Hindawi Publishing Corporation
Journal of Probability and Statistics
Volume 2012, Article ID 804691, 6 pages
doi:10.1155/2012/804691

Research Article

Not Significant: What Now?

Gerhard Marinell,[1] Gabriele Steckel-Berger,[1] and Hanno Ulmer[2]

[1] *Department of Statistics, Faculty of Economics and Statistics, University of Innsbruck, Universitätsstraße 15, 6020 Innsbruck, Austria*
[2] *Department of Medical Statistics, Informatics and Health Economics, Innsbruck Medical University, Schöpfstraße 41, 6020 Innsbruck, Austria*

Correspondence should be addressed to Gerhard Marinell, gerhard.marinell@uibk.ac.at

Received 25 July 2012; Accepted 12 October 2012

Academic Editor: Man Lai Tang

In a classic significance test, based on a random sample with size n, a P value will be calculated at size n aiming to reject the null hypothesis. The sample size n, however, can retrospectively be divided into partial samples $(1, 2, 3, \ldots, n-2, n-1, n)$ and a test of significance can be calculated for each partial sample. As a result, several partial samples will provide significant P values whereas others will not show significant P values. In this paper, we propose a significance test that takes into account the additional information from the P values of the $n-1$ partial samples of a random sample. We show that the $n-1$ P values can greatly modify the results of a classic significance test.

1. Introduction

In this day and age testing for significance has become a ritual which, if it leads to a significant result, still opens the doors to many well-known journals in nearly every scientific field. This is the case even though for a long time the application of null hypothesis significance testing has been criticized and even rejected [1]. What will be shown here is that by extending the classic significance test additional information from a random sample can be obtained and a "not significant" result can possibly be made "significant". A misuse of null hypothesis significance testing can however not be prevented with this method [2].

In the significance test as defined by Fisher [3, 4] the probability that a specific sample will occur is calculated based on the validity of the null hypothesis. This probability is usually abbreviated with P and is compared with a conventionally determined level of significance which is normally 5%. When P is equal to or less than this level of significance then the null hypothesis is rejected. If this is not the case then, as defined by Fisher, the null hypothesis cannot be rejected but also not accepted [5]. This procedure is valid for a given

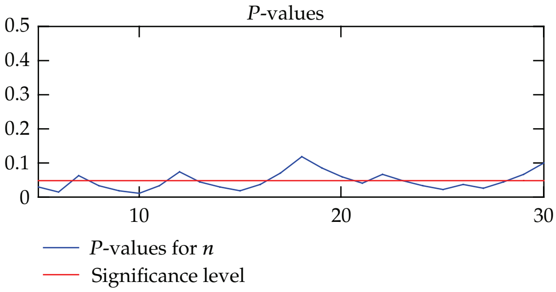

Figure 1: P values of 30 partial samples with $\pi = 0.55$ and $H_0 : \pi = 0.50$. The partial samples 1 to n are given on the abscissa, on the ordinate the corresponding P values. The red line indicates the significance level of 5%.

sample provided it is a random sample. This means that the units of the sample are drawn from the population randomly and the probability with which a unit is drawn out of the population is given. If you presuppose, as is customary, a simple random sample ("idd" assumption = "independent and identically distributed" assumption), then we have the same probability for each unit being part of the sample and the drawings from the population occur independently of each other.

Even though the randomness of the sample is a premise for a test of significance, it is seldom certified. Additionally, the "iid" assumption requires that the order of drawings from the population is known and this allows the split of a random sample into a series of smaller subsamples. The sample size n can thus be retrospectively divided into partial samples $(1, 2, 3, \ldots, n-2, n-1, n)$ and a test of significance can be calculated for each partial sample. As a result, several partial samples will provide significant P values whereas others will not show significant P values and the third category will lead to nearly only significant P values.

2. Illustrative Examples

A series of examples with randomly drawn samples should illustrate the typical situations. In a first example, the null hypothesis $H_0 : \pi = 0.50$ should be verified at a significance level of 5% with the help of a significance test and a random sample size of $n = 30$. If a random sample of this size is created by a random generator for the binomially distributed random variable and the "true" value $\pi = 0.55$, then consequently the corresponding random sample with a P value of 0.100 does not suggest a significant result. However, if the partial samples 1 to $n - 1 = 29$ of this random sample are examined then we get a different result. The P values of this partial sample are depicted in the Figure 1 below. One can see that the P values lie both above and under the level of significance. Yet in the case of $n = 18$, $P = 0.119$ is clearly above the 5% level which is similar to what can be seen at $n = 30$.

