PERGAMON

Computers in Biology and Medicine 33 (2003) 509–531

Computers in Biology and Medicine

www.elsevier.com/locate/compbiomed

Accurate confidence intervals for binomial proportion and Poisson rate estimation

Timothy D. Ross*

Air Force Research Laboratory COMPASE Center, AFRL/SNAR, 2241 Avionics Cl, Building 620, WPAFB, OH 45433, USA

Received 9 August 2002; accepted 25 February 2003

Abstract

Estimates of proportion and rate-based performance measures may involve discrete distributions, small sample sizes, and extreme outcomes. Common methods for uncertainty characterization have limited accuracy in these circumstances. Accurate confidence interval estimators for proportions, rates, and their differences are described and MATLAB programs are made available. The resulting confidence intervals are validated and compared to common methods. The programs search for confidence intervals using an integration of the Bayesian posterior with diffuse priors to measure the confidence level. The confidence interval estimators can find one or two-sided intervals. For two-sided intervals, either minimal-length, balanced-tail probabilities, or balanced-width can be selected.
Published by Elsevier Ltd.

Keywords: Confidence intervals; Significance; Binomial; Poisson; Proportion; Rate; Differences; Minimal length; Exact; Bayesian; Classifier evaluation

1. Introduction

Classification system assessments typically involve performance measures in the form of proportions and rates. Commonly used measures include proportions or probabilities, such as probability of detection, probability of identification, and false alarm rates [1]. Similar measures are of interest in many other disciplines. Performance assessment provides the basis for a variety of decisions related to programmatic planning, technical design, and transition of classification technology. This assessment invariably requires estimating the measures of interest. Proper use of these estimates depends on an accurate characterization of the estimates' uncertainties. There are several sources of uncertainty,

* Tel.: 1-937-255-5668; fax: +1-937-656-4027.
E-mail address: t.ross@ieee.org (T.D. Ross).

0010-4825/03/$ - see front matter Published by Elsevier Ltd.
doi:10.1016/S0010-4825(03)00019-2

but this paper is concerned solely with statistical uncertainty. We define "statistical uncertainty" as that which is appropriate when we have a *random sample* from a *representative population* that is *well separated from training data.* In practice, we rarely meet any of these conditions, much less all three. However, proper treatment of "statistical uncertainty" is still of interest. It provides a foundation upon which to build more general concepts and will have direct application in some instances. There are a variety of ways in which an estimate's uncertainty may be characterized, but this paper is concerned only with characterizations based on confidence intervals and significance of differences. After some brief background, Section 3 develops expressions for confidence interval estimators for proportions and rates. Section 4 explains their computer implementation and Section 5 reports on the validation process. Section 6 develops the confidence intervals for differences of proportions and rates.

2. Background

Confidence intervals (CIs) have an intuitive meaning, that is, the interval within which you can be confident that the true value lies, but this intuitive concept is surprisingly difficult to formalize. Particularly for a non-statistician, it is difficult to find an accessible explanation of confidence intervals that is appropriate to the distributions and conditions of their application along with algorithms for computing them. The ideal CI estimator (CIE) would be accurate across the conditions of interest, easily coded, and validated. The conditions of interest for classifier performance assessment include proportions near zero and one, rates near zero, and small numbers of samples. Proportion estimates may involve a small number of samples, often less than 100 and occasionally only a half-dozen or so. These proportions are also often close to 1.0. A handful of, or even zero, events are also an important basis for rate estimation, which may occur when the area of the available test data is on the order of one over the true false alarm rate. CI accuracy is important since significant resources go towards test execution and consequential decisions are made based on the results' uncertainties; therefore, a significant effort is warranted towards accurate confidence intervals.

There are two main perspectives on confidence intervals, "classical" and "Bayesian." Although these two perspectives are fundamentally different and generate some controversy within the statistics community, that difference is of little consequence under an assumption of limited prior information. From the "classical" or sampling theory perspective [2], there is some unknown, but "true" fixed value of the estimated parameter. The confidence interval is then a pair of random variables L and U (some function of the sampling random variable), not the computed limits themselves. These random variables cover the true value of a parameter θ being estimated with a certain probability, i.e., $\Pr(L_{\text{r.v.}} < \theta_{\text{fixed}} < U_{\text{r.v.}}) = 1 - \alpha$. When we compute particular confidence limits, we only have the realizations l and u of the random variables L and U. These either do or do not enclose θ. It is not correct to say that they probably enclose θ. Another statement of classical confidence intervals is that $100 \times (1 - \alpha)\%$ of all samples will result in confidence limits which enclose the true parameter value. From the "Bayesian" perspective [3], the confidence interval is the computed limits and it is the parameter that is a random variable. From this perspective, it is correct to say, $\Pr(l_{\text{fixed}} < \theta_{\text{r.v.}} < u_{\text{fixed}}) = 1 - \alpha$. We use the Bayesian perspective here because we can compute precise confidence intervals for the full range of conditions of interest. If the prior probabilities are "diffuse" (i.e., do not contribute significantly to the resulting posterior probabilities, which is assumed throughout this paper) then the two perspectives result in the same interval values.

When estimating a parameter θ from measurement x, the Bayesian formulation [3] is that the posterior distribution $f_{\theta|x}(\theta\,|\,x)$ is the product of the likelihood $f_{x|\theta}(x\,|\,\theta)$, which is often determined by standard models, and the prior distribution $f_\theta(\theta)$ with an appropriate constant to ensure that the posterior is a true distribution (i.e., integrates to 1.0). That is,

$$f_{\theta|x}(\theta\,|\,x) = \frac{f_{x|\theta}(x\,|\,\theta) f_\theta(\theta)}{\int f_{x|\theta}(x\,|\,\theta) f_\theta(\theta)\,\mathrm{d}\theta}.$$

For proportion and rate estimation, the standard likelihood models are the binomial ($f_{x|p}(x\,|\,n,p) = \binom{n}{x} p^x(1-p)^{(n-x)},\ x = 0,1,2,\ldots,n$) and Poission ($f_{x|\lambda}(x) = \lambda^x \mathrm{e}^{-\lambda}/x!,\ x = 0,1,2,\ldots$) distributions respectively, where x is the measurement and p, n and λ are distribution parameters.

We are assuming here that there is no significant prior information, so the prior distribution is taken to be essentially uniform across all possible values. Bayesian theorists favor a particular form of prior distribution, even when they are diffuse, known as conjugate priors. These are distributions, from a family of distributions, which result in posterior distributions that are also from that same family. We are numerically integrating the posterior, so priors of conjugate form are not essential, but beta and gamma distributions are conjugate priors for binomial and Poisson likelihoods, respectively. The beta distribution has the form

$$f_p(p) = \begin{cases} \dfrac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha,\beta)} & 0 < p < 1, \\ 0 & \text{otherwise}, \end{cases}$$

where $B(\alpha,\beta)$ is the beta function, which for α and β positive integers may be expressed

$$B(\alpha,\beta) = \frac{(\alpha-1)!(\beta-1)!}{(\alpha+\beta-1)!}.$$

A diffuse beta distribution may be defined with $\alpha = \beta = 1$, which is exactly the uniform distribution on the interval (0,1). The gamma distribution has the form

$$f_\lambda(\lambda) = \begin{cases} \dfrac{\lambda^{\alpha-1}\mathrm{e}^{-\lambda/\beta}}{\beta^\alpha \Gamma(\alpha)}, & x > 0, \\ 0, & x \leqslant 0, \end{cases}$$

where $\Gamma(\alpha)$ is the gamma function, which for α a positive integer may be expressed $\Gamma(\alpha) = (\alpha-1)!$. A diffuse gamma distribution may be defined with $\alpha = 1$ and $\beta = K_\beta$, for K_β some very large value. With these parameters, the prior distribution for rates becomes

$$f_\lambda(\lambda) = \begin{cases} \dfrac{\mathrm{e}^{-\lambda/K_\beta}}{K_\beta} & x > 0 \\ 0 & x \leqslant 0 \end{cases}$$

which is now of the form of an exponential distribution.

The posterior for the proportion estimate

$$f_{p|x}(p\,|\,x) = \frac{f_{x|p}(x\,|\,p) f_p(p)}{\int f_{x|p}(x\,|\,p) f_p(p)\,\mathrm{d}p} \quad 0 < p < 1, \quad x = 0, 1, 2, \ldots, n$$

for the binomial likelihood and uniform prior is then

$$f_{p\,|\,x}(p\,|\,x) = \frac{\binom{n}{x} p^x (1-p)^{(n-x)}}{\int f_{x\,|\,p}(x\,|\,p) f_p(p)\,\mathrm{d}p} = (n+1)\binom{n}{x} p^x (1-p)^{(n-x)}, \quad 0 < p < 1.$$

The value of the normalizing integral is apparent from the properties of the beta function with integer parameters. The proportion posterior is of beta distribution form with parameters $\alpha = x + 1$ and $\beta = n - x + 1$.

