OPEN ACCESS

GAVIN PUBLISHERS

Journal of Biostatistics & Biometrics

Research Article

Girondot M and Guillon JM J Biostat Biom 2018: JBSB-101.
DOI: 10.29011/JBSB-101. 000001

The *w*-value: An Alternative to *t*- and χ^2Tests

Marc Girondot*, Jean-Michel Guillon

Ecologie, Systematique, Evolution, Université Paris-Sud, AgroParisTech, Centre National de la Recherche Scientifique, Université Paris Saclay, Orsay, France

***Corresponding author** : Marc Girondot, Laboratoire Ecologie, Systematique, Evolution, Université Paris-Sud, AgroParisTech, CNRS, Université Paris Saclay, 91405 Orsay, France. Tel: +33-169157230 ; Fax:+33-169157353; Email: marc.girondot@u-psud.fr

Citation: Girondot M, Guillon JM (2018) The *w*-value: An Alternative to *t*- and χ^2 Tests. J Biostat Biom 2018: JBSB-101. DOI: 10.29011/JBSB-101. 000001

Received Date: 02**,** February 2018; **Accepted Date: 20**, February, 2018; **Published Date:** 02, March, 2018

Abstract

Reproducibility is central in science and statistics are one of the tool used to ensure that conclusions are supported by data. However, high false positive rate is a common problem denounced regularly when hypotheses testing using null and alternative hypotheses are used. The origin of this high false positive rate is deeply anchored in the method because the reported *p*-value is not what generally the researcher was thought to test: the *p*-value is the probability to observe the statistics summarizing the data if the null hypothesis H_0 is true, but the researcher would generally like to know what the probability is that H_0 is true given the observed data. We propose an alternative to *t*-test and χ^2-test, two of the most common tests, that used model comparison using Schwarz's Bayesian information criterion weights. Thus, we were able to answer the question that generally the researcher wants an answer: what is the probability that these *k*-series are obtained from a single set of parameters and then what is the probability that they are identical.

Keywords: BIC; BIC Weight; Hypothesis Testing; Schwarz's Bayesian Information Criterion

Introduction

Despite the numerous papers published over the last few years warning against the high rate of false positive results due to *p*-value statistics use [1-3], *p*-value-based tests are still widely used, and it is rare to see a quantitative study without them. Among the problems noted, one is particularly serious: the *p*-value does not generally respond to the question that the researcher asks: the *p*-value is technically the probability to observe the statistics summarizing the data if the null hypothesis H_0 is true, but the researcher would generally like to know what the probability is that H_0 is true given the observed data. These two probabilities are linked by Bayes' theorem and there is a clear relationship between both [4] but the first one can be much lower than the second one [5]. Further, the lack of reproducibility in scientific studies can be largely attributed to the conduct of significance tests at unjustifiable levels of significance [1] (classically type I error rate α=0.05).

To understand why this situation occurs, we must recall that statistics are not only a science, but also a toolbox that many researchers use without a profound knowledge in this scientific field. This situation has been reinforced by the emergence of user-friendly software or cooking-like recipes in various web sites (e.g. https://stats.stackexchange.com). Users of statistical tools are sometimes not aware that the *p*-value is a poor measure of evidence and, even if they are, they may continue to use the *p*-value out of habit or because alternative solutions require in-depth mathematical and statistical knowledge [6] with no ready-to-use out-of-the-box method being available. It is, however, the responsibility of authors, referees, and editors to ensure that statistics do not induce false levels of evidence.

We have developed one general methodology to be used and applied it to two common tests: the χ^2 and the *t*-tests. Our aim is to keep this method as simple as possible so as to be an alternative to the null hypothesis significance tests. The methodology is based on model selection using Schwarz's Bayesian Information Criterion (BIC) [7]. Given a set of models and postulating that the prior probabilities of these models are identical, the model with the highest posterior probability is the one that minimizes the BIC [8].

This statistic allows us to rank models based on the quality of fit at the same time as they penalize for the number of parameters used for fitting. For the alternatives to the $\chi^2$and *t*-tests presented here, the first model is fitted globally for all series, while the second uses series-specific parameters. The support for each model ("all series are obtained from the same distribution", hereafter

named "similar models", or "series are obtained from different distributions", hereafter named "different models") comes from the Schwarz weights [9]. For each of the models considered, Schwarz weights are the posterior probability of the model, given the data, the model set, and equal prior model probabilities.

Materials and methods

Series as collection of observations

Let us take k series each with l_i values with i from 1 to k. When a t-test is used to detect a significant difference among the series, the result cannot be used as a measure of evidence due to previously described pitfalls. Here we calculate the maximum likelihood of the hypothesis that each series was obtained from a specific distribution, for example a normal distribution as for t-test [10], with a series-specific mean and possibly a series-specific standard deviation; it is named L_k. Alternatively, the maximum likelihood of the hypothesis that all series come from the same distribution can be calculated; it is named L.