In this specific case we know however that the true value is 0.55, hence the null hypothesis does not apply, but the rejection of the null hypothesis based on the given sample is not possible. This would nevertheless be possible if one would, for example, only take the first 28 units into consideration. As a result the P value (0.044) is smaller than the significance level of 5% and the null hypothesis can be rejected. Such a rejection of the null hypothesis, however, requires that the information of the two following units is ignored. A method that simply ignores valid information is not sensible and can hardly be accepted. So in our approach we do not intend to ignore valid information, as we show below. We try to capture

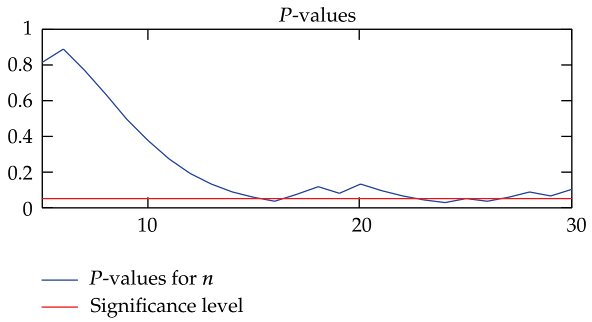

Figure 2: P values of 30 partial samples with $\pi = 0.50$ and $H_0 : \pi = 0.50$. The partial samples 1 to n are given on the abscissa, on the ordinate the corresponding P values. The red line indicates the significance level of 5%.

all the information that comes with a valid drawn sample. The null hypothesis in our example can only be rejected if all P values from the $n - 1$ subsamples are considered.

In a second example random values for the "true" value $\pi = 0.50$, and thus the same value as the null hypothesis, were created with the help of a random generator. The corresponding sample with a size of $n = 30$ has a P value of 0.100 and thus does not indicate a significant result on the 5% level of significance. The null hypothesis cannot be rejected. The P values of the $n - 1 = 29$ possible partial samples are shown in Figure 2.

What is striking here is that the P value for the partial sample $n = 16$ is $P = 0.04$ and therefore lies under the 5% level of significance and consequently leads to a rejection of the null hypothesis even though the "true" value upon which the random sample was created is $\pi = 0.50$ and thus the same as the null hypothesis. The same is true for the sample sizes $n = 23$, 24, and 26. Their P values also lead to a rejection of the null hypothesis. Once again the temptation is great in these cases to report a significant result by choosing sample sizes of 16, 23, 24, or 26 in which we have significant P values. This would once again mean that valid information is dropped; this is inacceptable. If, in contrast, the P values from all subsamples are considered, it appears likely that the null hypothesis cannot be rejected, as discussed below.

A final example should clarify the situation further: the "true" value that the binomially distributed random value creates is $\pi = 0.70$. The P value of the corresponding random sample with a size of $n = 30$ is $P = 0.100$. The null hypothesis can therefore not be rejected at a significance level of 5%. A graph of the resulting $n - 1 = 29$ partial samples' P values can be seen in Figure 3.

From a sample size of $n = 13$ onwards nearly all of the P values until $n = 27$ are smaller than the level of significance. Only when $n = 27$ is the P value greater than 0.05. In this case, in contrast to the previous examples, it would not be possible to reject the null hypothesis at a sample size of $n = 27$ as we have a P value of 0.062 which is greater than the level of significance. The same is true for $n = 29$. In this example again, using all P values will lead to the correct result, namely, the rejection of the null hypothesis.