The posterior for rate estimate

$$f_{\lambda|x}(\lambda\,|\,x) = \frac{f_{x|\lambda}(x\,|\,\lambda) f_\lambda(\lambda)}{\int f_{x|\lambda}(x\,|\,\lambda) f_\lambda(\lambda)\,\mathrm{d}\lambda} \quad 0 < \lambda, \quad x = 0, 1, 2, \ldots, n$$

for the Poission likelihood and exponential prior is then

$$f_{\lambda|x}(\lambda\,|\,x) = \frac{(\lambda^x \mathrm{e}^{-\lambda}/x!)(\mathrm{e}^{-\lambda/K_\beta}/K_\beta)}{\int f_{x|\lambda}(x\,|\,\lambda) f_\lambda(\lambda)\,\mathrm{d}\lambda} = \frac{(\lambda^x \mathrm{e}^{-(1+1/K_\beta)\lambda}/K_\beta x!)}{\int f_{x|\lambda}(x\,|\,\lambda) f_\lambda(\lambda)\,\mathrm{d}\lambda} \approx \frac{\lambda^x \mathrm{e}^{-\lambda}}{x!} \quad 0 < \lambda,$$

$$x = 0, 1, 2, \ldots, n.$$

That final approximation is better and better for larger and larger K_β. The value of the normalizing integral (i.e., $1/K_\beta$) is apparent from the posterior being of gamma distribution form, particularly gamma distributed with parameters $\alpha = x + 1$ and $\beta = 1$.

3. Confidence interval development

This section develops expressions for the confidence interval estimators, first for proportion estimates and then for rate estimates. The development is from a Bayesian perspective; but as noted above, for the assumptions made, the classical confidence intervals have the same numerical values. The difference CIEs will be developed in Section 6.

3.1. Confidence intervals for proportion estimates

Many common performance measures are simple proportions, sometimes referred to as probabilities. The positive outcome of a binary event occurs with some probability p. Let x represent the number of positive outcomes in n independent identically distributed "Bernoulli" trials. The number of positive outcomes (x) has a binomial distribution with parameters p and n. The standard problem then is: when given x and n, what is the best estimate of p and how confident are we in that estimate? The estimation of p by $\hat{p} = x/n$ has most of the desired properties of an estimator (unbiased, efficient, maximum likelihood and maximum a posteriori probability) for reasonable assumptions.

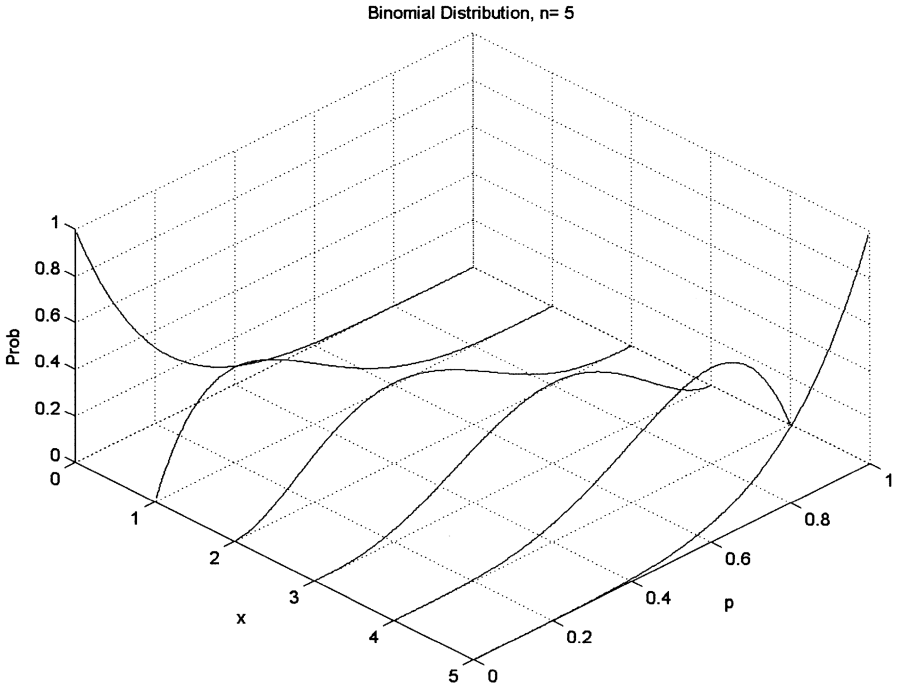

Fig. 1. Joint distribution.

The more interesting issue is the characterization of the estimate's uncertainty. A complete characterization of this uncertainty is represented by the distribution of the true values conditioned on the estimate, i.e., $f_{p|\hat{p}}(p\,|\,\hat{p})$, which is equivalent to $f_{p|x}(p\,|\,x)$ for $\hat{p} = x/n$ and n fixed. Fig. 1 is an example joint distribution of x and p, given $n = 5$.

Our likelihood, as the familiar binomial distribution $f_{x|p}(x\,|\,n, p) = \binom{n}{x} p^x(1-p)^{(n-x)}$, $x = 0, 1, 2, \ldots, n$, is the cross-section of this plot at a fixed p. Similarly, the cross-section of this plot for a fixed x is proportional to the posterior distribution

$$f_{p|x}(p\,|\,x) = \begin{cases} (n+1)\binom{n}{x} p^x(1-p)^{n-x}, & p \in [0,1] \\ 0 & \text{otherwise.} \end{cases}$$

Suppose $n = 5$ and we happen to realize an x of 2 (as in Fig. 2). In this case, $\hat{p} = 2/5$. Given x, we are after the confidence intervals (CIs) such that α is the probability that the p that produced this x was from outside the CI. That is, we want to find the interval $[a, b]$ such that $\Pr\{p \in [a, b]\,|\,x\} = \int_a^b f_{p|x}(p\,|\,x)\,\mathrm{d}p = 1 - \alpha$. For the assumed diffuse prior, $f_{p|x}(p\,|\,x)$ is the beta distribution with parameters $(x+1, n-x+1)$ [3]. The problem of finding a confidence interval is one of finding a and b such that the integral above equals $1 - \alpha$.

3.2. Confidence intervals for rate estimation

Another principal measure in classifier performance assessment is a rate, especially the false alarm (FA) rate (FAR). The FAR might be in FAs per unit area, unit time, or some other dimension,

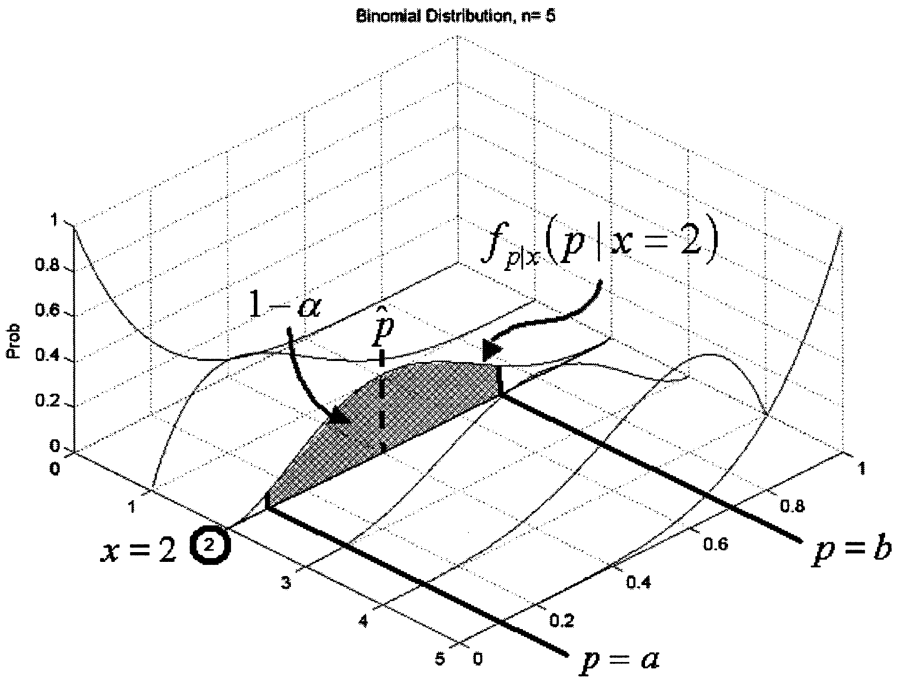

Fig. 2. Posterior distribution.

depending on the application. The following discussion will be in terms of area, however other bases for the rate would result in the same arguments. Our basic likelihood model for FAs is the Poisson distribution, i.e., $f_{x|\lambda}(x) = (\lambda^x e^{-\lambda}/x!)$, $x = 0, 1, 2, \ldots$ where x is now the number of FAs observed in a test area of size A. The parameter λ may be thought of as the number of FAs expected to occur in an area of size A. The natural estimator of rate $R = \lambda/A$ is simply $\hat{R} = x/A$. For a diffuse prior, we found the posterior distribution to be

$$f_{\lambda|x}(\lambda \mid x) = \begin{cases} \dfrac{\lambda^x e^{-\lambda}}{x!}, & \lambda \in [0, \infty] \\ 0 & \text{otherwise} \end{cases}$$

which is a gamma distribution with parameters $(x + 1, 1)$. An α confidence interval in terms of λ is an interval $[a, b]$ such that $\Pr\{\lambda \in [a, b] \mid x\} = 1/x! \int_a^b \lambda^x e^{-\lambda} \, d\lambda = 1 - \alpha$, assuming $0 \leqslant a \leqslant b$. The confidence interval on the rate $\hat{R}$ is then $[a/A, b/A]$.

