Some Aspects of Statistical Significance in Statistics Education

Pranesh Kumar
University of Northern British Columbia, Prince George, CANADA
pranesh.kumar@unbc.ca

Abstract

Statistical significance in the null hypothesis testing is the primary objective method for representing scientific data as evidence and for measuring strength of that evidence. Statistical significance is measured by calculating the probability value (P-value) generated by the null hypothesis test of significance. Several interpretations of P-values are possible. For example, P-value is interpreted as the probability that the results were obtained due to chance. A small P-value would recommend that the null hypothesis is not supported by the sample data and the research hypothesis is strongly favored by data. Alternatively, effect size can be considered as a measure of the extent to which the research hypothesis is true or to the degree to which the findings have practical significance in context of the study population. Effect size measures seem to have advantages over statistical significance because they are not affected by the sample size and are scale-free. The effect size measures can be uniquely interpreted in different studies regardless of the sample size and the original scales of the variables. In this paper we will present some aspects of statistical significance, practical significance and their computations. We will consider statistical significance measures for some commonly used statistical parameters. In conclusion, we present discussions and remarks.

Key Words: Statistical evidence, significance test, practical significance, effect size.

1. Introduction

Statistics primarily as a decision making science has made fundamental contributions to the scientific studies by providing objective and quantitatively measurable alternatives to the personal judgment for interpreting the evidence produced by experimental and observational data. Statistical inference deals with drawing conclusions about a population from the evidence provided by observations collected through a randomly selected sample. Because the sample is not the entire population, statistical conclusions are uncertain and, therefore, statistical inference must state conclusions and provide a measure of how uncertain they are. The uncertainty is measured using the rules of probability theory. Statistical null hypothesis testing and the probability-value generated by this procedure are widely used to answer a key question: Does the given set of sample data provide enough evidence to support one statistical hypothesis over another hypothesis? When is it correct to infer that sample observations are evidence in favor of one hypothesis *vis-à-vis* another? Standard statistical methods for testing of hypotheses are widely used; however, all is not well. Statistical hypothesis testing has received enormous criticism from researchers in various disciplines (Clark 1963, Bakan 1966, Deming 1975, Carver 1978, Guttman 1985, Cohen 1994, Barnard 1998, Johnson 1999, Moore and Notz 2009). There are often concerns that these methods frequently lead to misinterpretation of results of scientific investigations. This paper briefly describes what statistical null hypothesis testing is and how these tests are often viewed. We discuss shortcomings of hypotheses testing and some common misinterpretation of probability-values. Then some alternatives like effect size to indicate practical significance are described. Further, we summarize statistical significance measures for some commonly used statistical parameters. Finally, paper concludes with some discussions and remarks.

2. Reasoning of Tests of Significance

Statistical inference draws conclusions about a population from a given set of sample observations. Statistical test attempt to answer the question: Do the sample data provide good evidence against a claim? The conclusions from a statistical test are: “If we took many samples and the claim were true, we would rarely obtain a result like this.” Probability-value (P-value) is used as a numerical measure of how strong the sample evidence is.

In statistical null hypothesis testing, first a null hypothesis (H_0) about some parameter or phenomena is set up. The claim being tested is what null hypothesis is. Usually this is a statement of “no effect” or “no difference”. Alternatively, null hypothesis is generally the opposite of the research objective which the researcher believes true and wants to seek support for. Next relevant data are collected typically by an experiment or by random sampling. A statistical test is designed and conducted to assess the strength of the evidence against the null hypothesis. This test generates a P-value which is the probability, computed assuming that null hypothesis is true, that the sample outcome would be as extreme or more extreme than one actually observed. Finally what the P-value means relative to null hypothesis is considered and a decision to reject or fail to reject null hypothesis H_0 is made.

