

Significance tests or confidence intervals: which are preferable for the comparison of classifiers?

Daniel Berrar^{a*} and Jose A. Lozano^b

^aInterdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology, 4259 Nagatsuta, Midori-ku, Yokohama 226-8502, Japan; ^bDepartment of Computer Science and Artificial Intelligence, Intelligent Systems Group, University of the Basque Country UPV/EHU, Manuel de Lardizabal, 1, 20018 Donostia–San Sebastián, Gipuzkoa, Spain

(Received 22 July 2011; final version received 26 February 2012)

Null hypothesis significance tests and their p -values currently dominate the statistical evaluation of classifiers in machine learning. Here, we discuss fundamental problems of this research practice. We focus on the problem of comparing multiple fully specified classifiers on a small-sample test set. On the basis of the method by Quesenberry and Hurst, we derive confidence intervals for the effect size, i.e. the difference in true classification performance. These confidence intervals disentangle the effect size from its uncertainty and thereby provide information beyond the p -value. This additional information can drastically change the way in which classification results are currently interpreted, published and acted upon. We illustrate how our reasoning can change, depending on whether we focus on p -values or confidence intervals. We argue that the conclusions from comparative classification studies should be based primarily on effect size estimation with confidence intervals, and not on significance tests and p -values.

Keywords: null hypothesis significance testing; p -value; confidence interval; classification; reasoning

1. Introduction

Comparative classification studies generally focus on a single performance metric such the area under the ROC curve or the error rate. In the past, single point estimates of such measures were often compared directly. For example, assume that classifier A achieves an error rate of $\epsilon_A = 0.10$, and classifier B achieves an error rate of $\epsilon_B = 0.18$. As $\epsilon_A < \epsilon_B$, model A would be declared the winner.

Recently, however, comparisons of point estimates are less frequently used, and null hypothesis significance tests are now gaining increasing popularity in machine learning (Demšar 2006; Cawley and Talbot 2010), although they were introduced to the machine learning and AI community many years ago (Kibler and Langley 1988). In his recent editorial, Langley (2011) criticised that many researchers in machine learning seem to be pre-occupied with statistics.

A null hypothesis test takes the observed performance measure as input and assesses whether the difference between the classifiers is significant or not (Dietterich 1998;

*Corresponding author. Email: berrar.d.aa@m.titech.ac.jp

Nadeau and Bengio 2003; Bouckaert and Frank 2004; Berrar, Bradbury, and Dubitzky 2006; Demšar 2006; Garcia and Herrera 2008). This significance is quantified in the p -value. If the significance test yields a p -value smaller than a pre-defined threshold (generally, 0.05), then the difference is considered ‘significant’; hence, one model is declared significantly superior to another. In the example, the difference $\delta_{AB} = 0.08$ might meet this criterion, and A might be considered significantly better than B . More recent research in this field has focused on methods for addressing the problem of multiple testing, i.e. methods that correct individual p -values for multiplicity effects when several classifiers are compared (Demšar 2006; Garcia and Herrera 2008).

On the other hand, confidence intervals for the *effect size*, i.e. intervals for the difference between the performances, are still less frequently reported. In general, those publications that do report such intervals also include a significance test and base the conclusions only on the dichotomous outcome of the statistical test. Thus, the current practice of evaluating machine learning algorithms is dominated by statistical testing and p -values.

However, in other fields of science, notably in epidemiology, the practice of significance testing has led to heated debates for several decades (Goodman 2008). There are critical voices that caution against the use of significance tests and that try to ban the use of p -values altogether (Rothman 1978). In machine learning, however, these critical voices have been largely unheard. Do these criticisms have any relevance for machine learning? A debate of this question is timely, given the ever-increasing popularity of statistical testing in this field.

In this article, we are interested in the question whether significance testing and interval estimation for effect sizes are just two sides of the same coin. Or could they have a different influence on our interpretation of classification results? If that is the case, which approach would be preferable? In this study, our conclusions about the significance of differences in performance are essentially the same, regardless of whether we use the significance test or the confidence intervals. However, we conclude that confidence intervals are preferable to significance testing. Confidence intervals shift the focus of our attention on the effect size and its reasonably bounds. In contrast, the scalar p -value encourages dichotomisations into ‘significant’ and ‘non-significant’ results, which obscure important aspects of the classifiers that confidence intervals bring to the forefront. Effect size estimation should therefore guide our interpretation of classification results.

This article is organised as follows. First, we describe the problem of statistical evaluation of classifiers. Next, we review fundamental problems of significance testing. We then consider a concrete classification problem to illustrate some of these problems. We focus on the small-sample scenario, which is common in the life sciences such as *omics* research, where the number of test cases is frequently in the order of a few dozen only. We consider fully specified classifiers that are compared on an independent test set. After some formal preliminaries and mathematical details for the statistical test and confidence interval, we carry out a classification study. We then discuss how the two approaches to statistical inference shift the focus of our attention to different aspects of the results.

2. Problem statement

Figure 1 schematically illustrates the evaluation process of two models, A and B , that are compared in a classification task on the basis of a performance measure ε . Without loss of

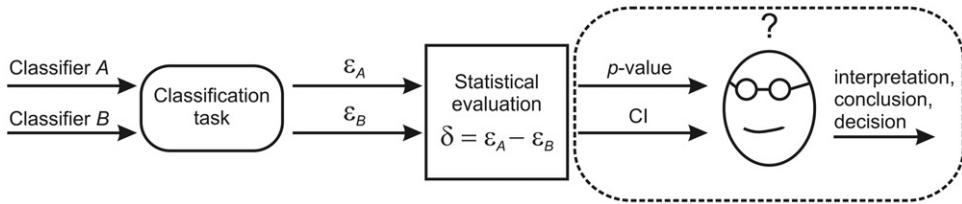


Figure 1. Evaluation process of two classifiers, A and B , based on the performance measure ε .

generality, we may assume that this measure is the error rate. This measure is then subjected to a statistical evaluation, which assesses the difference $\delta = \varepsilon_A - \varepsilon_B$. If this evaluation involves a significance test, then the null hypothesis is that no difference in performance exists between A and B . The outcome of this test is a p -value. If the evaluation involves the construction of a confidence interval (CI), then a lower bound (LB) and upper bound (UB) on δ define a reasonable range for the true difference in performance, i.e. $CI = [\delta_{LB}, \delta_{UB}]$. In either case, the output of the statistical evaluation is then the input to a human reasoning process that involves interpretations, draws conclusions and finally makes decisions. Does the outcome of this process depend on whether the p -value or the CI is the input?