Series as contingency tables

A table of contingency is a type of table in a matrix format that displays the (multivariate) frequency distribution of the variables. Two kinds of tests of performed in routine on such a table, namely homogeneity or conformity tests. The test for homogeneity is evaluating the equality of several populations of categorical data whereas the test for conformity is evaluating the equality of several populations of categorical data as compared to a known distribution. The χ^2 test used classically for such a purpose is the approximation of a likelihood ratio test. However, again the result cannot be used as a measure of evidence due to previously described pitfalls of the p-value measure of evidence.

As an alternative we estimate the likelihood for each k series of c categories fitted separately or altogether. The likelihood L_k of a series k of N_k total observations (separated in c categories, $n_{k.1}$ to $n_{k.c}$) is based on a multinomial distribution

$$L_k = \frac{N!}{n_{k.1}!\, n_{k.2}! \dots n_{k.c}!}\, p_1^{n_{k.1}} p_2^{n_{k.2}} \dots p_c^{n_{k.c}}$$

The p values are based on external information for conformity test (p_1 to p_{c-1}) or based on structure of the data for homogeneity test with $p_i = \sum_{j=1}^{k} n_{j.i} / \sum_{j=1}^{k} N_j$

Comparison of hypotheses

The Schwarz's Bayesian information criterion for a series of l_i observations fitted using a model with d parameters (d=2 for mean and standard deviation for Gaussian distribution for example) with likelihood L_i is $BIC = -2\ ln(L_i) + d\ ln(l_i)$ [7]. When all the series are fitted using a same set of d parameters, the Schwarz's Bayesian information criterion is:

$$BIC_{similar} = -2\ ln(L) + d.ln(\sum l_i)$$

When the k series are fitted using series-specific sets of parameters the BIC is:

$$BIC_{different} = -2 \sum ln(L_i) + d.k.ln(\sum l_i)$$

Model weights

$BIC_{similar}$ and $BIC_{different}$ are compared using Schwarz weights. The Schwarz weight reported here as the w-value (for weight-value) is the posterior probability that the single common model is the best [11].

These model comparisons were implemented as R functions series.compare() and contingencyTable.compare() using Akaike Information Criterion (AIC), AICc [11], or BIC in HelpersMG R package available in CRAN (https://cran.r-project.org/web/packages/HelpersMG/index.html). We recommend the use of BIC that supposes that the true model is among the tested models. Obviously, the true model is tested in the procedure that we describe here, the two alternatives being A=B or A≠B. To make the usage even easier, a web page http://134.158.74.46/compare/ has been published to perform the estimates. The same logic can be applied to other tests such as a regression analysis.

Results

The discrepancy between the p-value and the probability that the series need series-specific modelling can be demonstrated using two simple simulations. Let us use a series of random numbers obtained from normal distributions: series A comprising 100 numbers with a mean of 10 and a standard deviation of 2, and series B also comprising 100 numbers with a standard deviation of 4 and a specific mean varying from 8 to 13 by step of 0.1. Then the hypothesis that series A and B are similar is tested using a p-value based on a t-test or a w-value using BIC. We clearly see that, for usual levels of evidence, the p-value too often rejects the two series as being similar (Figure. 1). A w-value of 0.05 is obtained for a p-value of around 0.002, while a p-value of 0.05 corresponds to a w-value of 0.69 (Figure. 1).

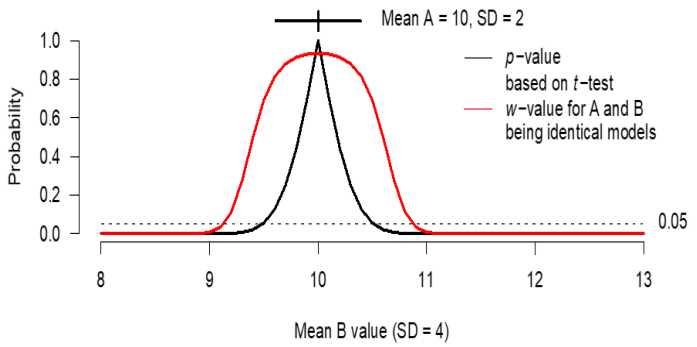

Figure 1: Level of evidence for two series A and B being similar using *p*-value or *w*-value (series A, n=100, mean=10, SD=2; series B, n=100, mean from 8 to 13, step 0.1, SD=4).

