# Small Sample Statistics for Classification Error Rates II: Confidence Intervals and Significance Tests

J. Kent Martin and D. S. Hirschberg
(jmartin@ics.uci.edu) (dan@ics.uci.edu)
Department of Information and Computer Science
University of California, Irvine
Irvine, CA 92717
Technical Report No. 95-43
November 12, 1995

## Abstract

Several techniques for estimating the range of uncertainty of estimated error rates and for estimating the significance of observed differences in error rates are explored in this paper. Textbook formulas which assume a large test set (*i.e.*, a normal distribution) are commonly used to approximate the confidence limits of error rates or as an approximate significance test for comparing error rates. Expressions for determining more exact limits and significance levels for small samples are given here, and criteria are also given for determining when these more exact methods should be used. The assumed normal distribution gives a poor approximation to the confidence interval in most cases, but is usually useful for significance tests when the proper mean and variance expressions are used. A commonly used $\pm 2\sigma$ significance test uses an improper expression for $\sigma$, which is too low and leads to a high likelihood of Type I errors.

# 1 Introduction

There is a substantial body of literature on estimating classifier error rates, and a clear consensus that some type of resampling technique is necessary to obtain unbiased estimates. In the companion paper [13] we dealt with methods for estimating a classifier's accuracy and the bias and variance of the estimates obtained from various methods. In this paper, we deal with confidence intervals (*i.e.*, the range of likely values of a classifier's true error rate given an estimated value), and with significance tests for the difference in the estimated error rates of alternative classifiers for the same population. The thesis of both papers is that *"...the traditional machinery of statistical processes is wholly unsuited to the needs of practical research ...the elaborate mechanism built on the theory of infinitely large samples is not accurate enough for simple laboratory data. Only by systematically tackling small sample problems on their merits does it seem possible to apply accurate tests to practical data." — R. A. Fisher [7] (1925)*

Among the significant findings reported in this paper are: (1) that the traditional formulas for error rate confidence intervals commonly found in introductory statistics textbooks assume an asymptotically large sample and are not accurate enough for machine learning research (an alternative formula is given), (2) that textbook formulas for significance tests are generally accurate enough (a more exact formula is also given), provided that the proper expression for the variance of an estimated difference in error rates is used, (3) that there are many pitfalls in estimating the variance of a difference in error rates, leading to common mistakes in significance testing, and (4) that the common practice in machine learning research of estimating significance from observations on a single sample from a test population is neither rigorous nor reliable.

Throughout this paper, the terms *error* and *error rate* (meaning misclassification rate) will be used interchangeably. The term *bias* (rather than error) is used to refer to a systematic difference between an error rate estimate and the true error rate (non-zero average difference), the term *precision* is used to refer to the rms variability of such differences, and the terms variance or standard deviation to refer to the variability of a particular estimate. Also, the function $E(x)$ denotes the expected (mean) value of the random variable $x$ and $\Phi(Z)$ denotes the cumulative standard normal distribution (zero mean, unity variance).

## 1.1 Hypothesis Testing

In this section we provide a brief tutorial on the statistical inference issues relating to confidence intervals and significance tests, and on their common foundation, *statistical hypothesis testing.* We also give a short outline of the organization of the paper.

Given a classifier and an estimate of its error rate, the true error rate might be substantially higher or lower than the estimate. In view of this, the point estimate (single value) is of little utility unless its *reliability* is also somehow indicated. One way to do this is to give the precision or the standard deviation of the estimate's sampling distribution. Another way is to specify a *confidence interval*, a region which contains the relatively plausible values of the true error. When the sampling distribution is skewed (asymmetric), as is usually the case for error rates, a correctly defined confidence interval is more informative than the standard deviation.

Given two unbiased estimators, if one has a lower variance it has a greater *power* to discriminate between different classifiers and is the preferred estimator for that reason. If two unbiased estimators have equal power, the least computationally expensive method is preferred. An unbiased estimator may sometimes be less powerful than a biased estimator (if the bias is the same for all of the classifiers being compared and the biased estimator has lower variance than the unbiased estimator).

The reliability and power of the various estimators have received relatively little attention in the machine learning literature, as compared to the literature on estimating error rates. In the first paper [13], we were concerned with the applications of statistical inference for *estimation* — using sample characteristics to infer population characteristics. That is, inferring a classifier (a set of rules predicting the classifications of items in the population) and estimating its true error. The topics dealt with in this paper concern a different aspect of statistical inference: *hypothesis testing* — using sample information to answer questions about the population and the inferred classifier.

One such question is whether the classifier correctly predicts the classes. The various methods for estimating error can be thought of as alternative methods for assessing the truth of the hypothesis that the classifier's predictions are correct. If we knew or assumed that the population data were free of any measurement, observation, or labeling errors, then the occurrence of a single prediction error would serve to refute the hypothesis. If we know or can reasonably assume that the population

data are imperfect, as is typically the case, then a single prediction error is not sufficient to refute the hypothesis (it could be that the prediction is right and the data are wrong). In the latter circumstance, we must accept or reject the hypothesis based on an inference regarding the strength of the contradictory evidence relative to the reliability of our data.

Another hypothesis that we frequently wish to test is that the true error rates of two alternative classifiers are different (*i.e.*, that one classifier predicts more accurately than the other). This question is more conveniently posed as a test of the ***null hypothesis*** that the true error rates are equal. Again, typically we must accept or reject the hypothesis based on an inference regarding the strength of the contradictory evidence relative to the reliability of our data.

Thus, the ability to answer the following two questions is particularly important: (1) how reliable is our estimated error rate, *e.g.*, within what interval is the true error rate to be found with a 95% (or 99%) likelihood? and (2) given another classifier having a different estimated error rate, how confident can we be that its true error rate is different from that of the first classifier?

We deal with the first of these questions, confidence intervals, in Section 2, dealing separately with traditional, textbook methods in Section 2.1 and with more exact methods derived from Bayesian analysis in Section 2.2 (a brief tutorial on the Bayesian methods is provided in the Appendix). We deal with the second question, significance tests, in Section 3, presenting first, in Section 3.1, a common mistake which confuses the confidence level of a significance test with the confidence interval for an estimate. Section 3.2 presents more correct formulations: a traditional, textbook method and a more exact method derived from Bayesian analysis.

Section 4 presents an extended example, comparing nearest neighbor and three nearest neighbor classifiers, which also illustrates several pitfalls in designing and analyzing such experiments (notably, use of biased error rate estimates and use of improper expressions for the variance). Section 5 discusses the particular difficulties encountered in single-sample tests for significance and, continuing the extended example of Section 4, illustrates the tendency to over-estimate significance inherent in the common approaches to applying such tests. Section 5.1 summarizes experiments using CART-style decision trees which test the generality of the results in Sections 4 and 5.

A summary of the significant findings and recommended methods is given in Section 6.

## 2 Confidence Intervals

As we have said, a classifier's true error rate might be somewhat higher or lower than the estimated rate ($\epsilon$) obtained by observing the number of errors that occur when the classifier is tested on a random sample. We quantify this by specifying a confidence interval $(\tau_a, \tau_b)$ such that this interval is expected to contain the true error ($\tau$) with high likelihood (in at least 95% of our experiments, for instance). Typically, we also balance the risk on either side of the interval, so that

$$P\{\tau_a < \tau < \tau_b \mid \epsilon\} \approx 0.95 \quad P\{\tau \leq \tau_a \mid \epsilon\} \approx 0.025 \quad P\{\tau \geq \tau_b \mid \epsilon\} \approx 0.025$$

These equations can be solved only if we specify a probability relationship between the true error rate and our observations.

All of the confidence interval analyses given here assume that the number of errors is binomially distributed, *i.e.*, that the probability of $m$ misclassified items in a sample of size $N$ drawn at random from a distribution with a true error rate of $\tau$ is given by the binomial distribution:

$$P(m \mid \tau, N) = \frac{N!}{m!\,(N-m)!}\,\tau^m\,(1-\tau)^{N-m} \tag{1}$$

such that the expected number of errors is $E(m) = N\tau$ and its variance is $\mathrm{Var}(m) = N\tau(1-\tau)$.

Cross-validation methods are somewhat different from the simple random sampling scheme from which the binomial is derived. To test whether the binomial assumption is reasonable for cross-validation, 1,000 samples of size $N = 100$ leading to linear discriminant classifiers with virtually identical true error ($\tau = 0.0203 \pm 0.0001$) were accumulated by repeatedly simulating samples from a population having 0.02 inherent error[1] until 1,000 classifiers in the target range were obtained (many simulated samples led to classifiers with true error rates outside the target range, which were not included). The number of errors found in 10-fold cross-validation of these selected samples was compared to the frequencies expected for a binomial with $N = 100$ and $\tau = 0.0203$. The differences are small ($\chi^2 = 9.05$, with 7 degrees of freedom, which is not statistically significant).

[1]The inherent error is the true error of the hypothetical classifier which has the lowest error rate possible for the population. In these experiments there are two equally likely classes, each normally distributed on a single attribute, with the same standard deviation ($\sigma$) and with $2.053\sigma$ distance between the class means.

## 2.1 Textbook Confidence Limits

A commonly used expression for the approximate $(1-\alpha)$ confidence interval is

$$\epsilon \pm \left(\frac{0.5}{M} + t\, s\right) \qquad \text{where } s = \sqrt{\epsilon(1-\epsilon)/M} \tag{2}$$

($M$ is the test set size, $m$ the number of errors found in the test set, $\epsilon = m/M$ the estimated error rate, and $t$ is Student's $t$ for probability level $\alpha/2$ and $M-1$ degrees of freedom; for the 95% confidence level and $M > 20$, $t \approx 2$).