4. Confidence interval estimators

We now consider the computation of CIs. CIs were of interest long before modern computers, so practical computation options drove the approach to CIEs. The availability of standard normal distribution tables and their appropriateness for large samples and non-extreme

measurements lead to that well known approach. An alternative Clopper–Pearson method [4] is also well known. This method was initially implemented by an iterative approach and various authors have made code available [5–7]. Although this second method only provides for a bound on α for parameters of discrete distributions, it was such an improvement over the normal approximation that it is known as an "exact" method. Finally, Bayesian concepts [3] for CIs have also become easier to implement with higher order languages. Brenner and Quan [8] reported Bayesian posterior proportion CIs and their comparison with Clopper–Pearson CIs. We assume that such "integration of the Bayesian posterior" (IBP) approaches have also been implemented for rate CIs, but we are not aware of the CIEs for either having been made generally available.

The CIEs made available here are two scriptable functions, prop_ci for proportions and rate_ci for rates. They have similar input and output functionality. There are up to five inputs: x-number of positive outcomes or events, n or A-number of trials or test area, desired α, *method*, and whether to be *verbose.* The acceptable range of inputs are n: any integer from 1 to 10^5, x: any integer from 0 to n (or 10^5 for rates), α: between 10^{-4} and 1, *method*: 1 through 6, and *verbose*: 0 or 1. The inputs x, n, A, α have the same meanings as previously introduced. The six *methods* are those described below. If there are only four input parameters, they are assumed to be x, n or A, α and *method.* Three input parameters are assumed to be x, n or A, and α. If *verbose* is not specified, the default value of 0 is used. If *method* is not specified, the default is 2. The output depends on the *verbose* setting. If *verbose* is 1 then there are thirteen outputs: estimated value (either x/n or x/A), lower confidence limit, upper confidence limit, input x value, input n or A value, desired α, method, CI length, lower tail probability, upper tail probability, actual α, α error, and run time in seconds. If *verbose* is 0 then only the first three of those values are output. When *method* 1 is specified, both the lower one-sided limit and the upper one-sided limit are returned, even though the user will only be interested in one or the other. The CIEs were implemented in MATLAB (© The Mathworks, Inc.) code, Release 13, utilizing the Statistics Toolbox. The MATLAB code for the CIEs is available at [9].

The remainder of this section develops some of the particulars of the six *methods.* There are four versions of the IBP CIE: one-sided, minimal-length [10], balanced-width [5], and balanced-tail two-sided. These four IBP versions are *methods* 1–4. The IBP methods are implemented by a binary search for a and b that produce the desired α. The IBP CIEs were designed to yield an actual α within 0.00005 of the desired α. The other two *methods* (5 and 6) are implementations of the normal approximation and Clopper–Pearson approaches.

Method 1, *one-sided IBP CIs*: One-sided CIs are of interest when we are only concerned about whether the true value is on one side of a limit, e.g., greater than some lower limit. Using the proportion CIs as an example, the one-sided lower limit is $a \ni \Pr\{p \geqslant a \,|\, x\} = 1 - \alpha$. The one-sided IBP method first searches with respect to the one-sided lower limit then separately with respect to the one-sided upper limit. Method 1 returns both the one-sided lower and the one-sided upper limits.

Two-sided CIs are not uniquely determined by α, so three additional criteria are considered here.

Method 2, *minimal-length IBP CIs*: The "minimal length" criteria is that the CI be as short as possible for the given α. That is, for minimal-length CI $[a_{\mathrm{ml}}, b_{\mathrm{ml}}]$ and θ the quantity of interest, $\forall[a,b] \ni \Pr\{\theta \in [a,b] \,|\, x\} = 1 - \alpha, (b_{\mathrm{ml}} - a_{\mathrm{ml}}) \leqslant (b - a)$. Let f_θ be the posterior distribution. If

$x \neq 0$ and $x \neq n$ (proportions) then CI $[a_{\mathrm{ml}}, b_{\mathrm{ml}}]$ is of minimal-length, for a given α, if and only if $f_\theta(a_{\mathrm{ml}}) = f_\theta(b_{\mathrm{ml}})$. That is, the posterior distribution value is the same at both limits, except when a limit is zero or one. The two-sided minimal-length CIE searches with respect to the posterior distribution value. For each hypothesized value of the posterior distribution, a search is conducted for a corresponding limit on either side of the posterior's mode.

Method 3, *balanced-width IBP CIs*: The "balanced-width" criterion is that the CI be centered on the point estimate. That is, for CI $[a, b]$ and estimated value $\hat{\theta}, \hat{\theta} - a = b - \hat{\theta}$. The two-sided balanced-width method searches with respect to interval width. For extreme values of x, this criterion could result in CIs that extend beyond the possible values of the parameter, e.g., negative lower limits on p. Method 3 sacrifices balance when the CI would extend below zero or above one (for proportions), in which case the appropriate limit is set to zero or one and the other limit is set for the desired α.

Method 4, *balanced-tail IBP CIs*: The "balanced-tail" criterion is that half of α should be below the lower limit and half above the upper limit. That is, for CI $[a, b]$, measurement x, and quantity of interest θ, $\Pr\{\theta \leqslant a \,|\, x\} = \alpha/2$ and $\Pr\{\theta \geqslant b \,|\, x\} = \alpha/2$. When $x = 0$ or $x = n$ (for proportions), the balanced-tail criteria will result in limits that do *not* enclose the point estimate; however, they are still meaningful and are provided as such. The two-sided balanced-tail CIE is implemented by calling the one-sided CIE (method 1) with $\alpha/2$.

Method 5, *Clopper–Pearson CIs*: Method 5 is based on Clopper and Pearson [11]. Wilks [12], has a more recent expression of the key result, that for our discrete distributions, $\Pr(\theta \in [a, b]) \geqslant 1 - \alpha$, rather than with equality. We use the MATLAB implementation, which is based on Daly'92 [7] using Eqs. (4) and (5) for binofit.m and Eqs. (8) and (9) for poissfit.m, except the normal approximation is used in poissfit for $x \geqslant 100$. Our method 5 is exactly the MATLAB functions, except we correct the NaN returned by poissfit.m when $x = 0$ for earlier MATLAB releases. In principle, the tail probabilities may be set independently for this method. The intent of the MATLAB implementation is for the CIs to have balanced-tails; however the errors are asymmetric, so the resulting CIs may not have balanced-tails.

Method 6, *normal approximation CIs*: A proportion estimate's confidence interval width with this CIE is $\pm Zc\sqrt{\hat{p}(1-\hat{p})/n}$, where Zc is the value along a standard normal distribution with cdf equal to $\alpha/2$, $\hat{p}$ is the estimate, and n the number of samples. Some sources suggest that the approximation is adequate if $n > 30$, $np > 5$ and $n(1-p) > 5$. The confidence interval for a FAR estimate may be approximated as $\pm Z_C\sqrt{x}/A$. This is thought to be adequate for $n > 30$. The appropriateness of these conditions is considered in Section 5.2.2.

As examples, the following calls within MATLAB produce the indicated output. In the first example, the default values of *method* = 2 and *verbose* = 0 are used.

```
≫prop_ci(90,100,0.05)
ans=0.9000    0.8313    0.9485
≫rate_ci(10,50,0.05,4,1)
r_hat, Lower CI Bound, Upper CI Bound, x, A, Desired alpha, Method, Length, Lower Tail,
Upper, Tail, Actual alpha, Delta alpha, Run Time
ans=0.2000 0.1098 0.3678 10.0000 50.0000 0.0500 4 0.2579 0.0250 0.0250 0.0500 −0.0000
0.8600
```

5. CIE validation and performance comparison

5.1. Introduction

We claim improved accuracy for the prop_ci and rate_ci CIEs. This section attempts to support that claim and validate the theory and software coding by empirical tests. In addition to accuracy, there are performance issues involving CI length and runtime. These issues are addressed in this section as well.

5.2. Accuracy

Attempting to complement previous results [8], we will consider the accuracy of the CIEs with respect to α. Is the chance that the CI does not enclose the true p equal to the desired α or not? This is investigated by first establishing that "the chance that the Cl does not enclose the true p" may be computed by integrating the posterior distribution. That is, we first claim that the actual α associated with a given CI may be computed by integrating the posterior distribution, i.e., $\alpha = \int_0^a f_{\theta|x}(\theta)\,\mathrm{d}\theta + \int_b^1 f_{\theta|x}(\theta)\,\mathrm{d}\theta$, where a is the lower CI limit, b the upper, and $f_{\theta|x}(\theta)$ the posterior distribution. Although this is true by definition, the IBP CIEs and this test method have similar dependencies on the conceptual framework, the mathematical derivations, and the implementation of the distribution integrations. Therefore, we first do a Monte Carlo test of IBP that is free of many of these dependencies. Separate tests are conducted for proportion and rate CIEs. Once we are satisfied with the legitimacy of our implementation of the posterior integration, we use that implementation to directly compare the α realized by a given method with the desired α. The difference between these two is α_{error}.