The same comparison can be made with a contingency table using a multinomial distribution for likelihood. A total of 100 observations belonging to group A shows 25 being of type P and 75 being of type Q, while group B also comprises 100 observations with P numbers ranging from 0 to 100. These observations are then compared using a χ^2-test (*p*-value) or *w*-value with BIC. Again, the *p*-value calculated after a classical χ^2 test too often rejects the hypothesis that groups A and B are similar (Figure. 2). A *w*-value less than 0.05 is obtained for a *p*-value between 0.01 and 0.03, while a *p*-value of 0.05 corresponds to a *w*-value of 0.16 (Figure. 2). However, a *p*-value < 0.01 should not be used as a rule of thumb to perform a test, because this significance level needs to be estimated for each dataset.

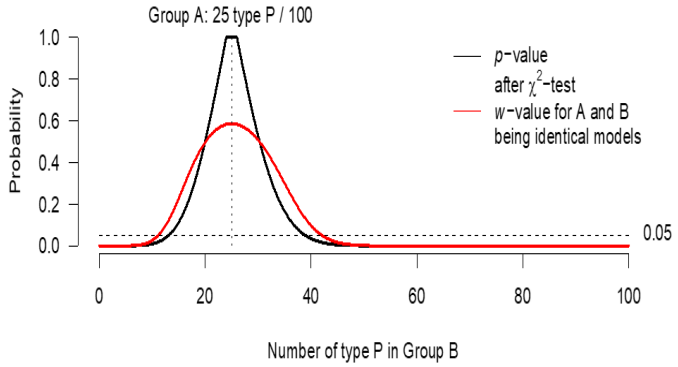

Figure 2: Level of evidence for two groups A and B being similar using *p*-value and *w*-value (series A, P=25 and Q=75; series B, P=0 to 100 and Q=100-P).

Discussion

When performing a test as described here, researchers want to investigate whether several series are identical or not. The *p*-value does not answer this question, but rather gives the probability that these data could have been obtained under the null hypothesis. It can be considered just as an index of evidence [12]. Contrary to the *p*-value, the *w*-value has a simple interpretation as a measure of evidence: it gives the posterior probability that the different series could have been generated from a single distribution [13]. The *w*-value is clearly closer to the researcher's expectation when performing a test. In any case, we recommend against the use of a cut-off value for the *w*-value (such as $w < 0.05$) because the test does not aim to reject a hypothesis but allows ranking different hypotheses according to their probabilities. Furthermore, the use of *w*-values does not prevent the need for better practices in test design [14]. We hope that these ready-to-use out-of-the-box tools will make statistics users more concerned about the nature of statistical evidence used for hypothesis testing.

Acknowledgements: We would like to thank Olivier Martin (Laboratoire de Génétique Quantitative et Evolution) and Franck Courchamp (Laboratoire Ecologie, Systématique, Evolution) for their critical reading of this manuscript and Victoria Grace (http://www.english-publications.com) who helped us in correcting our English text.

Author contributions: M. G. and J.-M. G. contributed to method development, reviewed literature and wrote the paper.

Competing financial interests. None.

References

1. Nuzzo R (2014) Statistical errors. Nature. 506: 150-152.
2. Baker M (2016) Statisticians issue warning on P values. Nature. 531: 151-152.
3. Wasserstein RL, Lazar NA (2016) The ASA's statement on p-values: context, process, and purpose. The American Statistician. 70: 129-133.
4. Murtaugh PA (2014) In defense of P values. Ecology. 95: 611-617.
5. Johnson VE (2013) Revised standards for statistical evidence. Proc Natl Acad Sci USA. 110: 19313-19317.
6. Raftery AE (1995) Bayesian model selection in social research. Sociological Methodology. 25: 111-163.
7. Schwarz G. (1978) Estimating the dimension of a model. The Annals of Statistics. 6: 461-464.
8. Kass RE, Raftery AE (1995) Bayes factors. J Am Stat Assoc. 90: 773-795.
9. Wagenmakers EJ, Farrell S (2004) AIC model selection using Akaike weights. Psychonomic Bulletin & Review. 11: 192-196.
10. Casella G, Berger R (2001) Statistical Inference. Duxburi: Pacific Grove. CA.
11. Burnham KP, Anderson DR (2002) Model selection and multimodel inference: A practical information-theoretic approach. Springer Verlag.

12. Burnham KP, Anderson DR (2014) P values are only an index to evidence: 20th- vs. 21st-century statistical science. Ecology, 95: 627-630.

13. Burnham KP, Anderson DR (2004) Multimodel inference: Understanding AIC and BIC in model selection. Sociological Methods & Research. 33: 261-304.

14. Forstmeier W, Wagenmakers EJ, Parker TH (2017) Detecting and avoiding likely false-positive findings - a practical guide. Biol Rev Camb Philos Soc. 92: 1941-1968.