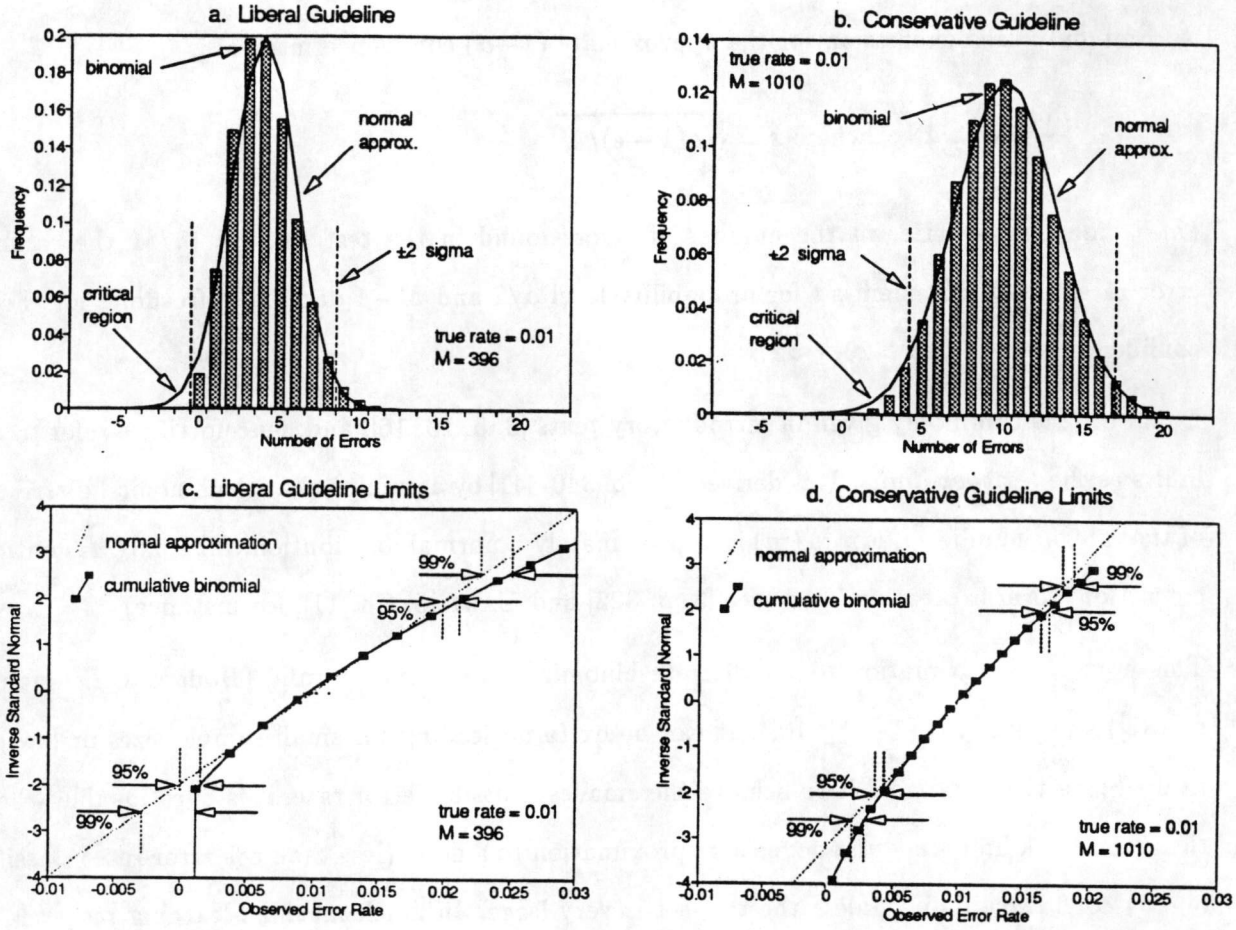
Equation 2 is commonly given in introductory texts [3, 5, 10, 16], and henceforth we refer to these limits as the *textbook limits*. It is derived [3, pp. 340-341] by assuming that the binomial distribution of the integer number of errors ($m$) is approximately a normal distribution. The $0.5/M$ 'continuity correction' term is often omitted (see [3, p. 322] and CART [5, Ch. 11], for instance).

This normal approximation to the discrete binomial distribution is valid (Hodges & Lehman [10, p. 187]) whenever $M\epsilon(1-\epsilon) \geq 10$, but can be quite misleading for small sample sizes or low error rates. Since the usual goal is to achieve the smallest possible error rate, it is questionable whether these textbook limits are an adequate approximation to a good (less than 5% error rate) classifier's $(1-\alpha)$ confidence limits unless the test set is very large. In much machine learning research, test set sizes of 200 or less are the rule, and this is certainly borderline regarding applicability of the normal approximation when the error rate is low. Breiman, *etal* [5, p. 308], for instance, report that noticeably more than 5% of the data fall outside $\pm 2\sqrt{\epsilon(1-\epsilon)/M}$ limits.

Somewhat different guidelines for the validity of the normal approximation are given by different authors — Anderson & Sclove [3, p. 322] give this criterion as $M\epsilon \geq 5$ and $M(1-\epsilon) \geq 5$, while Mendenhall, *etal* [14, p. 326] give the rule of thumb that the approximation is valid provided that $0 < \epsilon \pm 2\sqrt{\epsilon(1-\epsilon)/M} < 1$ (and also that, in estimating the probability of an error of $\epsilon$ or less, we should use the area under the normal curve below $\epsilon + 0.5/M$, *i.e.*, a continuity correction). The rule given by Hodges & Lehman is more conservative, and requires about twice the minimum sample size implied by Anderson's rule (2.5× Mendenhall's minimum).

We prefer the more conservative guidelines when estimating confidence intervals, because they are more precise in the crucial tail of the distribution. This is illustrated in Figure 1 — in Figures 1a

Figure 1: Liberal *vs.* Conservative Guidelines




and 1b, we see that the normal approximation is good in an overall sense under either rule, but better in the critical tail of the distribution under the more conservative rule. The $x$-axes ranges shown in Figure 1 include the absurdity of a negative number of errors or a negative error rate. Use of the normal approximation implies that such a thing is possible — in fact, under the liberal guidelines, that it has an appreciable (about 1 in 40) likelihood.

In Figures 1c and 1d, we illustrate the errors in the confidence intervals derived from the normal approximation under the different guidelines. (The $y$-axis in Figures 1c and 1d is $y = \Phi^{-1}(P(x))$, where $P(x)$ is the cumulative binomial distribution — under this transformation, the normal distribution appears as a straight line.) For the 95% limits, for instance, we show both the upper and lower limits, giving two positions for each. The right-hand position in each pair is correct, and the left-hand position is that derived from the normal approximation. The error in the limits is

Table 2: Empirical Tests of Confidence Intervals

| N | Percentage Outside Nominal 95% Limits Beta Distribution *vs* Textbook Limits | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | overall | | | $m = 0$ | | | $0 < m \leq M/2$ | | | $m > M/2$ | | |
| | Beta | | Text | Beta | | Text | Beta | | Text | Beta | | Text |
| | † | ‡ | book | † | ‡ | book | † | ‡ | book | † | ‡ | book |
| 10 | 5 | 6 | 14 | 1 | 2 | 29 | 7 | 5 | 2 | 14 | 24 | 5 |
| 20 | 6 | 6 | 10 | 0 | 4 | 20 | 8 | 5 | 3 | 13 | 13 | 13 |
| 30 | 7 | 6 | 12 | 1 | 2 | 30 | 8 | 5 | 3 | 11 | 20 | 9 |
| 50 | 6 | 5 | 14 | 3 | 3 | 42 | 6 | 5 | 3 | 15 | 15 | 11 |
| 100 | 6 | 7 | 10 | 1 | 13 | 39 | 7 | 5 | 3 | 10 | 10 | 9 |
| avg | 6 | 6 | 12 | 1 | 4 | 30 | 7 | 5 | 3 | 13 | 16 | 9 |

† from solving the Incomplete Beta Function
‡ using the normal approximation (Eqn. 4)

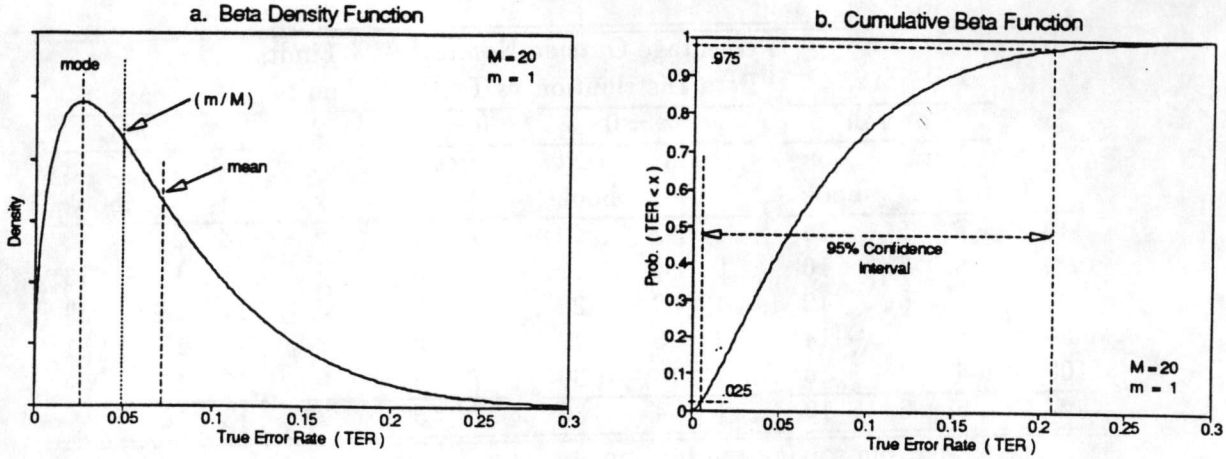
given by the difference between the pairs, and we can see that the error is non-symmetric (greater for the lower bound than for the upper), and much smaller and more nearly balanced under the conservative than under the liberal rule. Empirically, the textbook limits are a poor approximation, especially for low error rates (see Table 2, where overall 12% of the data fall outside the textbook nominal 95% confidence interval, and 30% of the data at $\epsilon = 0$ lie outside the textbook interval).

It is clearly inappropriate to use the textbook limits without first applying the textbook tests for determining whether they are applicable, yet this is commonly the case. Three likely causes for this are that many texts and handbooks omit or do not stress criteria for applicability, that methods for estimating confidence limits when the normal approximation is not valid are beyond the scope of introductory texts, so the user is given no alternative limits, and that the $\pm 2\sigma$ 95% confidence limits rule is ingrained and its underlying assumptions are rarely recalled or questioned.

## 2.2 More Exact Confidence Limits

The binomial distribution (Equation 1) expresses the probability, $P\{m \mid M, \tau\}$, of $m$ errors given the test set size, $M$, and the true error rate, $\tau$. Confidence intervals for $\tau$ require that one be able to answer such questions as how likely is it that $\tau$ is less than some particular value $x$, given $M$ and $m$? That is, what is $P\{\tau < x \mid M, m\}$, the *posterior distribution* of $\tau$?