5.2.1. Validation of the α_{error} estimation method

As a first step to testing the CIEs, we validate the approach used to estimate α_{error}. The following notation will be used for the various "α's". The desired α, as would be input to a CIE, is α_{desired}. The true α that results from a given CIE i over some range of conditions is α_i. The estimation of α_i by integrating the posterior is α_i-IBP. The α_i estimated from a Monte Carlo test is α_i-MC. Since α_i-MC is a statistically estimated proportion, we are also interested in confidence intervals about that estimate. Whenever such confidence intervals are reported for α_i-MC they are computed by method 2 using a confidence level of 0.95.

Our accuracy assessment is in terms of $\alpha_{\text{error}} = \alpha_{\text{desired}} - \alpha_i$. The question is, "Can we use α_i-IBP for α_i?" Not having direct access to α_i we must estimate it by α_i-MC and then answer the question by comparing actual α_i-MC's with α_i-IBP's. If the α_i-IBP's are the same as the α_i-MC's then α_i-IBP can be used for α_i because we are satisfied that α_i-MC is a good sampling theoretic estimate of α_i, "good" especially in the sense that it is independent of IBP CIE assumptions and implementations. In summary, $\alpha_{\text{error}} = \alpha_{\text{desired}} - \alpha_i$ is a measure of a CIE's accuracy. The α_i approximated by α_i-IBP is easily computed, but not immediately trustworthy. The α_i approximated by α_i-MC is trustworthy, but awkward to compute. We will show that α_i-IBP and α_i-MC are practically the same. We can then confidently use, in Section 5.2.2, the simple α_i-IBP in place of the awkward α_i-MC as α_i in the error measurement.

Overall: Monte Carlo alpha = 0.49555, with CI (0.0490, 0.0501)

n	Monte Carlo alpha	Lower CI	Upper CI
5	0.0567	0.0466	0.0680
10	0.0488	0.0418	0.0565
20	0.0544	0.0490	0.0602
50	0.0498	0.0464	0.0533
100	0.0488	0.0464	0.0513
200	0.0491	0.0474	0.0508
500	0.0489	0.0478	0.0500
1000	0.0499	0.0491	0.0507

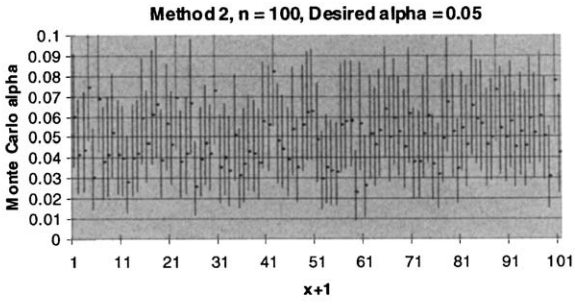

Fig. 3. Monte Carlo results for proportion CIs.

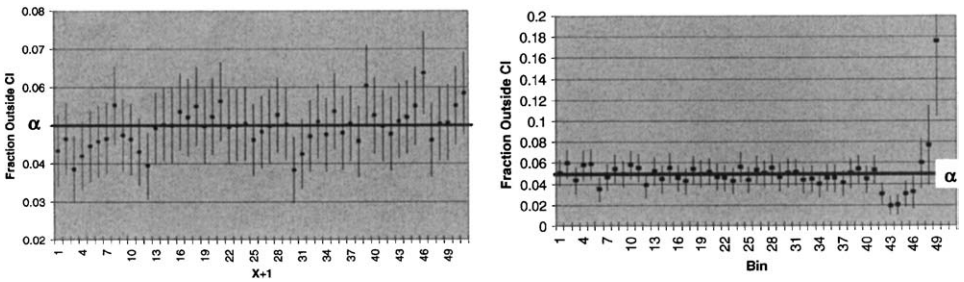

Fig. 4. Monte Carlo test results for proportion (left) and rate (right) CIs.

The question for proportion CIs is, "If for a given n, a randomly chosen p, and a randomly chosen x, will the computed CIs enclose the chosen p with a probability of $1-\alpha$?" So, for a given n, we repeatedly chose p from a uniform distribution on [0,1], randomly chose x from a binomial distribution with parameters (n, p), computed CIs (using method 2) with inputs (x, n, α), and then checked to see if the resulting CI enclosed p. The fraction of the time that the CI did not enclose p is the "Monte Carlo α" ($\alpha_{2\text{-MC}}$). Results were accumulated overall, by n, and by x. The values of n tested were those indicated in the table in Fig. 3. We attempted to get about 300 trials per x by running $300(n+1)$ trials for each n. There were a total of 567,900 trials. Of these, 28,142 did not include the true p value within the computed confidence interval. Therefore, overall $\alpha_{2\text{-MC}} = 0.049555$, compared to $\alpha_{2\text{-IBP}} = 0.05$. The 95% CI about $\alpha_{2\text{-MC}}$ is (0.0490, 0.0501), which contains $\alpha_{2\text{-IBP}}$ as would be expected for an accurate $\alpha_{2\text{-IBP}}$. The table in Fig. 3 contains the $\alpha_{2\text{-MC}}$'s for each n. It happens that in all cases the CIs enclose the desired α. This might be expected since there are 8 cases and there is only a 1-in-20 chance of a CI missing. Fig. 3 also plots the $\alpha_{2\text{-MC}}$ for each x at $n = 100$. There are four x's where the CIs do not enclose $\alpha_{2\text{-IBP}}$ (26, 43, 60, and 99). With 101 cases, that four CIs miss is also consistent with the confidence level. Therefore, overall and as a function of n, $\alpha_{2\text{-IBP}}$ may be trusted as an estimate of α_2 in assessing the accuracy of CIEs.

Monte Carlo tests were performed at other values of n and on other methods with similar results. For example, the Monte Carlo results for method 3 at $n = 50$ are shown in the left plot of Fig. 4.

With the 0.95 confidence level CIs and 51x's, we would expect, on average, that 2.55 of the CIs would not include α. There are actually 4 such cases, one where the fraction outside is too high (at $x = 45$) and three that are too low (at $x = 2$, 11, and 29). This is consistent with an accurate $\alpha_{3\text{-IBP}}$. We therefore trust the $\alpha_{i\text{-IBP}}$ values for proportions.

The Monte Carlo test for rate estimation is as follows. We fix $A = 1.0\,\text{km}^2$ (although A and its units are immaterial here) and $\alpha_{\text{desired}} = 0.05$. We then generate a random λ from the uniform distribution on [0,458]. The upper limit for λ was chosen to make the maximum x that we would likely see around 500. We then generated a random x from the Poisson distribution with parameter λ. For each x we computed an IBP CI and recorded whether the CI covered the true λ. This process was repeated 50,000 times. The x values were then grouped into 51 bins, that is 0 to 9, 10 to 19, 20 to 29, ..., 490 to 499, and 500 or larger. For each bin, the fraction of times that λ was outside the CI was computed along with the 0.95 confidence level CI and plotted on the right in Fig. 4. Note that the "fraction outside CI" diverges from the desired 0.05 with bin 42 and larger (i.e., for x's greater than about 410). This is an artifact of the Monte Carlo test methodology, particularly our arbitrarily choosing a maximum λ. To explain the dip down, consider bin 44. Bin 44 includes x values around 435. The maximum λ is 458. One standard deviation at $x = 435$ is around 21, so the CI cannot fail to cover λ because λ is too large. It can only fail on the other side, which it apparently does at a rate about half that of the 0.05 value. To explain the large "fraction outside CI" for the last few bins, consider bin 51. Bin 51 only has x values greater than 500, but the maximum λ is 458, so the CIs necessarily fail to cover the true λ (the actual fraction is 1.0). Therefore, we dismiss the Monte Carlo results for the last dozen or so bins. If we really wanted to see the performance for the x values in those bins, we would simply need to repeat the experiment with a larger maximum λ. With the 0.95 confidence level CIs and 40 meaningful bins, we would expect that about two of the CIs to miss α. There is actually one such case (bin 6); we therefore trust the $\alpha_{i\text{-IBP}}$ values for rates.

In summary, our implementation of IBP yields α values ($\alpha_{i\text{-IBP}}$) that are consistent with the Monte Carlo α values ($\alpha_{i\text{-MC}}$). Although assessing accuracy in terms of $\alpha_{i\text{-MC}}$ has attractions (particularly an independence from non-trivial assumptions and implementation details), $\alpha_{i\text{-IBP}}$ is more easily computed and we do not have to worry with the sampling error inherent in Monte Carlo methods. Therefore, we will use the foregoing Monte Carlo validation of $\alpha_{i\text{-IBP}}$ as justification for using $\alpha_{i\text{-IBP}}$ in our accuracy assessment below.