Consequently, we can draw the following conclusions. If as in usual practice only a P value at $n = 30$ is calculated, the null hypothesis cannot be rejected in the three examples even though this decision is wrong in two of the three situations. A decision based solely on a random sample's significance test, with the full sample size of n, does not take all the available information into consideration. What is missing for a well-rounded picture of any random sample is given in the $n - 1$ partial sample's P values as demonstrated in

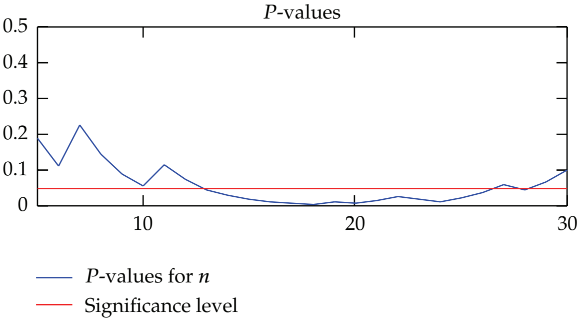

Figure 3: P values of 30 partial samples with $\pi = 0.70$ and $H_0 : \pi = 0.50$. The partial samples 1 to n are given on the abscissa, on the ordinate the corresponding P values. The red line indicates the significance level of 5%.

the examples and graphs above. The question that remains unanswered is how, based on the additional information, can we numerically make a distinction between a significant rejection or a nonrejection of the null hypothesis?

3. Bootstrap

One method to include the additional information given in the partial sample's P values is the bootstrap method [6]. This method does not require a particular type of P value distribution and still enables an estimation of the P value's unknown distribution function including mean and variance. Consequently, the confidence interval for the unknown "true" P value can also be determined, which contains not only information about the sample n, but also about its $n - 1$ partial samples.

In the first example given here the "true" value was $\pi = 0.55$, which created a random sample with a size of $n = 30$. In this sample the null hypothesis $H_0 : \pi = 0.50$ cannot be rejected at a significance level of 5%. The P value for this sample size equals 0.100. The following results are obtained if one does not only take the information from the sample size $n = 30$ into consideration, but also from the $n - 1 = 29$ partial samples calculated with the bootstrap method.

The null hypothesis can be rejected as the sample was taken from a population with $\pi = 0.50$. The probability of this sample result if the null hypothesis is valid is not 0.100 anymore but equals 0.044. The mean and standard deviation of the bootstrap distribution of P values are $E(p) = 0.044$, $\mathrm{St}(p) = 0.0301$.

In the second example the "true" value was $\pi = 0.50$, which created a random sample with a size of $n = 30$, the same as the null hypothesis. In this case the null hypothesis $H_0 : \pi = 0.50$ cannot be rejected at a significance level of 5%. The P value for this sample size equals 0.100. The following results are obtained if one does not only take the information from the sample size $n = 30$ into consideration but also from the $n - 1 = 29$ partial samples calculated with the bootstrap method.

The null hypothesis cannot be rejected as the sample was taken from a population with $\pi = 0.50$. The probability of this sample result if the null hypothesis is valid has risen to 0.1479. The mean and standard deviation of the bootstrap distribution of P values are $E(p) = 0.1479$, $\mathrm{St}(p) = 0.0228$.

In the last example the "true" value was $\pi = 0.70$. With a sample size of $n = 30$, the null hypothesis $H_0 : \pi = 0.50$ cannot be rejected at a significance level of 5%. The P value for this sample size equals 0.100. If one, however, does not only take the information from the sample size $n = 30$ into consideration, but also from the $n-1 = 29$ other sample sizes with the bootstrap method, then it is possible to reject the null hypothesis.

The probability of this sample result if the null hypothesis is valid equals 0.0448. The mean and standard deviation of the bootstrap distribution of P values are $E(p) = 0.0448$, $\mathrm{St}(p) = 0.0526$.

In contrast to our examples, one usually does not know the "true" value of π which created, with the help of a random generator, each of our binomially distributed random samples. Our aim was to estimate the "true" value of π in our samples. It does not play a roll whether we are dealing with a test problem for a $\pi(H_0 : \pi = 0.50)$ or whether we are looking at the famous null hypotheses of parameters such as for example, $H_0 : \pi_1 = \pi_2$ or $H_0 : \mu_1 = \mu_2$ or $H_0 : \rho = 0$.

4. Discussion

Consequently, should the classic significance test lead to a result that is "not significant", then this does not necessarily mean our analysis has come to an end (it does however also not necessarily indicate a "significant" result). The $n - 1$ P values can greatly modify the results of the classic significance test as our examples have shown. This kind of significance test offers the opportunity to not only take the information provided in a classic significance test based on a random sample with the size n into consideration, but also gives us additional information from the P values of the $n - 1$ partial samples of a random sample. These values can be considered as the realizations $p_1, \ldots, p_n$ of the random variables P [7], who also consider the P values random variables but are pursuing a different aim. According to Bernoulli's law of large numbers, P values will converge to the true P value with increasing n:

$$\lim_{n \to \infty} W\left(\left|p - P\right| < \varepsilon\right) = 1. \tag{4.1}$$

The bootstrap method is a possibility to estimate the mean and variance of this random variable.