Figure 3: The Jeffreys Beta Distribution




### 2.2.1 The Jeffreys Beta Distribution

Bayesian analysis (see, for instance, [8, pp. 76-78] and [9, pp. 18-33]) gives a Jeffreys Beta distribution, illustrated in Figure 3, as the posterior distribution of $\tau$:

$$P\{\tau < x \mid M, m\} = I(x, m+0.5, M-m+0.5) \equiv \int_0^x \mathrm{Be}(\tau, m+0.5, N-m+0.5)\, d\tau \qquad (3)$$

where $\mathrm{Be}(\tau, u, v)$ is the Beta probability distribution and $I(\tau, u, v)$ the Incomplete Beta function, with parameters[2] $u$ and $v$. For this distribution, the posterior mean ($\mu$) and variance ($\sigma^2$) of the true error are $\mu = (m+0.5)/(M+1)$ and $\sigma^2 = \mu(1-\mu)/(M+2)$, and the mode (most likely value) is

$$\text{mode} = \begin{cases} 0, & \text{if } m = 0 \\ (m-0.5)/(M-1), & \text{if } 0 < m < M \\ 1, & \text{if } m = M \end{cases}$$

Note the apparent paradox that, while the expected value of the estimated error rate $\epsilon = m/M$ is equal to the true error, $E(m/M \mid \tau) = \tau$, the expected value of the true error given $m/M$ is slightly different from $m/M$, $E(\tau \mid m, M) = (m+0.5)/(M+1)$. (See the Appendix for a discussion of this point.) The variance $\sigma^2$ is larger than $s^2 = \epsilon(1-\epsilon)/M$ for low error rates, $m/M < 0.1$, and for very high error rates, $m/M > 0.9$, and is less than $s^2$ for $0.15 < m/M < 0.85$ (provided that $M \geq 4$).

The $(1-\alpha)$ interval can be found by solving the Incomplete Beta function as shown in Figure 3b, though this can be very difficult in practice (see the Appendix). Empirical results for these exact Beta function limits compared to the textbook limits are summarized in Table 2. Overall, the Beta

[2] $u = v = 0.5$ is the Jeffreys prior, see the Appendix and the cited sources for information on these functions.

limits are a fairly good approximation to 95% confidence (6% of the data fall outside the limits, *vs.* 12% for the textbook limits). When the number of errors in the test set is zero, the Beta limits are too conservative (99% rather than 95%), but they are a great improvement over the textbook limits. For the bulk of the data ($0 < m \leq 50\%$) there is little to choose between the two sets of limits; they are about equally distant from 95%, too liberal for the Beta limits and too conservative for the textbook limits. At very high error rates, many points lie outside both sets of limits, reflecting the bias [13] of resampling estimates in that region.

An approximation to these Beta limits can be obtained by assuming that the Beta distribution is normal and using its known mean $\mu$ and variance $\sigma^2$. The resulting expression is similar in form to the textbook limits, but has a different mean and standard deviation:

$$\begin{aligned} \text{from Beta:} \quad & \mu \pm 1.96\,\sigma \quad \approx \quad \left(\epsilon + \frac{0.5}{M}\right) \pm 1.96\,\sigma \\ \text{textbook:} \quad & \epsilon \pm \left(\frac{0.5}{M} + t\,s\right) \qquad \text{(see Equation 2)} \end{aligned} \tag{4}$$

The differences between the two expressions are on the order of $1/M$, and lie in the sign of the $0.5/M$ term and in the magnitude of the standard deviation terms. The constant factor 1.96 (from the standard normal distribution) was used, rather than Student's $t$, because it provided a better fit to the empirical data. Table 2 also summarizes the results of applying these limits, which are an even better fit than was obtained from numerically solving the Incomplete Beta function.

### 2.2.2 Other Beta Distributions

Hartigan [8, pp. 48-50,76-78] gives a synopsis of arguments leading to $\text{Be}(\tau, 1/2, 1/2)$, the *Jeffreys prior* [11], as a good choice for finding confidence intervals from the binomial, which leads to Equation 3. (See the Appendix for the role of the prior distribution in Bayesian analysis.) The results of solving Equation 3 are shown in Table 2 in the previous section. While certainly an improvement over the textbook formula, these limits still only approximate the $(1-\alpha)$ interval (see [8, pp. 48-50] regarding limitations on the Jeffreys prior for two-sided confidence intervals). Considering the effort involved in solving Equation 3, it is fortunate that the normal approximation to the Incomplete Beta function used here (Equation 4) is adequate (and, in fact, gives a better fit to the empirical data than is furnished by the numerical solutions, see Table 2).

The *uniform prior*, $f(\tau) = 1$, is a special case ($f(\tau) = \mathrm{Be}(\tau, 1, 1)$) of the Beta priors, expressing complete ignorance as to $\tau$. The most important argument against the uniform prior is that the resulting posterior distribution does not fit the empirical data (both the mean and variance of the posterior are too large). A qualitative *a priori* argument against the uniform prior for classification problems takes note of the following facts:

- The true error rates being estimated are those of classifiers inferred from the sample. The inferred classifier always correctly predicts the class of some of the items in its training set. Those items are members of the population, therefore the error rate tested on the entire population cannot be 1.

- Populations involving actual measurements and observations, as opposed to hypothetical populations, always involve measurement, observation, and recording errors, and frequently have missing or inconsistent data. The inherent error of these real populations is unlikely to be zero. Even if it were zero, the inference method entails a language-intrinsic error[3] which is usually non-zero. Therefore, it is very unlikely that the true error rate will be zero.

- We know, *a priori*, that we can construct a classifier (always predict whichever class has the largest frequency in the sample) which makes minimal use of the sample data and has an expected† true error of less than 50% (typically less than 33%). Our *ex post*, more informed inference methods should certainly be able to do at least this well. Therefore, true error rates larger than 33% are relatively unlikely.

  † This can be derived from the multinomial distribution,

$$P\{N_1, \cdots, N_C \mid N, p_1, \cdots, p_C\} = N! \prod_{i=1}^{C} \left( p_i^{N_i} \; / \; N_i! \right)$$

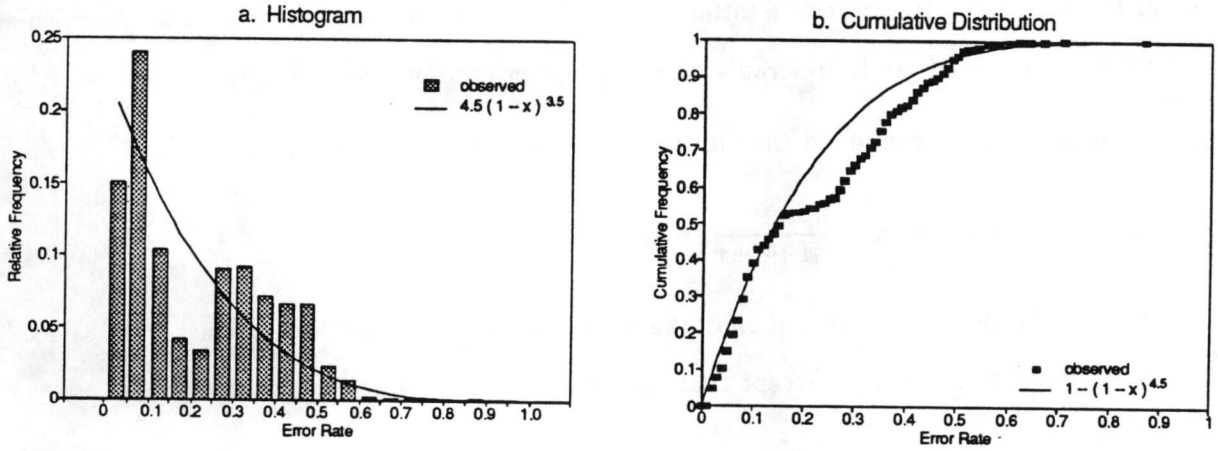
  (where $C$ is the number of classes in the population, $p_i$ is the fraction of class $i$ in the population, and $N_i$ is the number of class $i$ items in the sample) if we are given a suitable prior for the joint distribution of the $p_i$. It is more convenient to use Monte Carlo techniques, and some typical simulation results are shown in Table 4.

[3] If the correct classifier is a quadratic discriminant, then the linear discriminant which has the lowest possible true error can only approximate the correct classifier. Its error rate, though minimal for this kind of classifier, is greater than the inherent error. We term this hypothetical minimum true error for the chosen inference method the *language-intrinsic error*, denoting its dependence on the language used to represent a classifier.

Table 4: Mean Error of Largest Class Rule

| Sample Size $N$ | Average Error Rate Number of Classes $C = 2$ | $C = 3$ | $C = 4$ |
|---|---|---|---|
| 8 | .30 | .29 | .21 |
| 16 | .28 | .29 | .14 |
| 32 | .27 | .29 | .10 |
| 64 | .26 | .29 | .06 |
| 128 | .26 | .29 | .04 |

Figure 5: Empirical Error Rate Distribution



A non-uniform Beta distribution is consistent with these qualitative observations concerning the prior for $\tau$. Figure 5 shows some data on the relative frequency of various error rates compiled from a study [12] of 16 data sets from several problem areas. These empirical data are consistent with a Beta distribution with $u=1$ and $v=4.5$, which is also consistent with our qualitative observations regarding the prior for $\tau$. (For this Beta prior, $E(\tau)=18\%$, $\sigma_\tau=15\%$, the mode is zero, less than 5% of the error rates exceed 50%, and less than 20% exceed 33%. Also, the confidence interval for an observed error rate of $\epsilon=m/M<50\%$ is lower and narrower using this prior than those obtained using either the uniform or Jeffreys priors.)

The preceding discussions illustrate the important point that the confidence interval for an error rate estimate varies according to the prior distribution $f(\tau)$, that is, according to one's knowledge, beliefs, prejudices, or assumptions as to the likely values of $\tau$ prior to having inferred a classifier or estimated its error rate. (We note that the approach taken in arriving at the textbook limits implicitly assumes a uniform prior — the textbook limits differ from those of the uniform prior because of the additional assumptions of the normal approximation, *etc.*) The confidence interval also depends very strongly on the test set size $M$, such that the interval becomes narrower as $M$ increases and, importantly, such that the particular values assumed for $u$ and $v$ becomes less important as $M$ increases. For small samples, however, the influence of the assumed prior is very strong. Given the wide diversity possible even among the textbook methods, reported confidence intervals (or error bars) for error rate estimates are of little use unless the method used to calculate them is explicitly stated (and, preferably, the underlying assumptions, as well).

For small samples, we recommend the approximate limits from the Beta distribution

$$\mu \pm 1.96\,\sigma, \quad \text{where } \mu = \frac{m+u}{M+u+v} \text{ and } \sigma = \left(\frac{\mu(1-\mu)}{M+u+v+1}\right)^{0.5}$$

using the Jeffreys prior (we caution that the variety and range of the problems summarized in Figure 5 are much too sparse to accept that prior ($u=1$ and $v=4.5$) without reservation.)