5.2.2. *Application of the α_{error} measure in assessing CIE accuracy*

Our accuracy assessment is in terms of α_{error}. In principle, we could have used the Monte Carlo method for all of our testing, i.e., $\alpha_{\text{error}} = \alpha_{\text{desired}}\text{-}\alpha_{i\text{-MC}}$, but it is more efficient and, as argued in the previous section, as effective to use direct integration, i.e., $\alpha_{\text{error}} = \alpha_{\text{desired}} - \alpha_{i\text{-IBP}}$. Figs. 5–7 plot $\alpha_{\text{error}} = \alpha_{\text{desired}} - \alpha_{i\text{-IBP}}$ for proportion CIs. All of the plots on the left use the same z-axis scale. The plots on the right use varying z-axis scales. The n axis uses a log scale and runs from 1 to 10,000. The x/n axis runs from 0 to 1. The errors are computed on an array of 21 by 21 $(n, x/n)$ pairs. The lines simply connect the points for visualization and do not represent actual data. Since α_{desired} is 0.05, the maximum possible positive error is 0.05.

Fig. 5 is the plot for method 3, which is used as representative of all IBP methods (i.e., methods 1–4). The error is generally less than 10^{-5} and is roughly as likely to be positive as negative. There

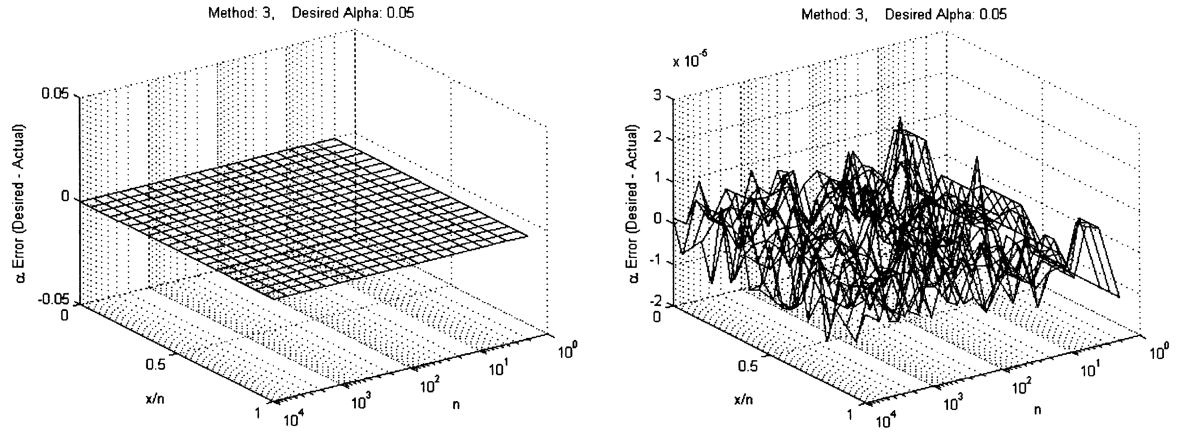

Fig. 5. Proportions method 3 α_{error}, -0.05–0.05 scale (left) -2×10^{-5}–3×10^{-5} (right).

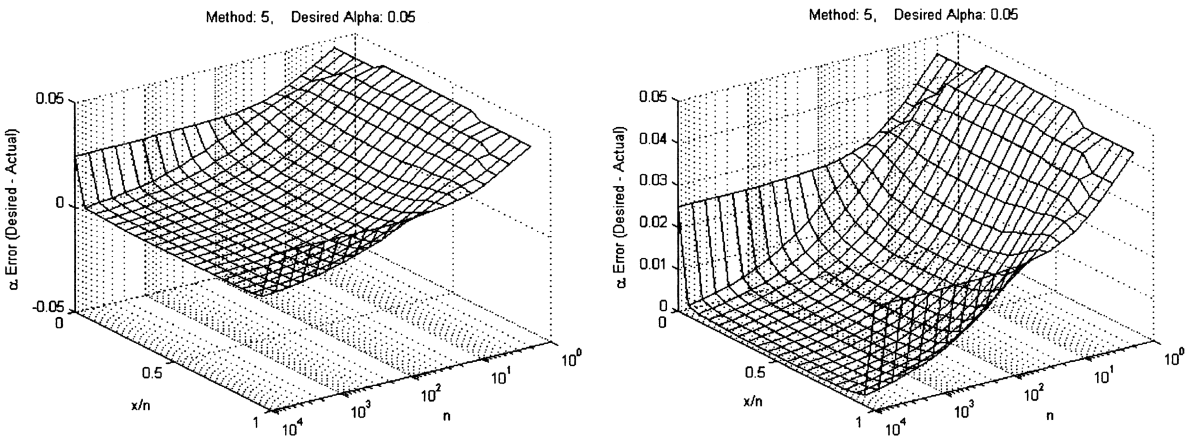

Fig. 6. Proportions method 5 α_{error}, -0.05–0.05 scale (left) 0–0.05 (right).

is no apparent pattern to the errors. The results for the other IBP methods (1, 2, and 4) were all similar to method 3's (same scale and largely random distribution of errors). This demonstrates that the IBP methods, as coded here, produce sufficiently accurate CIEs that the corresponding α's are generally within 10^{-5} of the desired value.

Fig. 6 is the error plot for method 5 (Clopper–Pearson). Method 5 is sometimes considered to be "exact" [5,7] and has been used when accurate CIs were important. The error is generally on the order of 0.01 (20%) or greater and is greater than 0.03 (60%) for some cases of interest in classifier performance assessment. The error is greatest at small n (less than 100) or x close to zero or n. As expected, the error is never negative, so the method 5 CIs are always conservative.

Fig. 7 is the error plot for method 6 (Gaussian Approximation). Method 6 is perhaps the most commonly applied CIE in classifier testing. On the left plot, error values less than -0.05 are truncated to that value. Method 6 produces zero width CIs at $x = 0$ or n, which makes $\alpha_6 = 1.0$. The error

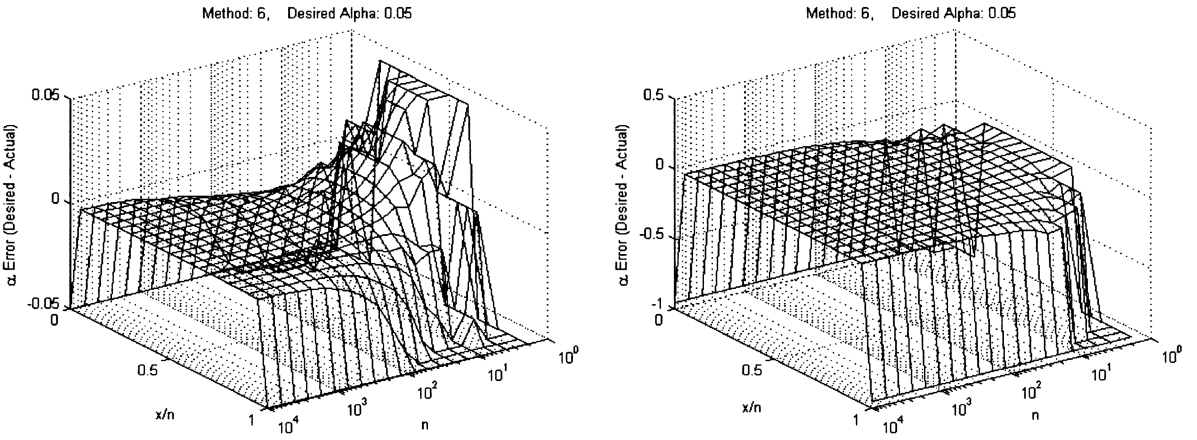

Fig. 7. Proportions method 6 α_{error}, −0.05–0.05 scale (left) −1.0–0.5 (right).

may be several times larger than the desired α. The error is greatest for small n, particularly when x is close to zero or n. The error tends to be negative for small or large x's, meaning the CIs are understated, so confidence may be assumed when it is not appropriate. This is the worst form of error for many applications.

In summary, one can see by scanning the left hand plots that the IBP CIEs are significantly more accurate than conventional approaches. Importantly, they are accurate without conditioning on a vague "*large n*" or "*x* not *too* small or *too* large." Method 6, which is commonly used in classifier performance assessment, has substantial errors in the CI for small n or, at any n, when x/n is far from 0.5. The common rules of thumb for the applicability of a normal approximation, i.e., $n > 30$, $np > 5$ and $n(1-p) > 5$ are not adequate constraints for comparable accuracy, e.g., at $\alpha = 0.05$, we only recommend using method 6 for $n > 600$ and $0.2 \leqslant x/n \leqslant 0.8$. For $\alpha = 0.01$, the minimum x is 2000. Method 5 avoids understating the uncertainty, but it appreciably overstates it for small n or large or small x.