To address this question, we consider the problem of comparing fully specified classifiers on an independent test set. By *fully specified classifier*, we mean a concrete model that has been induced from a data set using a learning algorithm. This model is the final model that will be used to predict new, unseen cases. In contrast, the term *learning algorithm* refers to a specific machine learning algorithm that was used to build this model. Commonly, the final model is assessed on a test set that was not involved in the model building process, which represents the ‘litmus test’ for classifiers. The performance on the independent test set determines the final verdict: whichever model performs best on this test set is declared the ‘winner’ and will be applied for similar data sets that become available in the future.

Figure 1 describes the evaluation process for only two models, but we extend the comparison to multiple models in our analysis. Our goal is not to find the ‘best’ classifier for our classification task. Instead, we are interested in elucidating the different impacts of p -values and CIs on our decision making process (cf. the dotted box in Figure 1). Hence, the concrete classification task represents only an illustrative example within which we analyse this process.

3. Significance tests and CIs: two sides of the same coin?

The p -value is defined as the conditional probability to observe a result as extreme as or more extreme than the actually obtained one, given that the null hypothesis is true (Goodman 1993). In our context, this null hypothesis is that no difference in performance exists between two classifiers, i.e. $p\text{-value} = \Pr(\text{classification result as observed or more extreme} \mid H_0 = \text{no difference in performance})$.

There are several problems with this value. The first problem becomes apparent when we inspect the formal definition of the p -value. The formula includes probabilities for events that, in fact, did not occur: classification results as observed *or more extreme than*

the actually observed ones. Johnson (1999) noticed this quandary in the context of wildlife management. Berger and Berry (1988) pointed out that the idea of data that are *more extreme* actually requires knowledge of the intentions of the scientist who designed the experiment. Thus, the idea of *being more extreme* entails an ill-defined problem that is prone to subjective interpretations.

The p -value is often a source of misinterpretations (Fralely and Marks 2007; Goodman 2008). Arguably, the most critical misinterpretation is the *fallacy of the transposed conditional*: the p -value is mistaken as the posterior probability that the null hypothesis is true. For example, assume that we obtain a p -value of 0.01 for the difference in performance between two classifiers. The following interpretations are incorrect: ‘*The probability that there is no difference in performance is 0.01*’ and ‘*The probability that one classifier performs better than the other is 0.99*’. Statements about posterior probabilities can only be based on a Bayesian analysis, but the commonly used parametric and non-parametric tests in machine learning such as, for example, ANOVA, t -tests, the Wilcoxon signed ranks test, or Friedman test (Dietterich 1998; Demšar 2006; Garcia and Herrera 2008) are not Bayesian. Thus, the p -value seems to be of rather limited help for our reasoning. Indeed, Goodman stressed the elusive meaning of the p -value, as it is not part of any formal calculus of inference (Goodman 1999, 2008). Johnson (1999) described the p -value even as irrelevant. There are of course also voices defending significance testing (Mulaik, Raju, and Harshman 2007). But the criticisms deserve to be taken very seriously in machine learning: the p -value resulting from a significance test cannot be interpreted as the probability that one classifier has a superior performance over its competitor.

Suppose that we observe a significant (p -value = 0.01) difference between A and B , where A outperforms B on a test set of n cases. Does this p -value tell us anything about the replicability of the results? For example, suppose that we compare the performance of A and B again, this time on a different test set of size n from the target distribution. What does the p -value of 0.01 from the first study reveal about the likelihood that we will observe a significant result in the second study? Alas, the answer is ‘nothing’. Under the assumption that the null hypothesis is false (here, that a difference in performance really exists), the statistical power of an experiment determines whether a significant result can be replicated or not. This power mainly depends on three factors: (i) the alpha level for the statistical test; (ii) the true effect size in the target distribution; and (iii) the size of the test set (Fralely and Marks 2007). As the power does not depend on the p -value, the p -value is indeed irrelevant for assessing the likelihood of replicability (Dixon 1998).

Another misinterpretation is that a very small p -value (say, <0.001) implies a large effect (i.e. a big difference between the true performances), whereas a p -value just below 0.05 implies a moderate effect. This interpretation, however, is not correct because the p -value is a function of both the sample size (and thereby the precision) and the effect size. A non-significant p -value could therefore mean that, most likely, no difference in performance exists, or that the sample size (i.e. the size of the test set used to benchmark the classifiers) was too small to allow the detection of the effect. Thus, if we fail to observe a significant difference in performance, then we cannot decide between (i) the classifiers indeed perform equally or (ii) the classifiers do have a different performance, but the size of the test set is too small to reveal this difference.

These problems of significance testing are not a new ones, though. In fact, they have fueled heated debates for at least 70 years (Goodman 2008). It is not a problem specific to machine learning but pesters other disciplines as well, including epidemiology (Rothman 1978; Poole 2001; Goodman 2008), psychology (Schmidt 1996; Denis 2003; Fralely and

Marks 2007), educational research (Nix and Barnette 1998) and wildlife research (Johnson 1999), to name but a few. The landmark papers by Dietterich (1998) and (Demšar 2006) may have contributed to the popularity of significance tests, although both authors clearly mentioned their caveats. Random permutation tests are an alternative to ‘classical’ null hypothesis testing and are now increasingly used by the machine learning community (Ojala and Garriga 2010). In short, permutation tests compute a statistic of interest (e.g. the error rate) and then compare this statistic to its empirical distribution, which we obtain under the null hypothesis. Hence, permutation tests differ from conventional null hypothesis testing with respect to the way in which the p -value is calculated. They do not resolve the problem of inferential interpretation. In machine learning, these problems have so far been largely ignored, with only a few recent exceptions.¹ Recently, Demšar pinpointed the problem; however, he currently sees no alternatives to significance testing (Demšar 2008). Drummond, too, criticises the current practice in machine learning but does not propose an alternative to significance testing either (Drummond 2008; Drummond and Japkowicz 2010).

An alternative is CIs for the effect size. A CI disentangles the effect size from the measure of uncertainty and can still be interpreted in terms of significance: in general, if a 95%-CI includes the null value (here, no difference in performance), then the test gives a p -value ≥ 0.05 ; if the interval does not include the null value, then the test gives a p -value < 0.05 . But the width of the CIs explicitly tells us something about the uncertainty of the estimate (and hence, we should refrain from limiting the interval to a mere test of significance). For example, consider the 95%-CI [0.03, 0.05] for the difference of the error rates of two models. This interval provides a more robust estimate of the difference in performance than, say, the interval [0.01, 0.25]. The interval [0.03, 0.05] does not include 0 and therefore implies statistical significance. However, the true error rate of one model may be only 5% lower than the error rate of another model. This difference in performance may be too small to be of any practical relevance, or it may be decisive – this depends on the concrete application. Thus, the CI shifts the focus of our attention to the effect size and the likely boundaries of the true effect size in the population (i.e. the set of cases similar to the test cases). Importantly, the interval also encourages us to think about the context or problem domain when we reason about the effect size. In contrast, a test’s p -value does not give any information beyond what is already conveyed by a confidence interval; moreover, p -values often invite dichotomisations into ‘significant’ and ‘non-significant’ results.