In advance, we decide how much evidence against null hypothesis H_0 we will accept. That means, how small a P-value we require. The decisive value of P is known as the significance level denoted by α. If the P-value is as small as or smaller than α, we conclude that data are statistically significant at level α. If α =0.05, we are requiring that the data give evidence against H_0 so strong that it would happen no more than 5% of the time (1 time in 20) when H_0 is true. By choosing α =0.01, we are aiming at an evidence so strong that it would appear only 1% of the time (1 time in 100) when null hypothesis H_0 is true. It may be noted that “significant” in statistical sense does not imply “important.” It means: “Not likely to happen just by chance.”

Several interpretations of P-values are possible. Sometimes P-value is viewed as the probability that the results were obtained due to chance. The smaller the P-value is, the stronger is the evidence against null hypothesis provided by the data. A large P-value say for a test of H_0: $\mu = 0$ would indicate that the observed sample mean was due to chance and μ could be assumed to be zero (Schmidt and Hunter 1997). Another interpretation of the probability 1-P is that it signifies the reliability of the results and is the probability of getting the same result if the experiment was repeated. Significant differences are often termed as reliable under this interpretation. The P-value can also be interpreted as the probability that the null hypothesis is true. This interpretation is the most direct one. These interpretations are termed as fantasies about the statistical significance (Carver 1978). None of them is actually true however they are treated as if they were true. Small values of P are considered strong evidence that the null hypothesis is false however researchers have demonstrated that it is not so. In an example (Berger and Selke 1987) for which a P-value was 0.05 for a sample of size $n = 50$, the probability that the null hypothesis was true was 0.52. It may be noted that the disparity between the P-value and the probability of the null hypothesis given the observed data increases as sample size increases. In reality, P-value is the probability of the observed data or data more extreme given that the null hypothesis is true.

What are more extreme data? Each set of more extreme outcomes has its own probability which alongwith the probability of the result actually obtained constitutes P. To determine which data set are more extreme than the observed so that a P-value can be computed requires knowledge of the intentions of the investigator (Berger and Berry 1988). Thus the P-value depends on results that were not obtained and what the intentions of the investigator were. For example (Johnson 1999), suppose a sample consists of 10 males and 3 females and the null hypothesis is about the balanced sex-ratio. What sample would be more extreme? The answer depends upon the sampling plan to collect the data. Suppose that the investigator has decided to sample 13 individuals and to stop as soon as there are 10 males encountered. Then, outcomes more extreme than observed sample of 10 males and 3 females would be 11 males and 2 females, 12 males and 1 female and 13 males and no females. Conversely, if the investigator decides to collect data until 3 females are reached; more extreme outcomes are infinite: (3 females and 11 males), (3 females and 12 males) and so on. Alternatively, the investigator collects data until the difference between the numbers of males and females is 7 or until the difference is significant at some level.

It needs to note that P-value is calculated under the assumption that the null hypothesis is true. Most null hypotheses under significance tests state that some parameter equals zero or some set of parameters are all equal. These hypotheses are almost invariably known to be “false” before any data are collected (Berkson 1938, savage 1957, Johnson 1995). If such types of null hypotheses are not rejected, it is often because the sample size is too small (Nunnally 1960). What remains of interest is whether or not the sample size considered will be sufficient to detect the difference. If the null hypothesis truly is false, P-value can be made as small as one wish by selecting a large enough sample. This strong dependence of P-values on the sample size suggests to standardize P-values to a sample size of 100 by replacing P by $\sqrt{n/10}$, or 0.5, if that is smaller (Good 1982). Another concern is about arbitrarily choosing the significance level α. Using a fixed significance level say $\alpha = 0.05$ makes a meaningless distinction between a significant finding for $P = 0.049$ and a non-significant finding if $P = 0.051$. As a matter of fact such minor differences are illusory as they are based on the tests with assumptions which are only approximately met (Preece 1990).