## 3 Significance Tests

Having established the range of plausible values for the true error given an estimate, we now shift focus to the second question posed in the introduction: given alternative classifiers for the

population and estimates of their error, how confident can we be in asserting that one classifier predicts more accurately than the other?

The null hypothesis for comparing the estimated error rates of two inferred classifiers is the hypothesis that the true error rates of the classifiers are equal. The *level of significance* $\alpha$ is the probability, given that the null hypothesis is true, of obtaining the observed difference or a more extreme value (in a two-sided test, a value having greater magnitude).

The level of significance is commonly expressed as its converse, $(1-\alpha)$, the *confidence level.* If the confidence level is sufficiently high (typically 95% or 99%), we reject the null hypothesis and assert that the classifiers' true error rates are different. If the confidence level is lower than our critical value, we accept the null hypothesis and assert that the classifiers' true error rates are equal. There is a risk associated with either assertion:

- In a *Type I error* we reject the null hypothesis when it is true (we wrongly assert that the true error rates are different). $\alpha$ is a measure of the risk of making a Type I error.
- In a *Type II error* we accept the null hypothesis when it is false (we wrongly assert that the true error rates are equal). The risk of making a Type II error is neither $\alpha$ nor $(1-\alpha)$, because the assessment of $\alpha$ explicitly assumes the proposition being asserted (see [10, pp. 370-376] for a discussion of this point). The risk of making a Type II error is not usually assessed in significance testing.

In classifier inference problems, the null hypothesis asserts only that the two classifiers' true error rates are equal. In order to assess $\alpha$, it is necessary to specify either the value or the probability distribution of this common true error rate $\tau$ and, typically, neither of these is known. Any value obtained for $\alpha$ is thus conditional on whatever assumptions are made concerning $\tau$ and its distribution. All subsequent use of the symbol $\alpha$ should be understood to represent a conditional estimate of the true significance level.

### 3.1 A Common Mistake

A common mistake in testing whether two classifiers' true error rates differ is to check whether one estimated error rate is outside the confidence interval for the true error rate of the other estimate.

Comparing the higher estimate to the upper bound for the lower estimate (by analogy to CART's 1-SE rule [5, pp. 78-80], which was developed for a more specialized use) is logically inappropriate[4] for the following reasons:

1. It is a one-sided test — a one-sided test is appropriate only if it is *known* that one classifier has a lower true error than the other. If that were known, of course, there would be no point in making the comparison unless one were willing to accept a slightly higher error rate in exchange for, say, reduced complexity (which, to be fair to CART, is part of the context in which their 1-SE rule was proposed).

2. If the textbook confidence interval is used, this interval is too narrow for low error rates, which leads to a high likelihood of Type I error. The problem is particularly severe when the samples are small and the continuity correction is not used.

3. Even if the improved confidence interval given in the previous section is used, this is still not the proper formulation for this significance test, because the quantity being tested is the difference between the two estimates. The variance of the difference of two random variables is the sum of their variances less twice their covariance. A very unique relationship between the two estimates is implied by the 1-SE rule. While this relationship might be assumed to hold in 1-SE's narrow context (it seems a reasonable heuristic there), that is certainly not a valid assumption for all contexts.

In addition to the inappropriate analogy to the 1-SE rule, this problem might also arise from confusing the 95% *confidence interval* or *confidence limits* with the 95% *confidence level* for the difference in two estimates.

The 1-SE rule was developed in the narrow context of selecting which members of a set of trees (each of which is derived by differently pruning a larger tree) have error rates comparable to the candidate which appears to be best, and should be evaluated for their complexity. The error rates of this series of related trees are not independent, and it is difficult to know the distribution of the differences in estimated rates. Including a pruned tree in the set to be studied when its estimated

[4]While the null hypothesis can be accepted if the higher value falls within the confidence interval of the lower estimate, the hypothesis cannot be rejected with the appropriate degree of confidence should the higher value fall slightly outside the interval.

error rate is within 1 standard deviation of the lowest error rate found is a heuristic which should be judged empirically in the narrow domain for which it was intended. The difficulty arises when something like the 1-SE rule (specifically the +1.645-SE and ±2-SE rules), where $\mathrm{SE} = \sqrt{\epsilon(1-\epsilon)/M}$, is used as a one-sided or two-sided significance test.

### 3.2 A Textbook Significance Test

An approximate $(1-\alpha)$ confidence level significance test for the difference between two independent error rate estimates ($\epsilon_1 = m_1/M_1$ and $\epsilon_2 = m_2/M_2$) is given by (see [3, pp 412-415]): $\alpha \approx 2\Phi(Z)$, where $Z = -|\epsilon_1 - \epsilon_2|/s$. (Here, $s^2 = \tau_0(1-\tau_0)(1/M_1 + 1/M_2)$, where $\tau_0 = (m_1+m_2)/(M_1+M_2)$ is a weighted average error rate, $M_i$ is the test set size for estimate $i$, and $m_i$ is the number of errors found in test set $i$.)

When the null hypothesis is true, this textbook normal approximation is fairly good even when the underlying binomial distribution of the two estimates is far from normal. The approximation is poorer when the sample sizes are unequal and one of them is small.

The relative likelihood of various values of $(\epsilon_1 - \epsilon_2)$ under the null hypothesis can be calculated directly from the binomial probabilities for $(M_1, \tau_0)$ and $(M_2, \tau_0)$. This procedure provides a means for estimating $(1-\alpha)$ for small test sets, even when the normal approximation is not good. Empirical tests were run comparing confidence levels calculated in this way to those calculated from the textbook formula. Summary statistics for their differences are given in Table 6. The textbook solution is very close, even when the test set sizes are unequal and very small. Notably, this is true in the crucial region $.9 < (1-\alpha) < 1$, where the critical values are to be found in most applications.

The $(1-\alpha)$ estimates resulting from the textbook formula are fairly precise; but, *are they accurate?* It must be remembered that these values are *conditional* on the assumption that the true error rate is closely approximated by $\tau_0 = (m_1+m_2)/(M_1+M_2)$. It might be argued from the Law of Large Numbers (Law of Averages) that, under the null hypothesis, the true error rate is probably closer to $\tau_0$ than to either $(m_1/M_1)$ or $(m_2/M_2)$. It might also be argued based on our definition of the sample as being all the data available that, if the test set comprises the entire sample (as in either cross-validation or bootstrapping), this is the best estimate of the true error available. Neither of these arguments is valid, however, if the individual $(m_i/M_i)$ estimates are biased.

Table 6: Greatest Observed Difference of $(1-\alpha)$ Values

| Estimated $1-\alpha$ | Approximate − Exact † | | | | | |
|---|---|---|---|---|---|---|
| | Textbook Formula | | Unbiased Formula | | Δ Normals ‡ | |
| | Test Set Size | | Test Set Size | | Test Set Size | |
| | equal | unequal | equal | unequal | equal | unequal |
| $(1-\alpha)=0$ | 0 | 0 | 0 | 0 | 0 | 0 |
| $0<(1-\alpha)\leq.9$ | .019 | .051 | .012 | .042 | .080 | .073 |
| $.9<(1-\alpha)<1$ | .009 | .017 | .009 | .012 | .017 | .014 |
| $(1-\alpha)=1$ | .000 | .000 | .000 | .000 | .000 | .000 |

† Normal Approximation − Exact Binomial Calculation

‡ Difference in Normal Approximations, (Textbook − Unbiased)

Earlier (see Section 2.2), we noted that the expected value of the true error given an observation of $m_i$ errors in a test set of size $M_i$ is given approximately by $\mu_i = (m_i + 0.5)/(M_i + 1)$, indicating that the $(m_i/M_i)$ components lying behind the textbook formula are both biased. Using the less biased estimates $\mu_i$ instead of $\epsilon_i$ gives a different normal approximation: $\alpha = \Phi(Z)$,

where $Z = -\,|\,(\mu_1 - \mu_2) - \overline{\Delta\mu}\,|\,/\sigma$ $\qquad \sigma = \sqrt{\tau_*(1-\tau_*)/M_*}$

$\tau_* = (m_1+m_2+0.5)\,/\,(M_1+M_2+1)$ $\qquad M_* = (M_1+2)(M_1+2)\,/\,(M_1+M_2+4)$

and $\overline{\Delta\mu} = (M_1\tau_*+0.5)/(M_1+1) - (M_2\tau_*+0.5)/(M_2+1)$

is the expected (mean) value of $(\mu_1-\mu_2)$ (zero, if $M_1 = M_2$)

Again, the probability distribution of $(\mu_1-\mu_2)$ under the null hypothesis can be calculated directly from the binomial probabilities for $(M_1, \tau_*)$ and $(M_2, \tau_*)$. Values of $(1-\alpha)$ from the normal approximation above are compared to these more exact calculations in Table 6. This normal approximation is slightly more precise than that found earlier for the textbook approximation. Table 6 also compares the textbook estimates of $(1-\alpha)$ to those obtained from the unbiased normal approximation. The textbook estimates of $(1-\alpha)$ are high, but they are quite good considering how small these test sets are. There is, however, a slightly (about 1.5%) higher risk of Type I error when the textbook 95% formula is used than when the unbiased formula above is used.

If we keep in mind that these values are, at any rate, only approximate and that they are conditional on our assumptions that $m_i$ is binomially distributed and that the normalized difference $Z$ is approximately normal under the null hypothesis, then the textbook formula seems to be accurate

enough[5] for most purposes (in the crucial region $\alpha < .1$), provided that the significance levels are reported as being only approximate (strictly, we should report simply that we accept or reject the null hypothesis, based on an approximate test at the 0.05 level).

# 4 A Nearest-Neighbors Example

The approximate methods given in Section 3.2 are conditional on several assumptions in addition to normality and the null hypothesis: that the classifiers and their error estimates are independent (*i.e.*, inferred from independent samples) and that the estimates are binomially distributed. As pointed out in Section 3.1, these methods are not appropriate when the classifiers and error estimates are not independent (as in decision tree pruning). In this section we present a methodology which is appropriate whether or not the independence and binomial assumptions hold, by means of an extended example using nearest-neighbors classifiers.

In this and following sections, TER denotes a classifier's true error rate (the rate which would be observed were the classifier tested using the entire population), 1-NN or subscript 1 denotes a nearest neighbor classifier, and 3-NN or subscript 3 denotes a three nearest neighbors classifier.