We now consider the accuracy of the various CIE methods for rates. The measures reported here and in the CI length section do not depend on the area (A), since the CI limits in terms of λ are simply divided by A. The runs below were all made with $A = 1$. Fig. 8 plots α_{error} for methods 2, 5, and 6 for $\alpha_{\text{desired}} = 0.05$ on the left. The MATLAB implementation of method 5 switches to the normal approximation at $x \geqslant 100$, as can be seen in the error plots. Method 6 returns zero as the lower and upper limits when $x = 0$. Method 2's accuracy is representative of that of the other IBP methods (1, 3 and 4). Although not evident from the plot, the errors for the rate IBP methods are comparable to that of the proportion IBP methods, i.e., on the order of 10^{-5} and with no particular pattern. The maximum error for both method 5 and 6 is at the smallest x. Method 6's error is larger and of the worst kind (reflecting an understatement of uncertainty). Method 5's error is about 50% of the desired α for $x = 1$. While n's in the single digits are rarely the basis for proportion estimates, x's in the single digits may well be of interest in rate estimation.

We have generally reported on CIE performance at $\alpha_{\text{desired}} = 0.05$. The above testing was also performed at $\alpha_{\text{desired}} = 0.1$ and $\alpha_{\text{desired}} = 0.01$ (Fig. 8 right); for all three values, the percentage α_{error} is about the same.

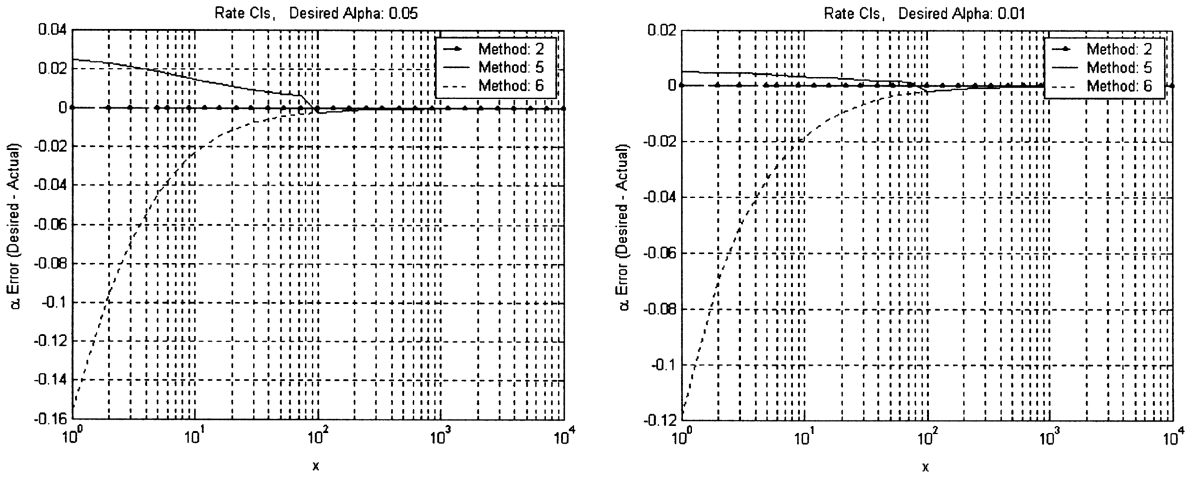

Fig. 8. α_{error} for rate estimation.

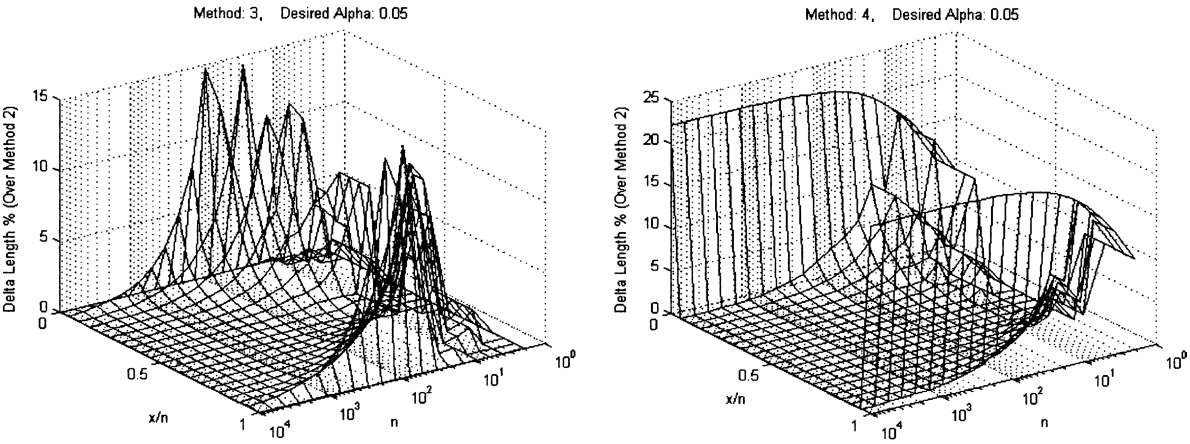

Fig. 9. Delta length proportions method 3 (left) and method 4 (right).

5.3. CI length

Another consideration in assessing a CIE is the "tightness" of the intervals for a given accurate α. We measure tightness as the length of the interval. Shorter intervals more tightly bound an estimate and are often preferred in our applications. Of the IBP CIEs, Method 2 is minimal length by design, but there are applications where methods 3 and 4 are of interest, so we characterize the compromise these methods make in tightness.

We compare the CI lengths of the three accurate two-sided proportion CIEs, methods 2–4 in Fig. 9. The n and x/n axes are as before. Method 2 is designed to produce an interval of minimal length, so we compared methods 3 and 4 to method 2. The z-axis is the percentage that a given

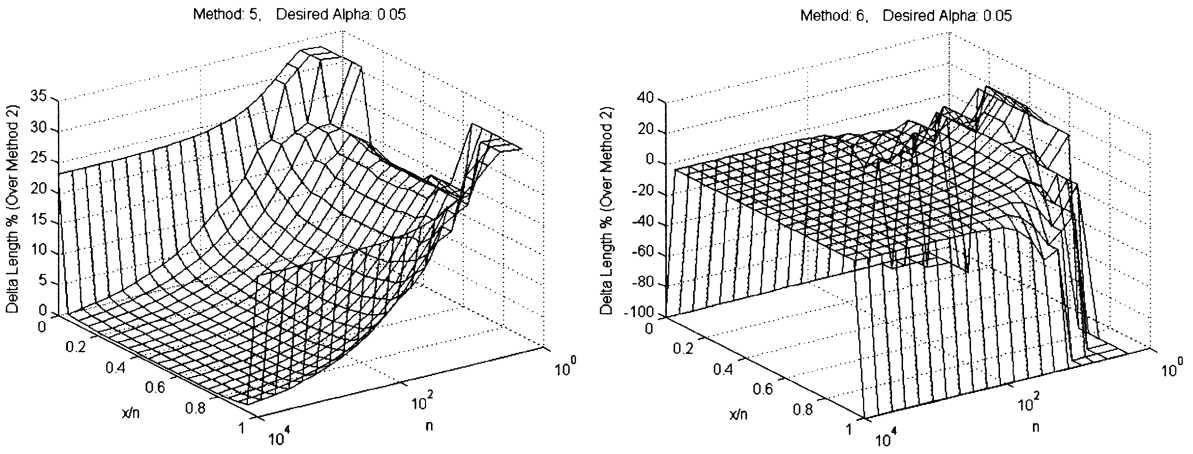

Fig. 10. Delta length proportions method 5 (left) and method 6 (right).

method's CI length is longer than method 2's ($100 \times$(method i's length − method 2's length)/method 2's length). For large n or medium x, the lengths are all roughly the same. This occurs when the posterior distribution is more symmetric or more gently sloping. Method 3 reverts to asymmetrical CIs, approaching method 2 CIs, when the CI would extend past zero or one. This explains method 3 having shorter lengths for smaller n and extreme x values. We suspect that the ridge apparent in method 3's relative length up through n's of a few hundred continues for larger n, it is just too narrow for our coarse sampling. Method 4 satisfies the balanced-tail criteria, even when that means the CI will not include the estimated value. When x is near zero or n, the CI must be lengthened significantly.

Fig. 10 plots methods 5 and 6's CI lengths relative to method 2's. There are two things causing method 5's lengths to be longer than method 2's. First, method 5 is actually providing a smaller α, which makes CIs longer. Second, it is not optimized for minimal-length, even for the α it is providing. If method 5 were accurate and with balanced-tails (as intended) its lengths would be the same as method 4's. Method 6's CIs are balanced-width, so for the correct α, they would behave the same as method 3's. Of course, method 6's CIs are shorter than method 2's only because their actual α is larger than desired. In summary, the other methods produce CIs up to 20% larger than method 2's for small n (less than a few hundred) and large or small x.