Prior to concluding this section, a relevant question which requires attention is: “Given so much of controversies about statistical null hypothesis testing, why it is so increasingly used?” Scientific hypothesis testing dates back to 17th century where Francis Bacon in 1620 (Quinn and Dunham 1983) discussed the role of proposing alternative explanations and conducting tests to distinguish between them for scientific understanding. The similar concepts are noted in Popper (1959) and Platt (1964) who emphasize setting alternative hypotheses that lead to different predictions. Results inconsistent with predictions from a hypothesis cast doubt on its validity. Researchers (Lindley referred in Matthews 1997) observed that: “People like conventional hypothesis tests because it’s so easy to get significant results from them.” Carver (1978) noted: “Statistical significance is generally interpreted as having some relation to replication which is the cornerstone of science.” Nester (1996) opined: “They appear to be objective and exact; they are readily available in many statistics software packages; everyone else seems using them and we are taught to use them; some research journals require them for publication.” Johnson (1999) attributed heavy use of significance testing in soft sciences like psychology, sociology, education etc. to “Physics envy”. In physics and hard sciences, a theory is postulated which generates several predictions. These predictions are treated as scientific hypotheses. Then experiments are conducted to falsify each hypothesis. If the experimental outcomes refute the hypothesis, that outcomes imply that the theory is incorrect. If the results do not refute the hypothesis, theory stands and gain support from the experiment.

3. Effect Size in Testing Significance

Statistical significance does not adequately address whether the results in a given study will replicate (Carver 1978). Thompson (1999) emphasizes that we must consider questions like: What the magnitudes of sample effects are? Whether these results will generalize? Statistical significance testing does not respond to either question.

Given two random samples from the same population, there will always be a difference between them. Effect size quantifies the size of the difference between two groups. Effect size emphasizes the size of the difference rather than confounding this effect with sample size (Coe 2002). Therefore, it effect size is a metric for quantifying the effectiveness of a particular intervention relative to some comparison. The statistical significance which is measured by P-value is the probability that a difference of at least the same size would have arisen by chance, even if there really were no difference between two populations. However statistical significance combines the effect size and sample size. The major concern in using statistical significance testing is that the P-value depends essentially on the effect size and the size of the sample. One may infer significant difference either if the actual effects were very large despite having only small samples, or if the samples were very large even if the actual effect sizes were small. However, we cannot ignore the statistical significance of a result since without it we may infer firm conclusions from studies where the samples are too small to justify such confidence.

Effect size is defined as the standardized mean difference between two groups. Assuming that population standard deviation is σ and control and treatment group sample means are M_1 and M_2 respectively, Cohen's (1969) effect size measure $d = \frac{M_1 - M_2}{\sigma}$. Another feature of the effect size is that it can be directly converted into statements about the overlap between the two samples in terms of a comparison of percentiles (Coe 2002). Effect sizes make use of an equivalence between the standardized mean difference (d) and the correlation coefficient (r). Using a dummy variable taking a value 0 for the control group, 1 for the experimental group and correlation r between dummy variable and outcome measure, metric is defined $r^2 = d^2/(4+d^2)$ (Cohen, 1969). Rosenthal and Rubin (1982) suggested a further interpretation, which they call the binomial effect size display. If the outcome measure is reduced to a simple dichotomy (for example, whether a score is above or below a particular value such as the median, which could be thought of as success or failure), r can be interpreted as the difference in the proportions in each category. McGraw and Wong (1992) have suggested a common language effect size measure which is the probability that a score sampled at random from one distribution will be greater than a score sampled from another. Another way to interpret effect size is to compare them to the effect sizes of differences that are familiar. For example, Cohen (1969) describes an effect size of 0.2 as small, an effect size of 0.5 is described as medium and an effect size of 0.8 as grossly perceptible and therefore, large. Cohen however does acknowledge the danger of using terms like small, medium and large out of context. Glass et al. (1981) are particularly critical of this approach, arguing that the effectiveness of a particular intervention can only be interpreted in relation to other interventions that seek to produce the same effect. They also point out that the practical importance of an effect depends entirely on its relative costs and benefits. For example, in education, if it could be shown that making a small and inexpensive change would raise academic achievements by an effect size of even as little as 0.1, then this could be a very significant improvement, particularly if the improvement applied uniformly to all students, and even more so if the effects were cumulative over time.