Are the error rates of 3-NN really different from those of 1-NN and, if so, which method yields the more accurate classifiers? Under the null hypothesis, the statistic $t = \overline{\Delta}/s_{\overline{\Delta}}$ is distributed approximately as Student's $t$, where $\overline{\Delta} = \overline{\text{TER}_3} - \overline{\text{TER}_1}$ and $s_{\overline{\Delta}}$ is the estimated standard deviation of $\overline{\Delta}$. The method for estimating $s_{\overline{\Delta}}$ and the number of degrees of freedom (dof) of the appropriate Student's $t$ distribution depend on the experimental conditions (see [16, pp. 348-377], for instance).

In the simplest experiments, $\overline{\text{TER}_3}$ and $\overline{\text{TER}_1}$ are each based on observed error rates for classifiers inferred from $\eta$ random samples from the population, and there are $\eta - 1$ degrees of freedom. When two different, independent sets of samples are used to infer the 1-NN and 3-NN classifiers, then $s_{\overline{\Delta}} = \sqrt{(s_1^2 + s_3^2)/\eta}$, where $s_1$ and $s_3$ are the unbiased[6] standard deviations of the 1-NN and 3-NN TER's. When both a 1-NN and a 3-NN classifier are inferred from the same sample, a paired $t$-test is more appropriate, $s_{\overline{\Delta}} = s_\Delta/\sqrt{\eta}$, where $s_\Delta^2 = \sum_i (\Delta_i - \overline{\Delta})^2/(\eta - 1)$, and $\Delta_i = \text{TER}_{3,i} - \text{TER}_{1,i}$ is the observed difference for the $i^{th}$ of $\eta$ samples.

[5] That is, the textbook formula is *robust* ($\alpha$ is approximately correct, even under departures from assumptions).

[6] The unbiased standard deviation of $\eta$ observations of a random variable $x$ is given by $s_x = \sqrt{\sum_i (x_i - \bar{x})^2/(\eta - 1)}$.

Table 7: Paired $t$-test of TER's for 3-NN *vs.* 1-NN

| values of $t$, two-sided test with 19 dof<br>$t > 2.093$ is significant at the 95% level<br>$t < 0$ indicates 3-NN is the more accurate classifier | | | | | | | |
|---|---|---|---|---|---|---|---|
| Sample | Population Inherent Error Rate | | | | | | |
| Size | 0.1% | 1% | 2% | 5% | 10% | 25% | 40% |
| 10 | .9 | 2.3 | .5 | − .5 | 1.0 | 1.0 | .1 |
| 20 | 2.0 | −1.2 | − .2 | −1.0 | −1.7 | −2.9 | −3.0 |
| 30 | 3.0 | .2 | − .4 | −3.6 | −3.7 | −6.5 | − .8 |
| 50 | 3.1 | −1.7 | −2.1 | −3.0 | −2.9 | −5.4 | −6.2 |
| 100 | 3.2 | −1.0 | −3.2 | −4.4 | −9.0 | −7.3 | −3.5 |

In Table 7 we summarize the $t$-test results of a paired test simulating 20 samples each of several different sizes from populations having different inherent error rates (two equally likely classes, each normally distributed on a single attribute, with the same variance but different class means, see [13]). As a rule, the 3-NN classifiers are more accurate. However, this is not the case when the inherent error is very low (0.1%) or the sample size very small ($N = 10$). Examination of the data in Table 7 also suggests that there is no difference at all in the error rates of 1-NN and 3-NN classifiers for these populations for a sample size of about 12 or an inherent error near 0.5% (we also note that, in this particular case, an inherent error of 50% corresponds to there being no difference between the classes' distributions, so that both 1-NN and 3-NN have a TER of exactly 50% for all samples regardless of size — since both the mean difference and its variance are zero, the $t$-statistic's value is undefined).

One explanation for these observations takes into account the data density around the critical region where the classes' distributions overlap. The inherent error rate is the relative density (fraction of the population) in this region and the sample size reflects the overall data density. The product of the sample size and inherent error is the expected number of items in this region (the number in any particular sample is random, with a binomial distribution). When this expected number is low, especially when it is less than one, a sample will contain very little information from which to infer the placement of the inter-class boundaries.

The smoothing effect of 3-NN heavily discounts the apparent information imparted by relatively isolated instances and, in very sparse data, there is little information to spare — while 1-NN

overfits[7] the sample, 3-NN underfits when the data are very sparse. These observations on the effects of smoothing in nearest neighbors classifiers are consistent with Schaffer's [17] observations that pruning (smoothing) decision trees may actually be harmful when the data are very sparse relative to the concept to be learned.

A sample size greater than 30 seems necessary for consistent results regarding whether there is a statistically significant advantage for 3-NN over 1-NN (or *vice-versa*) measured over 20 samples. The differences are neither more nor less real for larger or smaller samples, but there is so much variation in the results for smaller samples that we have but little confidence in our measurement of the differences — we would need to average over a larger number of samples (*i.e.*, classifiers) in order to have the same confidence for smaller samples. While 1-NN classifiers appear to be more accurate than 3-NN for very small samples (10 or less), we cannot confidently reject the hypothesis that 1-NN and 3-NN are equally accurate for small samples.

The results summarized in Table 7 would belie any assertion that 3-NN is universally superior to 1-NN. A decision to use 3-NN rather than 1-NN reflects a bias, an *a priori* belief or assumption that 1-NN will overfit the sample data (and that 3-NN will not underfit). Whether our decision will result in a more accurate classifier depends on how appropriate this bias is to the problem at hand. If the sample size is small or the inherent error very low, 1-NN tends to overfit but 3-NN has a stronger tendency to underfit. Since we have shown [13] that the 1-NN classifier in this case is entirely equivalent to an unpruned CART-style decision tree handling continuous attributes in the usual way and that decision tree pruning and nearest neighbors smoothing have similar effects, we expect that these observations are equally applicable to decision tree pruning and other such questions as to the differences between inference methods. The choice of one inference method or algorithm over another is simply a choice (albeit many times a tacit or unawares choice) of one set of *a priori* assumptions or beliefs about the data over another set of assumptions. Probably the most crucial step in any statistical inference process is matching assumptions to the problem.

In most real-world situations the biases underlying nearest neighbors smoothing and decision tree stopping or pruning (namely that the inherent or language intrinsic error of the population is

[7]The apparent error can be made arbitrarily low by considering very complex, *ad hoc* classifiers. This is called *overfitting* [18], which is described by CART [5] as inferring classifiers that are larger than the information in the data warrant, and by ID3 [15] as increasing the classifier's complexity to accomodate a single noise-generated special case.

Table 8: Unpaired Calculations Mis-Applied to Paired Observations

values of $t$, two-sided test with 19 dof
$t > 2.093$ is significant at the 95% level
⊘ indicates that the difference is significant, but does not appear to be so here

| Sample Size | Population Inherent Error Rate | | | | | | |
|---|---|---|---|---|---|---|---|
| | 0.1% | 1% | 2% | 5% | 10% | 25% | 40% |
| 10 | .7 | ⊘ 1.7 | .5 | − .4 | .8 | .6 | .0 |
| 20 | 1.7 | − .9 | − .1 | − .4 | −1.3 | ⊘ −1.9 | ⊘ −1.1 |
| 30 | 2.6 | .2 | − .3 | −2.1 | −2.3 | −5.3 | − .7 |
| 50 | ⊘ 1.4 | −1.5 | ⊘ −1.2 | −2.8 | ⊘ −1.4 | −3.1 | −3.1 |
| 100 | ⊘ 1.8 | − .8 | −2.7 | −3.8 | −5.6 | −4.0 | −2.5 |

significantly greater than zero, that the data will contain mistakes and measurement errors in addition to the random sampling variation, and that inference methods that are not stopped, smoothed, or pruned will overfit) are almost certainly more appropriate than the naive counter-assumptions that classes do not overlap significantly and that reported data may be relied on as gospel. However, we should be aware that we are relying on these assumptions (and the assumption that smoothing or pruning does not underfit) which may have unanticipated consequences, as in the interaction of sample size and inherent error in Table 7.

How important are the nuances of calculating $s_{\overline{\Delta}}$? Statistical packages and recipe books typically either give only one method for a significance test on means or they give a large variety of methods that may bewilder novice users. For illustrative purposes, Table 8 shows the results of mis-applying the formula for unpaired observations to analyze the data from our paired experiments. All of the $t$ values in Table 8 are lower than their correct counterparts in Table 7, enough so that the wrong conclusion is reached as to the significance of 3-NN *versus* 1-NN in 7 cases out of 35. This is so because we have overestimated $s_{\overline{\Delta}}$. The assumptions underlying the unpaired $t$-test formula are inappropriate for the paired experiment at hand, and such errors will likely result if the method of data analysis is not properly matched to the experimental conditions.

The analysis in Table 7 is possible only because we have perfect knowledge of the populations and TER's. In general, this is not the case, and we have to base our analysis on one or another method for estimating error. In Table 9 we show results of the paired $t$-tests in Table 7 using several estimators (the sets of samples are identical, only the method of estimating the error changes —

Table 9: Paired $t$-test of Estimate Means for 3-NN *vs.* 1-NN

⊘ indicates that the TER's are significantly different, but the estimates do not appear to be
⊙ indicates that the TER's are not significantly different, but the estimates appear to be
‡ indicates that the sign is opposite to the sign of the mean difference in TER's

| Sample Size | Population Inherent Error Rate | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 0.1% | 1% | 2% | 5% | 10% | 25% | 40% | 50% |
| | LOO Estimates | | | | | | | |
| 10 | | ⊘ | | ‡ | | | ‡ | |
| 20 | ‡ | | ‡ | ‡ | | ⊘ | ⊘ | |
| 30 | ⊘ | | | ⊘ | ⊘ | ⊘ | | |
| 50 | ⊘ | | | ⊘ | ⊘ | | ⊘ | |
| 100 | ⊘‡ | ‡ | ⊘ | | | | | |
| | 10-CV Estimates | | | | | | | |
| 10 | | ⊘ | | ‡ | | | ‡ | |
| 20 | ‡ | | ‡ | ‡ | | ⊘ | ⊘ | |
| 30 | ⊘ | | | ⊘ | ⊘ | ⊘ | | ⊙ |
| 50 | ⊘ | | ⊘ | ⊘ | ⊘ | | ⊘ | |
| 100 | ⊘ | ‡ | ⊘ | | | ⊘ | | |
| | 632b Estimates | | | | | | | |
| 10 | ⊙ | | ⊙ | ⊙‡ | ⊙ | ⊙ | ⊙ | ⊙ |
| 20 | | ⊙‡ | ⊙‡ | ⊙‡ | ⊙‡ | ‡ | ‡ | ⊙ |
| 30 | | ⊙ | ⊙‡ | ‡ | ‡ | ‡ | ⊙‡ | ⊙ |
| 50 | ⊘ | ⊙‡ | ‡ | ‡ | ‡ | ‡ | ‡ | ⊙ |
| 100 | ⊘ | ⊙‡ | ‡ | ‡ | ‡ | ‡ | ‡ | ⊙ |
| | LOO* Estimates | | | | | | | |
| 10 | ⊙ | | ⊙ | ⊙‡ | | | ‡ | |
| 20 | | ‡ | ‡ | ‡ | | ⊘ | ⊘ | |
| 30 | | | ‡ | ⊘ | ⊘ | ⊘ | | |
| 50 | ⊘ | | ⊘ | ⊘ | ⊘ | | ⊘ | |
| 100 | ⊘‡ | ‡ | ⊘ | | | | | |

LOO denotes leave-one-out, 10-CV 10-fold cross-validation, 632b Efron's [6] 632 bootstrap, and LOO* Weiss' [19] hybrid estimate — see [13] for a review of these methods).