Thus, p -values and intervals are of a different quality for our reasoning process. Significance tests and CIs are therefore *not* two sides of the same coin. In the following sections, we investigate how this different quality can influence our conclusions about differences in performance.

4. Formal preliminaries

Consider two fully specified classifiers A and B whose performance is to be compared on the same test set. The observed frequencies of correct/incorrect classifications are shown in Table 1. This 2×2 table summarises the classification results for a 0-1 loss function. Here, a is the number of cases misclassified by neither model A nor B ; b is the number of cases misclassified by A but not by B ; c is the number of cases misclassified by B but not by A ; d is the number of cases misclassified by both A and B and n is the total number of test cases.

Table 1. 2×2 table of the numbers of misclassified cases by model *A* and *B*.

<i>a</i>	<i>b</i>	<i>a + b</i>
<i>c</i>	<i>d</i>	<i>c + d</i>
<i>a + c</i>	<i>b + d</i>	<i>n = a + b + c + d</i>

The observed frequencies *a*, *b*, *c* and *d* result from a multinomial distribution with (unknown) underlying probabilities π_a , π_b , π_c and π_d . The sample estimates for these probabilities are $p_a = a/n$, $p_b = b/n$, $p_c = c/n$ and $p_d = d/n$. From Table 1, we can derive the statistic X^2 as shown in Equation (1), which approximately follows a χ^2 -distribution with one degree of freedom (May and Johnson 1997):

$$X^2 = \frac{n(|p_b - p_c| - |\pi_b - \pi_c|)^2}{(\pi_b + \pi_c) - (\pi_b - \pi_c)^2} \sim \chi^2_{1df}. \tag{1}$$

McNemar’s (1947) test is the standard hypothesis test for comparing two models on an independent test set with a 0-1 loss function (Dietterich 1998). We focus here on the error rate because it is arguably still the most widely used performance measure (Demšar 2006). The concrete measure, however, is irrelevant for our analysis. We are interested in contrasting the conclusions based on the test with the conclusions based on a CI for the effect size, i.e. the difference between the true error rates. As we compare the performance on the same test set, we need to take into account that the predictions made by the individual classifiers are correlated. As we compare differences in error rates, we need CIs for differences in binary correlated proportions. Furthermore, the intervals should have acceptable coverage probability even for small sample sizes. Quesenberry and Hurst (1964) proposed a method for constructing simultaneous CIs for multinomial proportions. May and Johnson (1997) have shown that these intervals are valid even for small sample sizes. In the next section, we follow the notation by May and Johnson (1997) and apply the interval estimation method to the problem of effect size estimation in classification. First, we consider the comparison of only two models, and then we extend the comparison to multiple models.

5. Significance test for the difference in performance on the same test set and a 0-1 loss function

To evaluate the differences in performance between any two classifiers, we focus on the probabilities π_b and π_c for discordance between the models: if the two models have the same performance, then these probabilities should be approximately the same. Thus, we can cast this problem into a statistical significance test by stating the null hypothesis $H_0 = |\pi_b - \pi_c| = 0$. This is the null hypothesis in McNemar’s (1947) test. Our best estimates for these probabilities are the sample estimates p_b and p_c , which we plug into Equation (1). The sample size, *n*, in the nominator is then cancelled out because $p_b = b/n$ and $p_c = c/n$. Thereby, we obtain McNemar’s test statistic as shown in Equation 2.

$$X^2 = \frac{n(|b/n - c/n| - 0)^2}{b/n + c/n - 0} = \frac{|b - c|^2}{b + c} \sim \chi^2_{1df}. \tag{2}$$

This statistic follows approximately a χ^2 -distribution with one degree of freedom. Equation (2) often includes Yates' correcting term -1 in the nominator to account for the fact that the χ^2 -distribution is continuous, whereas X^2 is discrete: $(|b - c| - 1)^2$. If the null hypothesis of equal performance is correct, then the probability that X^2 is greater than $\chi^2_{1,0.95} = 3.841$ is less than 0.05. Hence, if $X^2 > 3.841$, then we can reject the null hypothesis that the two classifiers perform the same. Note, however, that this rationale is strictly valid only for those comparative studies that include just two models. If we include more than two classifiers, then we enter the arena of multiple testing.

McNemar's test is conservative for small-sample data sets (May and Johnson 1997), specifically if it includes Yates' correction; however, this may actually be desirable in small-sample settings where a safeguard against false positive results (i.e. one model is erroneously declared superior to another) may not be a bad thing. We can calculate an exact binomial version of McNemar's test without the χ^2 approximation (Sheskin 2007), which is then equivalent to the binomial sign test. In our experiments, we focused on the standard McNemar test with continuity correction. For completeness, we mention the results of the exact test, but as these did not affect our conclusions, we focus on the standard McNemar's test (with Yates' correction).

6. CIs for the effect size

When we compare the performance of two classifiers, we are interested in their true performances, i.e. their expected performances on the entire population of data from which the test set is only a sample. Figure 2 visualises this idea. The observed performances on the test set are used to make inference for the true performances. Here, we quantify the performance in terms of error rate.

Definition (effect size): Let τ_A and τ_B be the true error rates and ε_A and ε_B be the observed error rates of classifiers A and B , respectively. The *effect size* is then the difference $\delta_\tau = |\tau_A - \tau_B|$ between the true error rates τ_A and τ_B of two models A and B .

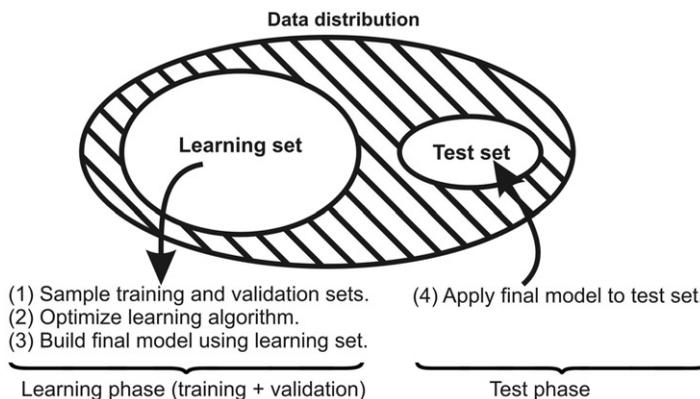


Figure 2. Learning set, test set and data distribution. The learning and test set are samples from an unknown target distribution. The fully specified classifier results from the application of the learning algorithm to the data in the learning set. This final model is applied to the test set, which serves as benchmark for comparing different models.