How do we measure the margin of error in estimating effect sizes? Clearly, an effect size calculated from a very large sample is likely to be more accurate than one calculated from a small sample. The margin of error can be quantified using the confidence interval which provides the same information as is usually contained in a significance test. For example, using a 95% confidence interval is equivalent to choosing a 5% significance level.

4. Concluding Remarks

Despite wide use of statistical significance testing in scientific studies and research journal publications, there have been many articles already published decrying its use. They suggest that statistical hypothesis tests add very little to the research products and this tool is overused, misused and often inappropriate. The debate on use and misuse of significance testing is not recent and there prevails a vast literature in favor and against its use in scientific investigations. In this article, we have re-emphasized the basic questions on statistical hypothesis tests in statistical inference making and addressed some of its important concerns. We agree that some of the questions raise valid concerns and need to be investigated thoroughly. We feel there is still a long way to go to understand how statistical theory along with the modern high computing facilities can be applied to reach sound, satisfactory and practical results from data that have been collected in scientific studies.

References

Bakan, D. (1966) “The test of significance in psychological research,” Psychologcal Bulletin, 66:423-437.

Barnard,G. (1998) “Pooling probabilities,” New Scientist, 157:47.

Berger, J. 0. and Berry, D. A. (1988) “Statistical analysis and illusion of objectivity,” American Scientist, 76: 159-165.

Berger, J. O. and Selke, T. (1987) “Testing a point null hypothesis: the irreconcilability of P values andEvidence, ” Journal of the American Statistical Association, 82:112-122.

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.

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

Clark, C. A. (1963) “Hypothesis testing in relation to statistical methodology,” Review of Educational Research 33: 455-473.

Cohen, J. (1969) Statistical Power Analysis for the Behavioral Sciences, NY: Academic Press.

Coe, R. (2002) “ It's the Effect Size, Stupid: What effect size is and why it is important,” Annual Conference of the British Educational Research Association, University of Exeter, England, 12-14

Cohen, J. (1994) “The earth is round (p < .05),” American Psychologist, 49, 997-100.

Deming, W.E. (1975) “On probability as a basis for action,” American Statistician 29:146-152.

Glass, G.V., McGaw, B. and Smith, M.L. (1981) Meta-Analysis in Social Research, London: Sage.

Good, I.J. (1982) “Standardized tail-area probabilities,” Journal of Statistical Computation and Simulation, 16:65-66.

Guttman, L. (1985) “The illogic of statistical inference for cumulative science,” Applied Stochastic Models and Data Analysis 1:3-10.

Johnson, D.H. (1999) “The insignificance of statistical significance testing,” Journal of Wildlife Management 63(3):763-772.

Matthews, R. (1997) “Faith, hope and statistics,” New Scientist 156:36-39.

McGraw, K.O. and Wong, S.P. (1992) “A Common Language Effect Size Statistic,” Psychological Bulletin, 111, 361-365.

Moore, D.S. and Notz, W. (2009) Statistics: concepts and controversies, W.H. Freeman.

Nester, M.R. (1996) “An applied statistician's creed” Applied Statistics 45:401-410.

Nunnally, C. (1960) “The place of statistics in psychology ,” Educational and Psychological Measurement 20:641-650.

Platt, J.R. (1964) “ Strong inference,” Science 146:347-353.

Popper, K.R. (1959) The Logic of Scientific Discovery. Basic Books, New York, USA.

Preece, D. A. (1990) “R. A. Fisher and experimental design: a review,” Biometrics, 46:925-935.

Quinn, J. F. and Dunham, A.E. (1983) “On hypothesis testing in ecology and evolution,” American Naturalist, 122:602-617.

Rosenthal, R, and Rubin, D.B. (1982) “A simple, general purpose display of magnitude of experimental effect,” Journal of Educational Psychology, 74, 166-169.

Savage, I.R. (1957) “Nonparametric statistics” Journal of the American Statistical Association 52:331-344.

Thompson, B. (1999) “Common methodology mistakes in educational research, revisited, along with a primer on both effect sizes and the bootstrap.” Annual meeting of the American Educational Research Association, Montreal. [http://acs.tamu.edu/~bbt6147/aeraad99.htm]