In 13 of our 40 experiments, the difference between the LOO estimates of 1-NN and 3-NN error rates is not significant at the 95% confidence level, though the difference in TER's is significant. The LOO estimates, though unbiased, are more variable than the TER's, and our $t$-test consequently less sensitive. The 10-CV results are similar, except that 10-CV is even more variable than LOO, and there are even more cases (15/40) where we cannot reject the null hypothesis at the 95% confidence level, even though the TER's are significantly different.

The LOO estimates are more variable than the TER's because LOO averages the error rates of $N$ classifiers, each slightly different from the reference classifier inferred using all of the sample, being inferred from one less instance. Let $\delta_i$ be the difference between TER and LOO for the $i^{th}$ sample, $\mathrm{LOO}_i = \mathrm{TER}_i + \delta_i$. LOO's variance is greater than the variance of TER by the mean square of the $\delta_i$. Similar considerations apply to 10-CV, except that the mean square $\delta$ is even larger than for LOO, because we average over fewer subsample classifiers, each differing even more from the reference than a LOO subsample (since 10% of the instances are omitted rather than a single instance), and because of randomness in selecting subsets.

These considerations also apply to all of the re-sampling estimators, with the important difference that bootstrapping and iterated cross-validation average over a very large number of subsample classifiers, which tends to reduce the variance to a level below that of LOO. For the biased estimators in this family, the difference in the variances of the estimator and TER is no longer simply the mean square $\delta$, and the estimator variance may even be lower than the variance of TER. For instance, the apparent error (APP, the rate observed when a classifier is tested on the same instances used to infer the classifier), which has zero variance for 1-NN.

Note the anomalous results for the 632b estimator. These high $t$-values and significance levels are not incorrect, but they are a potential pitfall for an unawares user. The 632b error rates of the 1-NN and 3-NN classifiers are different, with a very high degree of confidence, and the 632b rates for 1-NN are always better than those of 3-NN. This is because 632b is biased in both cases and has a different bias for 1-NN than for 3-NN. Knowing that the 632b estimates differ significantly tells us nothing about whether the TER's are different. This is the great danger inherent in using biased estimators — unless the magnitude and direction of the bias is known to be the same for the cases under study, or unless the magnitudes and directions are known and compensated for in each case, use of these biased estimators can (and almost certainly will) lead to fallacies in inferences about the differences between cases. This is so regardless of any apparent advantage for these estimators in terms of reduced variance.

The LOO* estimator is approximately unbiased for both 1-NN and 3-NN, and its behavior in the $t$-tests is similar to that of LOO or 10-CV, except that LOO* indicates that the advantage of 1-NN over 3-NN for very small samples is significant when the difference in the TER, LOO and 10-CV

values are not significant. Based on the results in Table 9, we do not see any advantage for LOO* over LOO or 10-CV in inferences about error rate differences.

## 5 Single-Sample Tests for Significance

In the analysis summarized in Tables 7 and 9, we assumed that we have the luxury of drawing 20 independent random samples from the population under study. In most real situations, there is but one small sample. In such a case, we can certainly infer both a 1-NN and a 3-NN classifier and estimate the error rate of each (though the classifiers and their error rate estimates are hardly independent) but we cannot obtain from this single sample any direct measurement of the sample-to-sample variability of the estimated error rates in general, nor of the variability of their difference, in particular. Can we, then, test whether the difference is or is not significant?

One approach would be to assume a value for the variance. Weiss & Indurkhya [20], for instance, adopt this approach in a ±2-SE test for pruned *vs.* unpruned decision trees. When two error rates ($\epsilon_1$ and $\epsilon_3$, with variances $s_1^2$ and $s_3^2$) are not independent, the variance of the difference between the rates is given by $s_\Delta^2 = s_1^2 + s_3^2 - 2s_1 s_3 r$, where $r$ is the correlation coefficient of the two rates, and the lack of independence means that $r \neq 0$. Under the null hypothesis, $\epsilon_1 \approx \epsilon_3 \approx \epsilon$, $s_1^2 \approx s_3^2 \approx s^2$, and $s_\Delta^2 \approx 2s^2(1-r)$. Since it is not at all clear how to obtain valid estimates of $s^2$ and $r$ from a single sample [5, p. 307], any heuristic for $s_\Delta^2$ must tacitly assume particular values for $s^2$ and $r$. The ±2-SE test assumes that $s^2 = \epsilon(1-\epsilon)/N$ and $r = +0.5$.

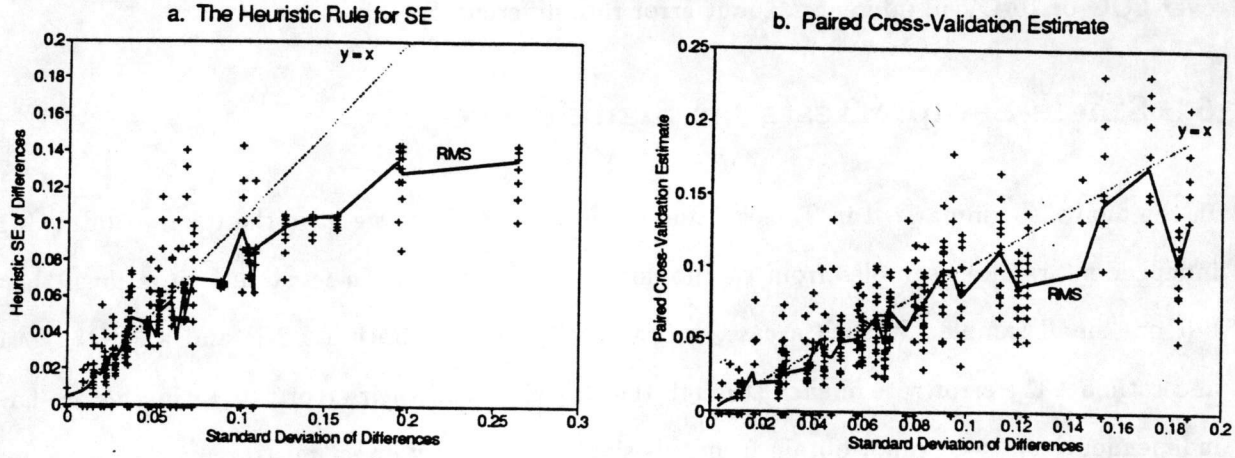
Table 9 summarizes 20 independent samples for each condition. Let $\epsilon_{1,i}$ and $\epsilon_{3,i}$ represent the 10-CV estimates for 1-NN and 3-NN for the $i^{th}$ sample, and

$$\epsilon_i = (\epsilon_{1,i} + \epsilon_{3,i})/2 \qquad \Delta_i = (\epsilon_{3,i} - \epsilon_{1,i}) \qquad \mathrm{SE}_i^2 = \epsilon_i(1-\epsilon_i)/N$$

In every case, $s_\Delta^2$ (the variance of the $\Delta_i$'s) is greater than the mean of the $\mathrm{SE}_i^2$'s, sometimes by as much as a factor of 4. As is shown in Figure 10a, $\mathrm{SE}_i$ is an optimistically biased estimate of $s_\Delta$, and the $\pm 2\mathrm{SE}_i$ test entails a significantly greater risk of Type I error than the intended 0.05 level.

A detailed examination of the data indicates that $r \approx +0.5$ is a reasonable assumption (*i.e.*, that $s_\Delta^2 \approx s_1^2 + s_3^2 - s_1 s_3$ is unbiased), but that the mean of $\mathrm{SE}_{1,i}^2 = \epsilon_{1,i}(1-\epsilon_{1,i})/N$ is optimistically biased as an estimator of $s_1^2$ (and likewise for the 3-NN estimates). We have found (see [13] and

Figure 10: Single-Sample Estimates of Variance




section 2 of this paper) that 10-CV is unbiased and binomially distributed, and that $\epsilon_i$ and $\mathrm{TER}_i$ are uncorrelated, *i.e.*, that

$$E(\epsilon_i) \approx E(\mathrm{TER}_i) \qquad E(\delta_i^2) \approx E(\mathrm{SE}_i^2) \quad \text{and} \quad E(\delta_i \nabla_i) \approx 0$$

where $\delta_i = \epsilon_i - \mathrm{TER}_i$ and $\nabla_i = \mathrm{TER}_i - E(\mathrm{TER}_i)$. From these, we can derive

$$E(s^2) \equiv E\left([\epsilon_i - E(\epsilon_i)]^2\right) \approx E\left([\delta_i + \nabla i]^2\right) \approx E\left(\mathrm{SE}_i^2\right) + E\left(\nabla_i^2\right)$$

That is, that $\mathrm{SE}_i$ is a biased estimator of $s_\Delta$ because it simply ignores the sample-to-sample variance, $E(\nabla_i^2)$, of the inferred classifiers' TER's. The magnitude of the bias ($s_\Delta - \widehat{\mathrm{SE}}$, where $\widehat{\mathrm{SE}} = \mathrm{rms}(\mathrm{SE}_i)$) increases as $\widehat{\mathrm{SE}}$ increases, and increases more rapidly than does $\widehat{\mathrm{SE}}$. The individual $\mathrm{SE}_i$ estimates are also highly variable, and their variance about $\widehat{\mathrm{SE}}$ is approximately proportional to $N^{-1}$.