As with the proportion CIEs, the relative lengths of the various accurate rate CIEs are compared in Fig. 11 for $\alpha_{\text{desired}} = 0.05$. All methods produce the same CI for large x. Method 3's lengths are relatively small as its lower CI limit bumps up against zero, where it is also no longer balanced-width. Method 5 has both the inaccuracy in α and non-optimized length working against it. When they both provide the same actual α, methods 4 and 5 have similar lengths. The step down in length for method 5 at $x = 100$ is due to the MATLAB implementation switching to the normal approximation at that point. Again, method 6's CI's are balanced-width, therefore would have lengths comparable to method 3's for a given actual α.

In summary, for conditions of interest in classifier testing, Method 2 provides accurate and tight CIs and is used as the default method.

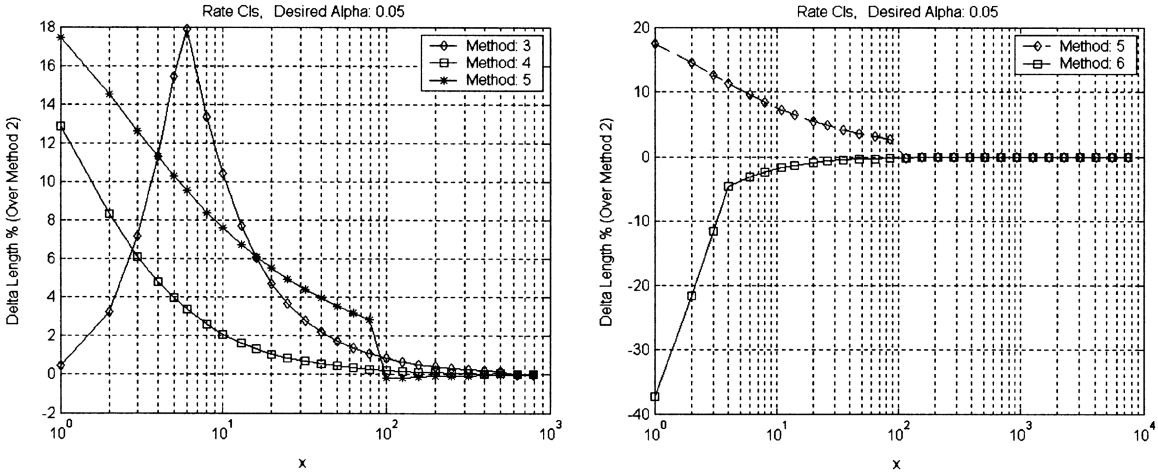

Fig. 11. Delta length for rate CIs.

5.4. Computational requirements

A final consideration in CIE selection concerns computational requirements. The run times for method 2, over all x, for n between 5 and 1000 averaged about 1 s (ranging from 0.5 to 2.5 s) on an 800 MHz Pentium III PC running interpreted MATLAB code. Method 3 has similar run times. Methods 1 and 4 take about twice as long. Methods 5 and 6 take virtually no time. The rate IBP CIEs are about twice as fast as the proportion IBP CIEs. The functions prop_ci and rate_ci are available as MATLAB source code, which may be used directly or as psuedo-code for implementation in other languages. Using the source code as is requires MATLAB Release 13 with the Statistics Toolbox. The source code was developed for multiple purposes (including non-conjugate priors); if this were not of interest, a simpler and quicker implementation is possible. For a proportion estimate, α is betacdf$(a, x+1, n-x+1) + (1\text{-betacdf}(b, x+1, n-x+1))$ and for a rate estimate, gamcdf$(a, x+1, 1) + (1\text{-gamcdf}(b, x+1, 1))$.

The functions betacdf and gamcdf are as used in the MATLAB Statistics Toolbox; returning the cdf value with the first argument being the independent variable and the other two arguments being parameters of their respective distributions. A binary search over a and b to get a desired $\alpha_{i\text{-IBP}}$, possibly with other constraints such as balanced-tails, is all that is needed to implement the CIE.

6. Estimates of differences

If we are simply interested in a performance measure's value, a CI tells us how well we know that value. However, we may be comparing two performance measures. Is one classifier performing better than another? Is one set of test data harder than another? Was a particular classifier design change an improvement? To answer such questions, we are interested in CI's for estimates of parameter differences [13], i.e., CIEs for $p_1 - p_2$ and $r_1 - r_2$. In principle, a difference CIE could be developed based directly on the posterior distribution and a search for integration limits. However, we here

first develop a tool that computes "significance of differences", e.g., $\Pr\{p_1 - p_2 > \delta\}$ and then use that tool for computing CIs. The significance of differences tool may be of interest in its own right and it made computation of the difference CIs more convenient. The difference CIEs (prop_diff_ci and rate_diff_ci) are much slower, taking tens of seconds in some cases, and the comparison with the Normal and Clopper–Pearson methods is limited to the results reported in [13].

6.1. Significance of differences

Suppose we have two sets of test results. We assume that each set of results is independently produced by a distribution (binomial-based for proportions and Poisson-based for rates). At issue is whether the parameters of the two distributions are significantly different.

For independent random variables $X_1 \propto f_{x_1}(x_1)$ and $X_2 \propto f_{x_2}(x_2)$ with some difference of interest δ, $\Pr(x_1 - x_2 \geqslant \delta) = \int_{-\infty}^{\infty} f_{x_1}(\theta) F_{x_2}(\theta - \delta)\,\mathrm{d}\theta$, where $F_{x_2}(x_2)$ is the cdf associated with pdf $f_{x_2}(x_2)$. This relationship may be developed from the properties of definite integrals and that the pdf of a sum of two independent random variables is the convolution of their pdfs, i.e.,

$$f_{x_1-x_2}(\tau) = \int_{-\infty}^{\infty} f_{x_1}(\theta) f_{-x_2}(\tau - \theta)\,\mathrm{d}\theta = \int_{-\infty}^{\infty} f_{x_1}(\theta) f_{x_2}(\theta - \tau)\,\mathrm{d}\theta$$

and then

$$\begin{aligned}
\Pr(x_1 - x_2 \geqslant \delta) &= \int_{\delta}^{\infty} f_{x_1-x_2}(\tau)\,\mathrm{d}\tau = \int_{\delta}^{\infty} \int_{-\infty}^{\infty} f_{x_1}(\theta) f_{x_2}(\theta - \tau)\,\mathrm{d}\theta\,\mathrm{d}\tau \\
&= \int_{-\infty}^{\infty} \int_{\delta}^{\infty} f_{x_1}(\theta) f_{x_2}(\theta - \tau)\,\mathrm{d}\tau\,\mathrm{d}\theta \\
&= \int_{-\infty}^{\infty} f_{x_1}(\theta) \int_{\delta-\theta}^{\infty} f_{x_2}(-\tau)\,\mathrm{d}\tau\,\mathrm{d}\theta = \int_{-\infty}^{\infty} f_{x_1}(\theta) \int_{-\infty}^{-\delta+\theta} f_{x_2}(\tau)\,\mathrm{d}\tau\,\mathrm{d}\theta \\
&= \int_{-\infty}^{\infty} f_{x_1}(\theta) F_{x_2}(\theta - \delta)\,\mathrm{d}\theta.
\end{aligned}$$

The integrand $f_{x_1}(\theta)F_{x_2}(\theta - \delta)$ is significantly greater than zero only in some finite range. The finite range must be provided to the numerical integration routine. The difficulty with selecting these integration limits was the principal factor in limiting the range of acceptable input values ($x \leqslant 10^5$ and $\alpha \geqslant 10^{-4}$) for the tools provided [9]. If the limits were not sufficiently tight around a range with non-zero integrand values, the numerical integration routine would fail. The lower integration limit was based on the larger of several standard deviations less than f_{x_1}'s mean, several standard deviations less than f_{x_2}'s mean, 0 (all the distributions of interest are non-zero only for non-negative arguments), and delta (which effectively shifts f_{x_2}'s starting point). The upper integration limit is set at several standard deviations above f_{x_1}'s mean. For the case of proportions only, the upper limit is also limited to 1.0. The f_{x_2} distribution does not affect the upper limit because it enters the integrand as a cdf.

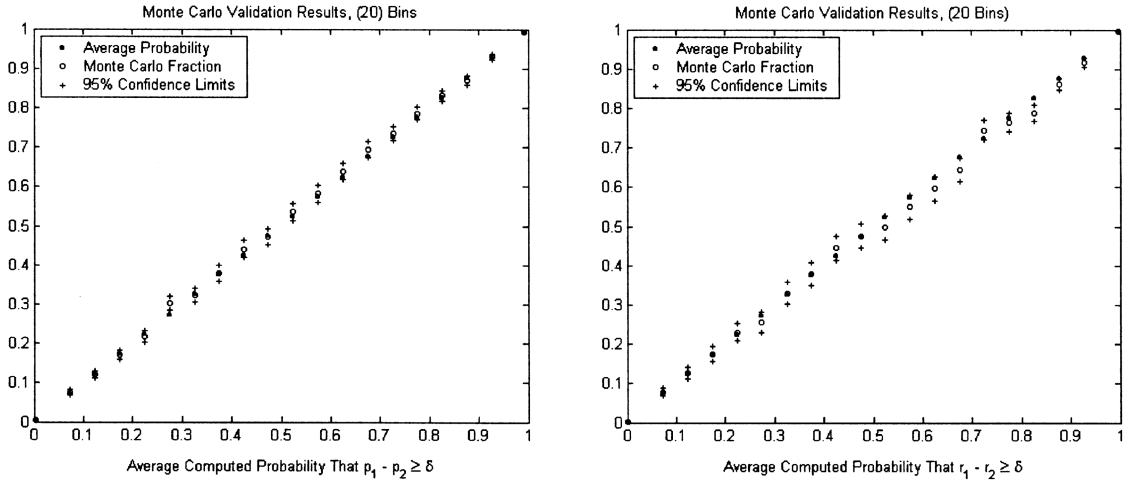

Fig. 12. Monte Carlo test results—proportions (left) and rates (right).