When the sample sizes are relatively small (i.e. in the order of a few dozen, as in our examples), variance estimations are notoriously difficult. We therefore need a CI that does not rely on a variance estimation of the true error rates. Quesenberry and Hurst (1964) showed that the limits of a $(1 - \alpha) \times 100\%$ CI for the difference between the probabilities of discordance, $|\pi_b - \pi_c|$, can be derived from Equation (1) without estimating the variances of π_c and π_d .

Proposition: *A CI for the effect size can be derived from Equation (1).*

Proof: Our best estimates for the true error rates are the sample estimates, so our best estimate for the effect size is $\delta_\varepsilon = |\varepsilon_A - \varepsilon_B|$. From Table 1, we see that the sample error rate of model *A* is $\varepsilon_A = (b + d)/n$, and the sample error rate of model *B* is $\varepsilon_B = (c + d)/n$. The absolute difference is $\delta_\varepsilon = |\varepsilon_A - \varepsilon_B| = |(b + d)/n - (c + d)/n| = |b/n - c/n| = |p_b - p_c|$. We recognise this difference as the sample estimate for the difference between the probabilities of discordance, $|\pi_b - \pi_c|$. Therefore, the difference between the true error rates is equivalent to the difference between the probabilities of discordance. \square

May and Johnson (1997) noted that solving the inequality (3) for δ_τ gives the CI for the difference between the probabilities of discordance. Thus, by following May and Johnson (1997), we find the CI for the effect size.

$$\frac{n(|p_b - p_c| - \delta_\tau)^2}{(\pi_b + \pi_c) - \delta_\tau^2} \leq \chi_{1-\alpha}^2 \quad (3)$$

Here, $\chi_{1-\alpha}^2$ is the critical value of the χ^2 -distribution with one degree of freedom. Using again p_b and p_c as estimates for π_b and π_c , respectively, we can solve the inequality (see Appendix for details). As lower and upper bound of δ_τ , we obtain δ_{LB} and δ_{UB} , respectively.

$$\delta_{LB} = \frac{n|p_b - p_c|}{n + \chi_{1-\alpha}^2} - \frac{\sqrt{\chi_{1-\alpha}^2[(p_b + p_c)(n + \chi_{1-\alpha}^2) - n(p_b - p_c)^2]}}{n + \chi_{1-\alpha}^2} \quad (4)$$

$$\delta_{UB} = \frac{n|p_b - p_c|}{n + \chi_{1-\alpha}^2} + \frac{\sqrt{\chi_{1-\alpha}^2[(p_b + p_c)(n + \chi_{1-\alpha}^2) - n(p_b - p_c)^2]}}{n + \chi_{1-\alpha}^2} \quad (5)$$

The interval $[\delta_{LB}, \delta_{UB}]$ defines a $(1 - \alpha) \times 100\%$ CI for the difference between the true error rates of two classifiers that were applied to the same test set. For example, $\chi_{1-\alpha}^2 = 3.841$ gives a 95% CI.

7. Materials and methods

We deliberately included only classifiers with a known true performance (such as the flip of a coin), so that we could contrast our expected differences with the experimental results. All experiments and all analyses were carried out using R 2.10.1 (R Development Core Team 2009).

We used a publicly available microarray data set of 119 breast cancer samples (Sotiriou et al. 2006). Each sample (i.e. case or instance) is characterised by the expression profile of 22215 genes, which are the attributes of the sample. The data were normalised as described

in the original study. According to the standard pre-processing for this type of data, all expression values were \log_2 -transformed and then median-centred. The data set comprises two classes of 34 estrogen receptor negative (ER-) and 85 estrogen receptor positive (ER+) cases.

We used stratified random sampling to generate one learning set of 69 samples (20 ER- and 49 ER+) and one test set of 50 samples (14 ER- and 36 ER+). All models were induced based on the learning set only, and no information from the test set was used for constructing the final models. As we are interested in comparing fully-specified classifiers, not learning algorithms, we did not use any data resampling approach such as cross-validation.² We are not interested in finding an overall best model with optimally calibrated parameters that could generalise well to future test cases. Indeed, we intentionally include models (e.g. the flip of a coin) that will certainly be unsuitable to predict test cases well. Our aim is solely to compare our conclusions based on the significance test and the CIs (Figure 3).

We used the following five models.

Model #1: We used diagonal linear discriminant analysis (*dlda*) as the learning algorithm, which assigns a sample \mathbf{x} to the class c (here, either ER- and ER+) that minimises the sum in Equation (6), where p is the number of genes, x_i is the value of the i -th gene, \bar{x}_{ci} is the mean of all values of i -th gene in class c , and s_i^2 is the pooled estimate of the variance of the i -th gene.

$$dlda(\mathbf{x}, c) = \sum_{i=1}^p \frac{(x_i - \bar{x}_{ci})^2}{s_i^2} \quad (6)$$

Then, using Welch's t -test, we selected $n = 30$ top-scoring genes, i.e. those with the smallest p -values for the discrimination between ER- and ER+ in the learning set. Using these 30 features, we built model #1, which was then applied to the test set.

Model #2: This model was built analogously to model #1, except that we used only the top 10 discriminatory genes.

Model #3: We deliberately corrupted the learning set as follows. We randomly selected five ER+ and five ER- cases from the learning set and swapped their class label (i.e. an ER+ case was mislabeled as ER- and vice versa). Thus, the conditional distribution of the variables in the learning and the test set is no longer the same. Then, we selected the 30 most discriminatory genes from the learning set to build a *dlda* model. The rationale for this deliberate data corruption was that we wished to include a model that must be truly inferior to the competing models #1 and #2 on the entire data distribution.

Model #4: This model was a fair coin, i.e. a model that predicts the class label of each test case as either ER+ or ER- with a probability of 0.5.

Model #5: This model used only the class prior information from the learning set to predict the test cases; no covariate information was used. As the learning set contains 69 cases of which 20 are ER-, the model predicted $20/69 \times 50 = 14$ test cases as ER- and 36 test cases as ER+.

The rationale for using these models was the following. The learning algorithm of *dlda* is deterministic. Unlike, for example, neural network algorithms, *dlda* does not include any random element; thus, any observed difference must be due to the different data inputs.