While it is possible to infer a heuristic rule for estimating $s_\Delta$ given $\mathrm{SE}_i$ from the data in Figure 10a (by, say, fitting a polynomial), we caution that the data underlying Figure 10a are all very similar and very simple — while the shape of the curve probably captures a general, qualitative relationship, the coefficients of a particular fitted polynomial might not adequately describe the relationship for situations involving more complex attributes data.

It is sometimes suggested that one might simply raise the threshold for rejecting the null hypothesis when using this heuristic formula (*e.g.*, $|t| > 2.5$, rather than $|t| > 2.0$, for 95% confidence). If this is done, however, we feel that it would be misleading to report a 95% significance level or

that $|t| > 2.5$. Rather, the result should simply be stated as apparently (heuristically) significant without quantifying it. (How is the value 2.5 to be justified? Why not 4.0? Reporting a level or $t$-value under these circumstances would lend the analysis an undeserved aura of rigor.)

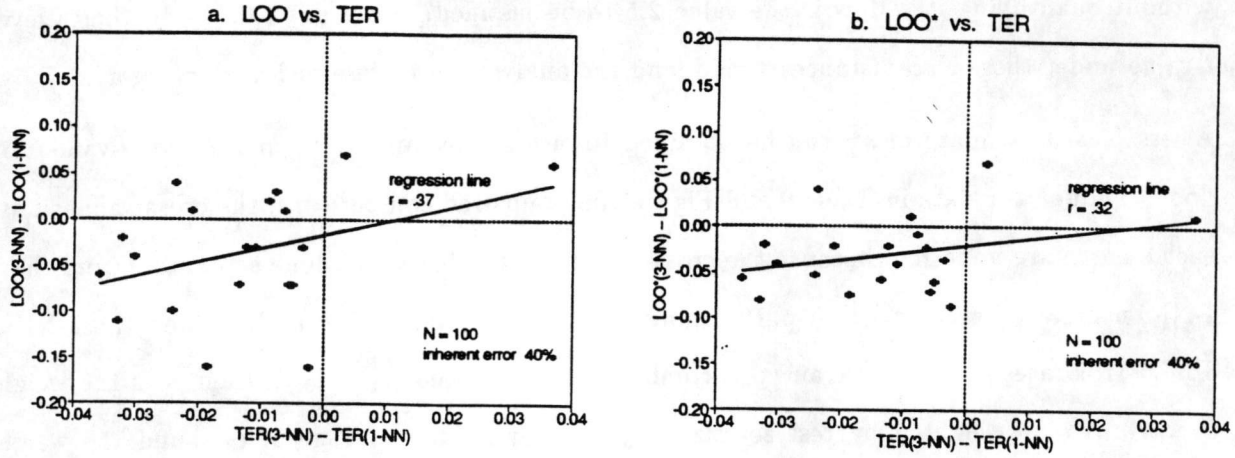
A less biased estimate of $s_\Delta^2$ can be obtained from a single sample if a paired cross-validation is done. Though the data in Table 9 and Figure 10a are paired comparisons, the cross-validations for each sample are unpaired, because the cross-validation for 3-NN was done separately from that for 1-NN, using a different random partitioning. If only one partitioning is done, and the 1-NN and 3-NN error rates for the $j^{th}$ train/test combination are paired ($\epsilon_{1,j}$, $\epsilon_{3,j}$), then $\epsilon_1$ is the weighted average of $\epsilon_{1,j}$ (weighted by test set size), $\epsilon_3$ is the weighted average of $\epsilon_{3,j}$, and the weighted variance of $\delta_j = (\epsilon_{3,j} - \epsilon_{1,j})$ provides a single-sample estimator for the sampling variance of the $v$-fold cross-validation estimate, $\mathcal{S}_\Delta^2 \approx \sum_j (\delta_j - \Delta)^2 m_j / (v-1)N$ (where $\Delta = \sum_j m_j \delta_j / N$). Figure 10b shows the individual and rms values of $\mathcal{S}_\Delta$ for 20 samples each for the 40 population/sample size combinations (as in Table 9). This estimator is much less biased than SE, but biased nontheless[8], and more variable than SE. The bias is greatest for very small samples ($N < 50$).

There are many other problems with these single-sample approaches, all deriving from the fallacy that conclusions about the differences between inference methods based on observations from a single sample generalize to other samples (are representative of the differences for the problem population at hand) or to other problems. First, as can be seen in Table 7, the sign and magnitude of the mean difference may depend on the sample size — an approach such as 3-NN that is beneficial for moderately large samples may actually be harmful for very sparse data. Secondly, as can also be seen in Table 7, the sign and magnitude of the difference may depend on the problem population at hand — 3-NN is beneficial when the inherent error is moderate to large (greater than 2%), but appears to be harmful when the inherent error is very low (less than about 1%). Thirdly, even for a single problem population and a fixed sample size, the classifier error rates and the difference in the paired error rates are so variable from one sample to another that we cannot draw a reliable inference even as to the sign of the difference from only one sample.

We illustrate this last point in Figure 11, where we show the paired differences for 20 samples of the same size from the same population ($N = 100$, inherent error 40%, where the average difference is

[8]This estimator is biased because, for any pair $(\delta_j, \delta_k) k \neq j$, 89% of the items in the two training sets are identical. These values simply are not free to vary as widely within a single sample as they would be from sample to sample.

Figure 11: Single-Sample Differences Between 1-NN and 3-NN




negative and significant at the 95% confidence level). In these figures we plot the paired difference in TER along the horizontal axis, and the paired difference of estimated error along the vertical axis (only LOO and LOO* are shown, 10-CV is very similar to both of these, but even more variable and less correlated). We can see that even here, where the average TER difference is strongest, the TER difference for 2 of our 20 samples is opposite in sign to the average difference. The difference in estimated error rates is but poorly correlated with the difference in TER, even more variable, and more likely to reverse the sign. LOO* is less variable[9] and less likely to reverse the sign or be grossly wrong as to the magnitude than LOO or 10-CV, but still poorly correlated with the difference in TER — but, even if it were perfectly correlated, the difference in TER for a single sample is not a reliable indicator of the expected difference in performance. Regardless of how we approach obtaining an estimate of $s_\Delta$, single-sample tests are apt to be misleading if we cannot be confident that at least the sign of the single-sample difference is correct. In the case of 3-NN *vs.* 1-NN (and, by analogy, pruned *vs.* unpruned decision trees), we cannot be sure of even that much, even for a population and sample size where the average difference is highly significant.

Bailey & Elkan [4] note this problem of high variation and poor correlation and suggest that it might be problematic as to the current machine learning approaches to determining the differences between inference methods. Our experiments show that their misgivings are absolutely correct.

[9] And, therefore, $SE = \epsilon(1-\epsilon)/N$ is not an appropriate estimate of LOO*'s variance.

Table 12: $t$-test for Stopped *versus* Unpruned Trees

| Inherent Error % | Sample Size | $t$-statistic TER | APP | LOO | 10-CV | 632b | LOO* |
|---|---|---|---|---|---|---|---|
| 0.1 | 24 | 4.2 | 4.4 | ⊘ 1.9 | 2.4 | 3.1 | 2.9 |
| | 36 | 2.3 | 6.0 | 4.6 | .5 | 3.9 | 3.9 |
| | 48 | 3.9 | 11.2 | 3.5 | 6.1 | 6.1 | 4.7 |
| | 96 | 1.0 | 1.5 | ⊙ 3.0 | ⊙ 3.1 | ⊙ 5.1 | ⊙ 4.8 |
| 5 | 24 | 2.9 | 3.9 | 2.6 | 2.7 | 2.9 | 3.3 |
| | 36 | 1.1 | ⊙ 7.3 | ⊙ 3.3 | 2.0 | ⊙ 3.1 | ⊙ 3.3 |
| | 48 | .9 | ⊙ 2.8 | 2.2 | 1.6 | 2.1 | ⊙ 2.8 |
| | 96 | 3.0 | 2.7 | ⊙ 2.0 | ⊘ 1.8 | 2.4 | 3.1 |
| 10 | 24 | 2.2 | 1.2 | .1 | .5 | .4 | .2 |
| | 36 | 1.2 | ⊙ 2.9 | 1.6 | 1.7 | ⊙ 2.5 | 1.6 |
| | 48 | .2 | 1.8 | − .2 | .2 | .9 | .2 |
| | 96 | 1.2 | 2.2 | 2.0 | 1.6 | 2.0 | 1.7 |
| 25 | 24 | − .3 | ⊙ 5.3 | 1.0 | .5 | 1.8 | 1.0 |
| | 36 | .2 | ⊙ 2.9 | 1.0 | .1 | 2.0 | 1.0 |
| | 48 | .6 | ⊙ 5.9 | 1.6 | 1.0 | ⊙ 7.3 | 2.3 |
| | 96 | .1 | 2.2 | .8 | 1.0 | 1.2 | .9 |

⊙ TER difference is not significant, but this is
⊘ TER difference is significant, but this is not

## 5.1 A Similar Decision Tree Experiment

A series of experiments was conducted to examine whether these results from nearest neighbors classifiers can be generalized to decision trees derived from nominal rather than continuous attributes. Eight samples each of sizes 24, 36, 48, and 96 were drawn from noisy contact lens populations [13] with inherent error rates of 0.1, 5, 10, and 25%. Both an unpruned and a stopped[10] decision tree were inferred from each sample, and the various error rate estimates were determined for each of the trees.

In Table 12 we show $t$-statistics for the paired differences between stopped and unpruned trees, highlighting those cases where the paired $t$-test using one of the estimators is misleading as to the significance of the difference in true error. Overall, the TER data indicate that the stopped trees are less accurate, but that the difference is not significant for higher inherent error rates or larger samples. (For each entry in Table 12, there are 7 degrees of freedom, and $t > 2.365$ is significant at the 95% level.)

[10]Stopped using the multiple hypergeometric probability test [12] (an extension of Fisher's exact test [2]).

The data summarized in Table 12 suggest that 632b and LOO* tend to exaggerate the significance of the difference in error rates between stopped and unpruned trees, and that the risk of committing a Type I error (falsely rejecting the null hypothesis) when using these estimators is markedly higher than 0.05 (the value implied by a test at nominally the 95% confidence level). LOO and 10-CV seem to lead to correct decisions regarding the null hypothesis for inherent error rates of 10% or more, but may sometimes overstate or understate significance for lower error rates. On the whole, 632b and LOO* do not appear to have any advantage over LOO or 10-CV for these data. 10-CV is recommended for these comparisons because it has the lowest computational cost among these four methods and also appears to have the least added risk of Type I or Type II error.