Our idea to extend the classical significance testing does not describe a sequential analysis in the sense of Wald's sequential probability ratio test [8], where, for example, the sampling can be stopped after reaching a certain decision limit. In our approach, the sample size n is fixed and there is no intention to stop sampling at an earlier stage. We further do not need to consider the type II error and the corresponding effect size since our approach relies on the Fisher model [3, 4] rather than that of Neyman and Pearson [9]. Based on the required randomness of a random sample, the sampling order is also fixed and we do not need to account for the theoretical possibility (sum of binomial coefficients at $n-1, n-2, \ldots$) to analyze more than $n-1$ partial subsamples. In addition, our approach does not constitute a typical multiple testing situation. A multiple testing situation usually requires the testing of multiple hypotheses. In our case, we test only one hypothesis, our main aim is to improve the estimation of the P value using the information of the data in the partial subsamples. In this sense our approach can also be regarded as an estimating problem.

In summary, in this paper, we propose a significance test that takes into account information from P values of partial subsamples of a random sample. We show that the use of the additional P values can greatly modify the results of a classic significance test as an "extended partial data" analysis approach to data mining.

Appendix

Realisation of Three Random Samples

(1) Sample: $1, 1, 1, 1, 1, 1, 2, 1, 1, 1, 2, 2, 1, 1, 1, 2, 2, 2, 1, 1, 1, 2, 1, 1, 1, 2, 1, 2, 2, 2$ $(\pi = 0.55)$.

(2) Sample: $2, 1, 2, 1, 2, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 1, 2, 1, 1, 1, 1, 2, 1, 2, 2, 1, 2$ $(\pi = 0.50)$.

(3) Sample: $2, 1, 2, 2, 2, 1, 1, 1, 1, 1, 2, 1, 1, 1, 1, 2, 1, 2, 1, 1, 1, 1, 1, 2, 2, 1, 1, 1, 2, 1$ $(\pi = 0.70)$.

References

[1] J. Cohen, "The earth is round ($p < .05$)," *American Psychologist*, vol. 49, no. 12, pp. 997–1003, 1994.

[2] G. Loftus, "On the tyranny of hypothesis testing in the social sciences," *Contemporary Psychology*, vol. 36, pp. 102–105, 1991.

[3] R. A. Fisher, *The Design of Experiments, (5th Ed., 1951; 7th Ed., 1960; 8th Ed., 1966)*, Oliver & Boyd, Edinburgh, UK, 1935.

[4] R. A. Fisher, *Statistical Methods and Scientific Inference*, Oliver & Boyd, Edinburgh, UK, 1956.

[5] G. Gigerenzer, "Mindless statistics," *Journal of Socio-Economics*, vol. 33, no. 5, pp. 587–606, 2004.

[6] B. Efron and R. Tibshirani, *An Introduction to the Bootstrap*, Chapman & Hall/CRC, 1994.

[7] D. Boos and I. Stefanski, "P-value precision and reproducibility," *The American Statistician*, vol. 65, no. 4, pp. 213–221, 2011.

[8] A. Wald, "Sequential tests of statistical hypotheses," *The Annals of Mathematical Statistics*, vol. 16, no. 2, pp. 117–186, 1945.

[9] J. Neyman and E. Pearson, "On the problem of the most efficient tests of statistical hypotheses," *Philosophical Transactions of the Royal Society of London. Series A*, vol. 231, pp. 289–337, 1933.

Advances in Operations Research
Advances in Decision Sciences
Mathematical Problems in Engineering
Algebra
Journal of Probability and Statistics
The Scientific World Journal
International Journal of Differential Equations
International Journal of Combinatorics
Advances in Mathematical Physics
Journal of Complex Analysis
Journal of Mathematics
International Journal of Stochastic Analysis
Abstract and Applied Analysis
Discrete Dynamics in Nature and Society
International Journal of Mathematics and Mathematical Sciences
Journal of Discrete Mathematics
Journal of Function Spaces
Journal of Applied Mathematics
Journal of Optimization