Table 2. Ten pair-wise comparisons (ranked in decreasing order based on X^2).

Comparison	b	c	X^2	α	χ^2	p -value	exact p -value
(1, 4)	3	21	12.0	0.005	7.9	5.3×10^{-4}	2.8×10^{-4}
(1, 3)	1	15	10.6	0.006	7.7	1.1×10^{-3}	5.2×10^{-4}
(2, 4)	4	20	9.4	0.006	7.5	2.2×10^{-3}	1.5×10^{-3}
(1, 5)	4	19	8.5	0.007	7.2	3.6×10^{-3}	2.6×10^{-3}
(2, 3)	2	14	7.6	0.008	7.0	5.8×10^{-3}	4.2×10^{-3}
(2, 5)	4	17	6.9	0.010	6.6	8.6×10^{-3}	7.2×10^{-3}
(4, 5)	13	10	0.7	0.013	6.2	0.40	0.68
(1, 2)	0	2	0.5	0.017	5.7	0.48	0.50
(3, 4)	11	15	0.3	0.025	5.0	0.58	0.56
(3, 5)	14	15	0.0	0.05	3.8	>0.99	>0.99

Notes: Tuples (i, j) denote pair-wise comparisons; b denotes the errors made by model i but not by model j ; c denotes the errors made by model j but not by model i ; X^2 is McNemar's test statistic; α is the Holm-adjusted significance level and χ^2 is the corresponding critical value.

each time with a decrement of n by 1. This approach controls the family-wise error rate at level $\alpha = 0.05$.

8. Results

Table 2 summarises the 10 pair-wise comparisons between the i -th and the j -th model, ranked in decreasing order based on X^2 . Unsurprisingly, we observed the largest difference between model #1 (*dlda* using 30 top-ranking genes) and #4 (fair coin). As expected, we observed a significant difference in performance for (1, 3), (2, 4), (1, 5), (2, 3) and (2, 5). The adjustments for multiple testing did not change our conclusions. For completeness, we also included the p -values that result from the exact binomial version of McNemar's test. The conclusions about significant differences remain the same, though.

Figure 4 shows a graphical representation of the CIs for all pair-wise differences. The adjustment for multiple testing increases the width of the intervals, but it does not affect the inclusion or exclusion of the null value. Therefore, the conclusions about the significance are the same with or without adjustment.

Consider the adjusted interval for the difference of error rates between *dlda* using 30 top-ranking genes and the fair coin, i.e. comparison (1, 4) in Figure 4. The point estimate for the difference is 0.31, and the adjusted 95%-CI is [0.10, 0.54]. This means that the true difference in performance for the entire population (from which the test set is only a sample) is likely to be within these bounds,³ so we can expect that *dlda* will make at least 10% and at most 54% fewer errors than the fair coin on similar data. In addition, as the null value is outside of the interval, we can readily see that this difference is significant. The small p -value for the difference between models #1 (*dlda* using 30 top-ranking genes) and #4 (fair coin) could invite us to highlight this result and pay little (if any) attention to other, non-significant results. But, despite its low p -value of 5.3×10^{-4} (here, the 'most significant'), the result for the comparison (1, 4) is imprecise, as shown in the wide CI. For most applications, and certainly for the development of genomic classifiers for a clinical application, this measurement is perhaps *too* imprecise to be of any practical use.

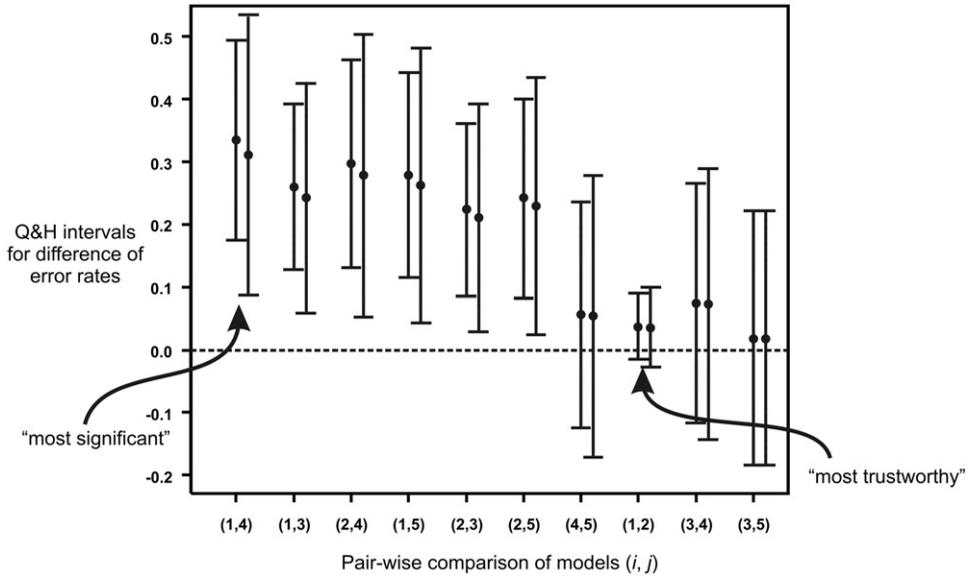


Figure 4. 95%-CIs based on Quesenberry and Hurst for differences in error rate on the independent test set ($n=50$). For each pair-wise comparison, two intervals are shown. The narrower interval is the unadjusted 95%-CI; the wider interval results from Holm’s adjustment. The horizontal line at $y=0$ indicates the null value.

Consider now the comparison between *dlda* using 30 and 10 top-ranking genes, i.e. comparison (1, 2) in Figure 4. The point estimate for the difference is 0.04, and the 95%-CI is $[-0.03, 0.10]$. As the null value is inside of the interval, we cannot reject the null hypothesis of equal performance between the two classifiers. But despite its non-significance, the result from (1, 2) is the most trustworthy. Model #1 can be expected to have a lower error rate (of about 10% at most) than model #2, and we need to assess whether the additional costs of including 20 marker genes justify this improvement. Here, we could argue that the improvement is only marginal and does not outweigh the additional costs of including more biomarkers.

Now let us compare this information with what the p -value of 0.48 (Table 2) tells us, besides the obvious result that the difference is not significant. Assuming that the two models indeed have the same performance, the probability of observing a result as extreme as or more extreme than the actually observed one is 0.48. This means that the value includes results that were not observed at all – the results that are as extreme as or more extreme than the observed ones, $(b, c) = (0, 2)$, are $(b, c) = (2, 0)$, $(b, c) = (0, 1)$, $(b, c) = (1, 1)$ and $(b, c) = (0, 0)$. These events did not happen, yet they enter the calculation of the p -value. Furthermore, the value 0.48 does not tell us anything about the expected size of the difference in performance for similar data sets.