The $t$-tests in Table 12 relate to the average paired difference over 8 samples. As was the case for nearest neighbor classifiers (see Figure 11), the difference observed for a single sample is highly variable, and not trustworthy even as to the sign of the difference. The $\sqrt{\epsilon(1-\epsilon)/N}$ heuristic for the standard error of the paired differences between stopped and unpruned decision tree error rates is strongly biased. Though similar in its shape to the corresponding relationship for 3-NN *vs.* 1-NN (see Figure 10a), the bias is quantitatively different for these discrete attribute trees than for those continuous attribute classifiers.

# 6 Conclusions and Recommendations

1. The textbook formula based on the normal approximation to the binomial is not a good approximation to the confidence interval of an error rate estimator for small samples or low error rates, even if a 'continuity adjustment' is made. When the number of observed errors is less than 10, the more exact limits calculated from the Beta distribution should be used.

2. The confidence interval for a single estimate (even the more exact Beta distribution limit) does not provide a good significance test for the difference between two estimated error rates. These limits, especially the ±2-SE limits, have a high additional risk of Type I error.

3. For comparing two independent rates on small samples, the textbook approximation using the combined variance entails a slightly greater risk of Type I error than the unbiased approximation given here.

4. For paired comparison of two classifier inference methods, Student's $t$ test for the average difference over several independent samples is appropriate. The 10-CV estimator is recommended for these tests because it corresponds most closely to the significance of the differences in TER. 632b is not recommended because it is biased, and differently biased for different inference methods. LOO* has no significant advantage for these tests, and is not recommended because of its higher computational cost and increased risk of Type I error.

5. Paired comparison of inference methods based on a single sample may be misleading, even as to the sign of the difference, because the difference in error estimates is highly variable and not correlated with the difference in true error, and also because the commonly used $\sqrt{\epsilon(1-\epsilon)/N}$ heuristic for the SE of the differences is biased and lacks rigor.

## 7 Acknowledgement

The authors are indebted to Dr. Sholom Weiss (Rutgers University) and Dr. Leo Breiman (University of California, Berkeley) for their assistance in sorting out the issues relevant to single-sample significance tests. We are also grateful to the editor and reviewers of an earlier version of these papers, whose suggestions have been invaluable in correcting many of the weaknesses and oversights.

## References

[1] M. Abramowitz and I. A. Stegun. *Handbook of Mathematical Functions.* Dover Publications, Inc., New York, 1972.

[2] A. Agresti. *Categorical Data Analysis.* John Wiley & Sons, New York, 1990.

[3] T. W. Anderson and S. L. Sclove. *The Statistical Analysis of Data.* The Scientific Press, Palo Alto, 2nd edition, 1986.

[4] T. L. Bailey and C. Elkan. Estimating the accuracy of learned concepts. In *Proceedings of the 13th International Joint Conference on Artificial Intelligence (IJCAI-93)*, volume 2, pages 895–900, San Mateo, CA, 1993. Morgan Kaufmann.

[5] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees.* Wadsworth & Brooks/Cole Advanced Books & Software, Pacific Grove, CA, 1984.

[6] B. Efron. Estimating the error rate of a prediction rule: Improvement on cross-validation. *Journal of the American Statistical Association*, 78:316–331, 1983.

[7] R. A. Fisher. *Statistical Methods for Research Workers.* Oliver & Boyd, Edinburgh, 14th edition, 1970. (the quotation is from the preface to the first (1925) edition).

[8] J. A. Hartigan. *Bayes Theory.* Springer-Verlag, New York, 1983.

[9] G. R. Iversen. *Bayesian Statistical Inference.* Sage University Paper series on Quantitative Applications in the Social Sciences, series no. 07-043. Sage Publications, Beverly Hills, 1984.

[10] J. J. L. Hodges and E. L. Lehman. *Basic Concepts of Probability and Statistics.* Holden-Day, Inc., Oakland, CA, 1970.

[11] H. Jeffreys. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London, A*, 186:453–461, 1946.

[12] J. K. Martin. An exact probability metric for decision tree splitting and stopping. Technical Report 95-16, University of California, Irvine, Irvine, CA, 1995.

[13] J. K. Martin and D. S. Hirschberg. Small sample statistics for classification error rates, I: error rate measurements. Technical Report 95-42, University of California, Irvine, Irvine, CA, 1995.

[14] W. Mendenhall, D. D. Wackerly, and R. L. Scheaffer. *Mathematical Statistics with Applications.* PWS-KENT Publishing Co., Boston, 4th edition, 1990.

[15] J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1:81–106, 1986.

[16] S. Rasmussen. *An Introduction to Statistics with Data Analysis.* Brooks/Cole Publishing Co., Pacific Grove, CA, 1992.

[17] C. Schaffer. Overfitting avoidance as bias. *Machine Learning*, 10:153–178, 1993.

[18] J. W. Shavlik and T. G. Dietterich, editors. *Readings in Machine Learning.* Morgan Kaufmann, San Mateo, CA, 1990.

[19] S. M. Weiss. Small sample error rate estimation for k-nearest neighbor classifers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13:285–289, 1991.

[20] S. M. Weiss and N. Indurkhya. Decision tree pruning: Biased or optimal? In *Proceedings of the 12th National Conference on Artificial Intelligence (AAAI-94)*, volume 1, pages 626–632, Menlo Park, CA, 1994. AAAI Press.

# A Appendix — The Beta Distribution

What is the *posterior* distribution of the true error $\tau$, $P\{\tau < x \mid M, m\}$, given a test set of $M$ items and that the observed number of errors $m$ is binomially distributed? The binomial distribution is

$$P\{m \mid M, \tau\} = \frac{M!}{m!\,(M-m)!}\ \tau^m\,(1-\tau)^{M-m}$$

The expected value (mean) of $m/M$ is $E(m/M) = \tau$, and its variance is $\sigma^2 = \tau(1-\tau)/M$. From this, one might surmise that the expected value of $\tau$ given $m/M$ would be $m/M$, but this is not so because it is possible to obtain the result $m = 0$ when $\tau \neq 0$:

$$\begin{aligned} P\{m = 0 \mid M = 10, \tau = 0.0\} &= 1.0000 \\ P\{m = 0 \mid M = 10, \tau = 0.1\} &= 0.3487 \\ P\{m = 0 \mid M = 10, \tau = 0.2\} &= 0.1074 \\ &\vdots \\ P\{m = 0 \mid M = 10, \tau = 1.0\} &= 0.0000 \end{aligned}$$

The expected value of $\tau$ given $m/M$ (in particular, and the posterior distribution in general) depends on the relative likelihood of finding various values of $\tau$, regardless of $m/M$. This *a priori*, unconditional probability function $f(\tau)$ is known as the prior distribution or simply the *prior* of $\tau$. The relationship between the prior and posterior distributions is given by Bayes' Theorem:

$$P\{\tau < x \mid M, m\} = \int_0^x P\{m \mid M, \tau\} f(\tau) d\tau \;\Big/\; \int_0^1 P\{m \mid M, \tau\} f(\tau) d\tau$$

or by the derivative of this expression evaluated at $x = \tau$, the *posterior density function*:

$$f\{\tau \mid M, m\} = P\{m \mid M, \tau\} f(\tau) \;\Big/\; \int_0^1 P\{m \mid M, \tau\} f(\tau) d\tau$$

It is convenient if $f(\tau)$ can be expressed in such a form that Bayes' Theorem is easily integrated; such priors are sometimes called *conjugate priors*. The Beta distribution with parameters $u$ and $v$, $\mathrm{Be}(\tau, u, v)$, is a family of conjugate priors (sometimes called *Beta priors* or *Dirichlet priors*) for the binomial which are capable of expressing, to at least a very good approximation, a very wide variety of plausible prior distributions of $\tau$ (see Iversen [9, pp. 18-33]).

$$\mathrm{Be}(\tau, u, v) = \tau^{u-1}(1-\tau)^{v-1} / B(u, v)$$

$$\text{where} \qquad B(u, v) = \int_0^1 \tau^{u-1}(1-\tau)^{v-1} d\tau = \Gamma(u)\Gamma(v) \,/\, \Gamma(u+v) \quad \text{is the Beta function}$$

$\Gamma(z)$ is the Gamma function, a generalization of the factorial. For positive integers $n$, $\Gamma(n) = (n-1)!$. In general, $\Gamma(z) = (z-1)\Gamma(z-1)$ and, in particular, $\Gamma(1/2) = \sqrt{\pi}$. See [1, pp. 255-258,944-945] for information on these functions. These Beta priors give a Beta distribution as the posterior:

$$f(\tau \mid M, m) = \mathrm{Be}(\tau, m+u, M-m+v)$$

$$P\{\tau < x \mid M, m\} = I(x, m+u, M-m+v) = \int_0^x \mathrm{Be}(\tau, m+u, M-m+v)\, d\tau$$

where $I(x, m+u, M-m+v)$ is the Incomplete Beta function. For this Beta distribution, the posterior mean (expected value, $\mu$) and variance ($\sigma^2$) of the true error rate $\tau$ are

$$\mu = (m+u)/(M+u+v) \qquad \sigma^2 = \mu(1-\mu)/(M+u+v+1)$$

and the most likely value (the mode) is

$$\text{mode} = \begin{cases} 0, & \text{if } m = 0 \\ (m+u-1)/(M+u+v-2), & \text{if } 0 < m < M \\ 1, & \text{if } m = M \end{cases}$$

Hartigan [8, pp. 48-50,76-78] gives a synopsis of arguments leading to $\mathrm{Be}(\tau, 1/2, 1/2)$, the *Jeffreys prior* [11], as a good choice for finding confidence intervals from the binomial. This leads to

$$P\{\tau < x \mid M, m\} = I(x,\ m+0.5,\ M-m+0.5) \tag{5}$$

$$= \frac{x^{m+0.5}(1-x)^{M-m+0.5}}{(m+0.5)\ B(m+0.5, M-m+0.5)} \left\{ 1 + \sum_{i=0}^{\infty} \frac{B(m+1.5, i+1)}{B(M+1, i+1)}\ x^{i+1} \right\}$$

For large values of $M$ and $m$, the series in Equation 5 converges very slowly. Solutions to determine the confidence interval are very sensitive to numerical precision errors (including the use of Stirling's [1, p. 257] approximation for $\Gamma(z)$). For $m > M/2$, advantage can taken of the symmetry of the Incomplete Beta function, $I(x,u,v) = 1 - I(1-x,v,u)$. For $m \gg 0$, advantage can be taken of the recurrence relation $I(x,u,v) = x\ I(x,u-1,v) + (1-x)\ I(x,u,v-1)$.