and relationships of practical or theoretical import. Otherwise, the research goal may be to actually support a null condition (to not reject H_0) as opposed to an alternative condition in a study that has been well-funded with power to reject H_0 if it is false. One may predict and hope that a null hypothesis will not be rejected, investing personal trust or faith in the decision not to reject H_0, and this can only be accomplished when H_0 is not likely to be supported when false. Modern emphasis on power analysis with consequent generalized statistical confidence of calculable levels provides students and practitioners the opportunity to consider worthy effects to detect, and to vary sample sizes in setting the occasion for data to illuminate the hypothesis that is true, and to trust the odds that their hypothesis decisions will be correct.

5. References

Carlin, B. P., & Louis, T. A. (2009), Bayesian methods for data analysis (3rd Ed.), Boca Raton, FL: Chapman & Hall.

Carver, R. P. (1978), The case against significance testing, Harvard Educational Review, 48(3): 378-399.

Cohen, J. (1969), Statistical power analysis for the behavioral sciences, New York, NY: Academic Press.

Cohen, J. (1977), Statistical power analysis for the behavioral sciences (Rev. ed.), New York, NY: Academic Press.

Cohen, J. (1988), Statistical power analysis for the behavioral sciences (2nd Ed.), Hillsdale, N.J.: L. Erlbaum Associates.

Cohen, J. (1990), Things I have learned (so far), American Psychologist, 45(12), 1304-1312.

Cohen, J. (1992), A power primer, Psychological Bulletin, 112(1), 155-159.

Erdfelder, E., Faul, F., and Lang, A. (2009), G*Power (Version 3.1.2)[Computer program], http://www.psycho.uniduesseldorf.de/aap/projects/gpower/

Glass, G. V., and Stanley, J. C. (1970), Statistical Methods in Education and Psychology, Englewood Cliffs, N. J.: Prentice-Hall, Inc.

Gould, R. and Ryan, C. (2013), Introductory Statistics, exploring the world through data, Boston, MA: Pearson.

Kirk, R. E. (1996), Practical significance: A concept whose time has come, Educational and Psychological Measurement, 56(5): 746-759.

Moore, D. S., Notz, W. I., and Fligner, M. A. (2015), Statistics In Practice, New York, NY: W. H. Freeman and Co.

Neyman, J. and Pearson, E. (1928a), On the use and interpretation of certain test criteria for purposes of statistical inference: Part I, Biometrika, 20A, 175-240.

Neyman, J. and Pearson, E. (1928b), On the use and interpretation of certain test criteria for purposes of statistical inference: Part II, Biometrika, 20A, 263-294.

Rossman, A. J., and Chance, B L. (2012), Workshop Statistics: Discovery with data (4th Ed.), Hoboken, NJ: John Wiley and Sons.

Starnes, D. S., Yates, D., and Moore, D. S. (2011), Statistics through Applications, New York, NY: W. H. Freeman and Company.

Utts, J. M., and Heckard, R. F. (2014), Mind on Statistics (5th Ed.), Stamford, CT: Cengage Learning