Small p -values may give rise to the illusion that chance plays a small role. However, the estimates that are least affected by chance are those estimates with narrow CIs, not small p -values (Poole 2001). This is indeed confirmed in our experiments: we observe the lowest p -value for the comparison between a ‘real’ classifier (i.e. a model built using an established algorithm and real predictors) and a random guesser. However, we observe the narrowest CI for the models derived from *dlda* with the top-30 and top-10 genes.

Thus, in the words of Poole, the ‘most durable’ results are the results with the narrowest CIs, and these results – not the results with the smallest p -value – deserve our attention because they are the most statistically stable (Poole 2001).

Let us now pretend that we do not know the true ‘nature’ of the models. Imagine that a researcher developed a novel model and compared it to state-of-the-art classifiers. Let this novel model be model #2. The significance test gives a non-significant result for the comparison with model #1. Thus, the researcher’s paper presenting model #2 is bound to be rejected, with comments on the non-significant, even inferior performance, compared to the state-of-the-art classifier #1. But now, contrast this decision with the reasoning based on the CI. This decision would certainly be a different one.

9. Discussion and conclusions

In machine learning, comparative classification studies have emphasised the role of significance testing and p -values (Dietterich 1998; Nadeau and Bengio 2003; Bouckaert and Frank 2004; Demšar 2006; Garcia and Herrera 2008). However, CIs for effect size estimation seem to play a minor role. In fact, their widespread use even appears to be declining (Demšar 2006). We consider this a rather unfortunate development. In fact, a CI is always more informative than a statistical test: the interval provides a measure of the effect size (such as the magnitude of the difference in performance) *and* a measure of its uncertainty (Cummings 2012). In contrast, a test lumps these (intrinsically different) measures together in the p -value.

It is, therefore, perhaps surprising that significance tests enjoy a great deal of popularity in machine learning. Why is that the case? Artificial intelligence and machine learning are sciences, and consequently, they are committed to the scientific method (Drummond and Japkowicz 2010). Null hypothesis significance testing might seem to contribute to scientific objectivity, which might explain why it is becoming increasingly entrenched in artificial intelligence and machine learning.

We may also speculate that the popularity is partly due to the increasingly interdisciplinary nature of modern research practice. The life sciences, for example, draw mainly upon the expertise of scientists from biology, statistics and computer science. Genomics posed new analytical challenges that offered fertile grounds for the development of new concepts and algorithms, and machine learning has made a tremendous contribution. It may, therefore, not be a surprise that the accompanying website⁴ for random forests (Breiman 2001), for example, lists one case study on microarray analysis and another on DNA analysis. We also find examples of microarray analysis in standard textbooks on machine learning (Hastie, Tibshirani, and Friedman 2008). In the life sciences, however, significance testing is still deeply entrenched (Goodman 2008), despite efforts to ban its use (Rothman 1978; International Committee of Medical Journal Editors 1997). Are still too many reviewers and publishers of academic articles demanding significance testing? Perhaps.

This article focused on the problems of null hypothesis significance testing in the current practice of evaluating machine learning algorithms. However, we need to stress that CIs for effect sizes are no ‘silver bullets’ because they, too, can lead to misinterpretations (Levin and Robinson 1999). While this article takes a supremely positive stance towards CIs in general, we also need to acknowledge the shortcomings of frequentist CIs. For example, under specific circumstances, it is possible to obtain a 95%-

CI with a probability of zero of containing the population parameter of interest (Robinson 1978; Leslie 2008).⁵ Also, we remember that a 95% frequentist CI is not the same as a 95% Bayesian *credibility* interval. If we construct 100 frequentist intervals from different random samples of test sets from the entire population, then 95 are expected to include the true difference in performance. But a single frequentist CI does not give the probability that the true difference is included. In contrast, a Bayesian interval does allow such an interpretation.

Also, the statistical evaluation is only *one* aspect of the evaluation process. Drummond (2006) mentioned two further aspects, which we did not elaborate on here: the adequacy of the performance measure and the benchmark data sets being used. Indeed, as Hand (2006) confirmed, there is all too often a discrepancy between the measure used to choose and evaluate a model and the measure that we actually care about. The use of benchmark data sets, such as the data sets from the UCI repository, is also problematic. How representative are these data sets for the data sets actually arising in reality? Arguably, not very much (Drummond 2006). We also need to remember that the performance of a classifier depends on both its own idiosyncrasies and the complexity of the data set. The true complexity of real-world data sets, however, is nearly impossible to characterise because of uncertainties such as sampling sparsity, missing values, etc. Thus, we may argue that the usefulness of real data for benchmarking purposes is limited.

Here, we focused on CIs for the effect size. However, CIs can also be constructed to accompany graphical representations of performance (Dugas and Gadoury 2010), i.e. confidence bands around ROC and cost curves. Such curves are always more meaningful than single scalar values because they paint a more complete picture of the performance. We wish to caution that any evaluation on the basis of a single scalar measure (be it error rate, accuracy, area under the ROC curve or any other) cannot do justice to a classifier's true performance. ROC analysis and other cost-sensitive evaluations are certainly better. But then, we also need to resist the temptation of reducing such representations again to scalar measures, for example, by reducing a ROC curve to its AUC.

Methodologically, we did not present anything new in this study. The CIs for the effect size – that is, the difference between the true error rates of two classifiers – are based on the method by Quesenberry and Hurst (1964) and May and Johnson (1997). The novelty of our work is that we offer an alternative to the current practice of performance evaluation and show how different our reasoning can be, depending on the inferential tools that we use. Reasoning based on CIs requires more intellectual effort: the researcher needs to argue why an effect is worth being taken seriously or not. This reasoning needs to take into account the specific context, too. Although the literature is replete with comparative classification studies, the domains rarely enter the discussion about differences in performance. In contrast, classifiers are generally compared on benchmark data sets from very different domains (Frank and Asuncion 2010), and a novel classifier may be highlighted if it significantly outperforms a number of competing classifiers on most of these benchmark sets. But in one domain, a significant result may not deserve our attention, whereas in another domain, a non-significant result may indeed deserve it. Only CIs, not p -values, can support this reasoning.

We invite the readers and notably the reviewers of scientific articles to compare how their own reasoning changes, depending on whether they focus on p -values or CIs. Our study is hoped to further raise the awareness of this issue. We consider the shift from

significance tests to CIs as an essential ingredient of a possible remedy to what Drummond and Japkowicz (2010) have called ‘a harmful addiction in performance evaluation’.

Acknowledgements

We thank the anonymous referees who clearly spent a lot of time and effort for their detailed reviews, which helped us improve our manuscript. Jose A. Lozano is supported by Saiotek and Research Groups 2007–2012 (IT-242-07) programmes (Basque Government) and TIN2010-14931 (Spanish Ministry of Science and Innovation).

Notes

1. <http://www.site.uottawa.ca/ICML08WS>
2. From the observed performances of the fully specified classifiers, we may of course draw conclusions about the suitability of the learning algorithms for the concrete task at hand.
3. More precisely, if we construct 100 such intervals from different random samples of test sets from the entire population, then 95 are expected to include the true difference in performance. A single (frequentist) CI does not give the probability that the true difference is included.
4. http://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm
5. We thank the anonymous reviewers for reminding us about this shortcoming.

References

- Berger, J.O., and Berry, D.A. (1988), ‘Statistical Analysis and the Illusion of Objectivity’, *American Scientist*, 76, 159–165.
- Berrar, D., Bradbury, I., and Dubitzky, W. (2006), ‘Avoiding Model Selection Bias in Small-sample Genomic Data Sets’, *Bioinformatics*, 22, 1245–1250.
- Bouckaert, R.R., and Frank, E. (2004), ‘Evaluating the Replicability of Significance Tests for Comparing Learning Algorithms’, *Advances in Knowledge Discovery and Data Mining*, 3056, 3–12.
- Breiman, L. (2001), ‘Random Forests’, *Machine Learning*, 45, 5–32.
- Cawley, G.C., and Talbot, N.L.C. (2010), ‘On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation’, *Journal of Machine Learning Research*, 11, 2079–2107.
- Cummings, G. (2012), *Understanding the New Statistics: Effect Sizes, Confidence Intervals, and Meta-analysis*, New York/London: Routledge, Taylor & Francis Group.
- Demšar, J. (2006), ‘Statistical Comparisons of Classifiers Over Multiple Data Sets’, *Journal of Machine Learning Research*, 7, 1–30.
- Demšar, J. (2008), ‘On the Appropriateness of Statistical Tests in Machine Learning’, in *Proceedings of ICML 2008 Workshop on Evaluation Methods for Machine Learning II*, Helsinki, Finland, 5–9 July 2008.
- Denis, D. (2003), ‘Alternatives to Null Hypothesis Significance Testing’, *Theory & Science*, 4(1). Available online at http://theoryandscience.icaap.org/content/vol4.1/02_denis.html. Accessed 19 April 2012.
- Dietterich, T.G. (1998), ‘Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms’, *Neural Computation*, 10, 31–36.
- Dixon, P. (1998), ‘Why Scientists Value p Values’, *Psychonomic Bulletin & Review*, 5, 390–396.
- Drummond, C. (2006), ‘Machine Learning as an Experimental Science, Revisited’, in *Proceedings of the 21st National Conference on Artificial Intelligence: Workshop on Evaluation Methods for Machine Learning*, AAAI Press Technical Report WS-06-06, pp. 1–5.

- Drummond, C. (2008), 'Finding a Balance Between Anarchy and Orthodoxy', in *Proceedings of ICML 2008 Workshop on Evaluation Methods for Machine Learning II*, Helsinki, Finland, 5–9 July 2008.
- Drummond, C., and Japkowicz, N. (2010), 'Warning: Statistical Benchmarking is Addictive. Kicking the Habit in Machine Learning', *Journal of Experimental and Theoretical Artificial Intelligence*, 2, 67–80.
- Dugas, C., and Gadoury, D. (2010), 'Pointwise Exact Bootstrap Distributions of ROC Curves', *Machine Learning*, 78, 103–136.
- Fraley, R.C., and Marks, M.J. (2007), 'The Null Hypothesis Significance Testing Debate and its Implications for Personality Research', in *Handbook of Research Methods in Personality Psychology*, eds. R.W. Robins, R.C. Fraley and R.F. Krueger, New York: Guilford, pp. 149–169.
- Frank, A., and Asuncion, A. (2010), 'Uci Machine Learning Repository', URL <http://archive.ics.uci.edu/ml>.
- Garcia, S., and Herrera, F. (2008), 'An Extension on Statistical Comparisons of Classifiers Over Multiple Data Sets for All Pairwise Comparisons', *Journal of Machine Learning Research*, 9, 2677–2694.
- Goodman, S. (1993), '*P* Values, Hypothesis Tests, and Likelihood: Implications for Epidemiology of a Neglected Historical Debate', *American Journal of Epidemiology*, 137, 485–496.
- Goodman, S. (1999), 'Toward Evidence-based Medical Statistics. 1: The *p* Value Fallacy', *Annals of Internal Medicine*, 130, 995–1004.
- Goodman, S. (2008), 'A Dirty Dozen: Twelve *p*-Value Misconceptions', *Seminars in Hematology*, 45, 135–140.
- Hand, D. (2006), 'Classifier Technology and the Illusion of Progress', *Statistical Science*, 21, 1–14.
- Hastie, T., Tibshirani, R., and Friedman, J. (2008), *The Elements of Statistical Learning* (2nd ed.), New York, Berlin, Heidelberg: Springer.
- Holland, B. (1991), 'On the Application of Three Modified Bonferroni Procedures to Pairwise Multiple Comparisons in Balanced Repeated Measures Designs', *Computational Statistics Quarterly*, 6, 219–231.
- Holm, S. (1979), 'A Simple Sequentially Rejective Multiple Test Procedure', *Scandinavian Journal of Statistics*, 6, 65–70.
- International Committee of Medical Journal Editors (1997), 'Uniform Requirements for Manuscripts Submitted to Biomedical Journals', *New England Journal of Medicine*, 336, 309–315.
- Johnson, D.H. (1999), 'The Insignificance of Statistical Significance Testing', *Journal of Wildlife Management*, 63, 763–772.
- Kibler, D., and Langley, P. (1988), 'Machine Learning as an Experimental Science', in *Proceedings of the 7th International Conference on Machine Learning*, pp. 1207–1211.
- Langley, P. (2011), 'Machine Learning As an Experimental Science', *Machine Learning*, 82, 275–279.
- Leslie, C. (2008), 'Exhaustive Conditional Inference: Improving the Evidential Value of a Statistical Test by Identifying the Most Relevant *p*-value and Error Probabilities', PhD Thesis, Australia, University of Melbourne.
- Levin, J.R., and Robinson, D.H. (1999), 'Further Reflections on Hypothesis Testing and Editorial Policy for Primary Research Journals', *Educational Psychology Review*, 11(2), 143–155.
- Manly, K.F., Nettleton, D., and Hwang, J.T. (2004), 'Genomics, Prior Probability, and Statistical Tests of Multiple Hypotheses', *Genome Research*, 14, 997–1001.
- May, W.L., and Johnson, W.D. (1997), 'Confidence Intervals for Differences in Correlated Binary Proportions', *Statistics in Medicine*, 16, 2127–2136.
- McNemar, Q. (1947), 'Note on the Sampling Error of the Difference Between Correlated Proportions or Percentages', *Psychometrika*, 12, 153–157.

- Mulaik, S.A., Raju, N.S., and Harshman, R.A. (2007), ‘There Is a Time and a Place for Significance Testing’, in *What If There Were No Significance Tests?*, eds. L.L. Harlow, S.A. Mulaik and J.H. Steiger, New Jersey (USA): Lawrence Erlbaum Associates, pp. 65–115.
- Nadeau, C., and Bengio, Y. (2003), ‘Inference for the Generalization Error’, *Machine Learning*, 52, 239–281.
- Nix, T.W., and Barnette, J.J. (1998), ‘The Data Analysis Dilemma: Ban or Abandon. A Review of Null Hypothesis Significance Testing’, *Research in the Schools*, 5, 3–14.
- Ojala, M., and Garriga, G.C. (2010), ‘Permutation Tests for Studying Classifier Performance’, *Journal of Machine Learning Research*, 11, 1833–1863.
- Poole, C. (2001), ‘Low p -values or Narrow Confidence Intervals: Which Are More Durable?’, *Epidemiology*, 12, 291–294.
- Quesenberry, C.P., and Hurst, D.C. (1964), ‘Large Sample Simultaneous Confidence Intervals for Multinomial Proportions’, *Technometrics*, 6, 191–195.
- R Development Core Team (2009), *R: A Language and Environment for Statistical Computing*, Vienna, Austria: R Foundation for Statistical Computing.
- Robinson, G.K. (1978), ‘On the Necessity of Bayesian Inference and the Construction of Measures of Nearness to Bayesian Form’, *Biometrika*, 65, 49–52.
- Rothman, J. (1978), ‘A Show of Confidence’, *New England Journal of Medicine*, 299, 1362–1363.
- Schmidt, F.L. (1996), ‘Statistical Significance Testing and Cumulative Knowledge in Psychology: Implications for Training of Researchers’, *Psychological Methods*, 1, 115–129.
- Sheskin, D.J. (2007), *Handbook of Parametric and Nonparametric Statistical Procedures*, New York: Chapman and Hall, CRC.
- Sotiriou, C., Wirapati, P., Loi, S., Harris, A., Fox, S., Smeds, J., Nordgren, H., Farmer, P., Praz, V., HaibeKains, B., Desmedt, C., Larsimont, D., Cardoso, F., Peterse, H., Nuyten, D., Buyse, M., van de Vijver, M.J., Bergh, J., Piccart, M., and Delorenzi, M. (2006), ‘Gene Expression Profiling in Breast Cancer: Understanding the Molecular Basis of HistoLogic Grade to Improve Prognosis’, *Journal of the National Cancer Institute*, 98, 262–272.

Appendix: Quesenberry and Hurst CI for the effect size

The effect size is the difference $\delta_\tau = |\tau_A - \tau_B|$ between the true error rates τ_A and τ_B of two models A and B . To derive a CI for δ_τ , we solve the following inequality for δ_τ as suggested by May and Johnson (1997).

$$\frac{n(|p_b - p_c| - \delta_\tau)^2}{(\pi_b + \pi_c) - \delta_\tau^2} \leq \chi_{1-\alpha}^2 \tag{A1}$$

Here, $\chi_{1-\alpha}^2$ is the critical value of the χ^2 -distribution with one degree of freedom. Using again p_b and p_c as estimates for π_b and π_c , respectively, we can rewrite the inequality as

$$\begin{aligned} n(|p_b - p_c| - \delta_\tau)^2 &\leq \chi_{1-\alpha}^2 [(p_b + p_c) - \delta_\tau^2] \\ \Leftrightarrow (n + \chi_{1-\alpha}^2) \delta_\tau^2 - (2n|p_b - p_c|) \delta_\tau &\leq \chi_{1-\alpha}^2 (p_b + p_c) - n(p_b - p_c)^2 \\ \Leftrightarrow \delta_\tau^2 - \frac{2n|p_b - p_c|}{n + \chi_{1-\alpha}^2} \delta_\tau + \left[\frac{n|p_b - p_c|}{n + \chi_{1-\alpha}^2} \right]^2 &\leq \frac{\chi_{1-\alpha}^2 (p_b + p_c) - n(p_b - p_c)^2}{n + \chi_{1-\alpha}^2} + \left[\frac{n|p_b - p_c|}{n + \chi_{1-\alpha}^2} \right]^2 \\ \Leftrightarrow \left(\delta_\tau - \frac{n|p_b - p_c|}{n + \chi_{1-\alpha}^2} \right)^2 &\leq \frac{\chi_{1-\alpha}^2 (p_b + p_c)(n + \chi_{1-\alpha}^2) - n(p_b - p_c)^2 (n + \chi_{1-\alpha}^2) + n^2 (p_b - p_c)^2}{(n + \chi_{1-\alpha}^2)^2} \\ \Leftrightarrow \left| \delta_\tau - \frac{n|p_b - p_c|}{n + \chi_{1-\alpha}^2} \right| &\leq \frac{\sqrt{\chi_{1-\alpha}^2 (p_b + p_c)(n + \chi_{1-\alpha}^2) - n\chi_{1-\alpha}^2 (p_b - p_c)^2}}{n + \chi_{1-\alpha}^2} \end{aligned}$$

Hence, as LB and UB of δ_τ , we obtain δ_{LB} and δ_{UB} , respectively.

$$\delta_{LB} = \frac{n|p_b - p_c|}{n + \chi_{1-\alpha}^2} - \frac{\sqrt{\chi_{1-\alpha}^2[(p_b + p_c)(n + \chi_{1-\alpha}^2) - n(p_b - p_c)^2]}}{n + \chi_{1-\alpha}^2}, \quad (\text{A2})$$

$$\delta_{UB} = \frac{n|p_b - p_c|}{n + \chi_{1-\alpha}^2} + \frac{\sqrt{\chi_{1-\alpha}^2[(p_b + p_c)(n + \chi_{1-\alpha}^2) - n(p_b - p_c)^2]}}{n + \chi_{1-\alpha}^2}. \quad (\text{A3})$$

Copyright of Journal of Experimental & Theoretical Artificial Intelligence is the property of Taylor & Francis Ltd and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.