6.1.1. Proportions differences

As previously developed, the posterior distribution for proportions with a diffuse prior is

$$f_{p|x}(p\,|\,x) = \begin{cases} (n+1)\binom{n}{x} p^x(1-p)^{n-x}, & p \in [0,1], x = 0,1,2,\ldots,n, \\ 0 & \text{otherwise.} \end{cases}$$

We are interested in computing $\Pr(p_1 - p_2 \geqslant \delta) = \int_{-\infty}^{\infty} f_{p_1|x_1}(\theta) F_{p_2|x_2}(\theta - \delta)\,\mathrm{d}\theta$. In MATLAB code this is simply the integral over t of betapdf$(t, a_1, b_1).\times$betacdf$(t\text{-delta}, a_2, b_2)$, where $a_1 = x_1 + 1, b_1 = n_1 - x_1 + 1, a_2 = x_2 + 1$, and $b_2 = n_2 - x_2 + 1$. As in previous sections, x_1 is the number of positive outcomes in n_1 trials from a population whose binomial parameter is p_1 (x_2, n_2, and p_2 are similarly defined for the second population). The code for this, as prop_diff$(x_1, n_1, x_2, n_2, \text{delta})$, returning $\Pr(p_1 - p_2 \geqslant \delta)$, is available on the web site [9]. If we are interested in whether $p_1 \geqslant p_2$, we can answer that by computing the probability with $\delta = 0$. If we are interested in whether the difference is likely to be greater than some threshold then we simply specify δ as that threshold. We may also search for the δ that yields a certain desired probability.

A Monte Carlo validation of prop_diff was performed using 200,000 trials. For each trial, p_1 and p_2 were randomly selected from uniform[0,1] (i.e., a uniform distribution on [0,1]). Samples sizes, n_1 and n_2 were then randomly selected from uniform$[1, N]$, where N was set to 100. If N is too large, prob_diff tends to be close to zero or one always. We then randomly pick x_1 and x_2 from binomial(p_1, n_1) and binomial(p_2, n_2), respectively. A δ was then randomly selected from uniform$[-1, 1]$. Finally, prop_diff was called with $(x_1, n_1, x_2, n_2, \text{delta})$ and all parameters and the result recorded. After all trials were complete, the results were grouped in 20 bins based on prop_diff's output (i.e., the computed $\Pr(p_1 - p_2 \geqslant \delta)$). For each bin, the fraction of the time that in fact $p_1 - p_2 \geqslant \delta$ was computed. Ideally, the computed probability and the fraction would be close. Fig. 12 (left) shows the results. Solid dots are the average (within a bin) of the computed

probabilities (i.e., $\Pr(p_1 - p_2 \geqslant \delta)$), circles are the fraction of the time that in fact $p_1 - p_2 \geqslant \delta$, pluses are the 95% CIs for the circles, as estimates of the true probabilities. Note that all but one (exactly 95%) of the CIs contain the solid dots; so we are inclined to trust prop_diff.

6.1.2. Rate differences

The posterior distribution for the Poisson λ parameter is

$$f_{\lambda|x}(\lambda \,|\, x) = \begin{cases} \dfrac{\lambda^x \mathrm{e}^{\lambda}}{x!}, & \lambda \in [0, \infty], x = 0, 1, 2, \ldots, \\ 0 & \text{otherwise,} \end{cases}$$

where x is the actual number of events realized. Let r be the random variable associated with a rate, i.e., $r = (1/A)\lambda$, where A is a constant area. The posterior pdf for the rate is then

$$f_{r|x,A}(r \,|\, x, A) = \begin{cases} \dfrac{(Ar)^x \mathrm{e}^{-Ar}}{x!}, & r \in [0, \infty]; x = 0, 1, 2, \ldots; A > 0, \\ 0 & \text{otherwise.} \end{cases}$$

We are interested in $\Pr(r_1 - r_2 \geqslant \delta) = \int_{-\infty}^{\infty} f_{r_1|x_1,A_1}(\theta) F_{r_2|x_2,A_2}(\theta - \delta)\,\mathrm{d}\theta$. The second term in our probability expression is a cdf; however it is awkward to deal with the cdf corresponding to this pdf for non-unit areas, so the program converts to normalized units such that A_2 is one. That is, we use $A_1' = A_1/A_2$, $A_2' = 1.0$ and $\delta' = \delta \times A_2$.

$\Pr(r_1 - r_2 \geqslant \delta)$ can be computed in MATLAB as the integral over t of $(A_1/A_2) \times \mathrm{gampdf}(t \times (A_1/A_2), x_1 + 1, 1) \times \mathrm{gamcdf}((t\text{-delta} \times A_2), x_2 + 1, 1)$. This computation was also Monte Carlo tested with the results shown in Fig. 12 (right). Although not significant in this test, there is a systematic bias caused by our having an upper limit when generating r_1 and r_2 samples. We take this as another reflection of limitations in Monte Carlo tests involving rates, as in Section 5.2.1, rather than as an indication of inaccurate difference significances.

The user may note the following property of difference significances. Using rate differences as an example, suppose the first test is on a small area, say 1.0 km^2, and the second test is on a larger area, say 100 km^2. In both tests there are zero events. Although $\hat{r}_1 = \hat{r}_2 = 0$, the posterior for r_1 and r_2 are different and $\Pr\{r_1 > r_2\}$ is *not* 0.5. In fact, $\Pr\{r_1 > r_2\} = 0.99$. In the example, we are confident that r_2 is small (we had zero events in a large test area) while we are quite uncertain about r_1 (a small test area). Since r_1 could be any number of values (all greater than zero) while r_2 is definitely close to zero, we may reasonably expect r_1 to be greater than r_2.

6.2. Confidence intervals

Confidence intervals are related to the significance of difference as follows, using proportions in the development that applies similarly to rates. The parameter of interest is $\Delta p = p_1 - p_2$. The confidence interval $[a, b]$ of interest is, by definition, $a, b, \ni \Pr\{\Delta p \in [a, b] \,|\, x_1, x_2\} = 1 - \alpha$. This expression is related to the quantity computed by the tools above, i.e.,

$\Pr\{\Delta p \in [a, b]\} = \Pr\{\Delta p \geqslant a \cap \Delta p \leqslant b\} = \Pr\{\Delta p \geqslant a\} + \Pr\{\Delta p \leqslant b\} - 1 = \Pr\{\Delta p \geqslant a\} - \Pr\{\Delta p \geqslant b\}$, so $\alpha = 1 - \Pr\{\Delta p \geqslant a\} + \Pr\{\Delta p \geqslant b\}$. As when directly integrating the Bayesian

posterior, we search over a and b for the desired α, where the lower tail contribution to α is $1 - \Pr\{\Delta p \geqslant a\}$ and the upper tail contribution is $\Pr\{\Delta p \geqslant b\}$.

The confidence interval tools prop_diff_ci and rate_diff_ci are implemented, based on the above development, and included in the posted tool set [9]. These CIEs have functionality similar to prop_ci and rate_ci, with the additional arguments for x_2, n_2 (or x_2, A_2), except method 2 is not implemented. The implementations of method 5 (Clopper–Pearson based) and method 6 (Normal approximation based) are those of [13] for proportion differences. Methods 5 and 6 are not implemented for rate differences. Note that the correction factor suggestion in [13] and used in method 5 is specifically for $\alpha = 0.05$ and may not be ideal for other α values. One example reported in [13] is for x_1, n_1, x_2, n_2 and α as 5, 12, 36, 112, and 0.05, respectively. The method 5 interval is -0.1878 to 0.4151 with coverage about 97.91%. The method 3 interval is -0.1665 to 0.3570 with the desired 95.0% coverage. Extensive Monte Carlo tests of method 5 reported in [13] demonstrate that "The average coverage using the method described here for differences in proportions gave identical coverage to that computed for single binomial proportions with the same denominators. When unequal denominators were used, the coverage was intermediate between that expected for the individual denominator sizes." [13, pp. 85–86]. The term "coverage" as used in [13] is the $1-\alpha_i$ of Section 5 above. Therefore, the single parameter CIE method comparisons of Section 5 above are likely to be indicative of the relative performance of IBP and other CIEs for differences. In particular, the previously demonstrated accuracy advantages for IBP methods in proportion and rate estimation are likely to apply to their difference CI's as well.

7. Conclusions

Proportion and rate estimates involve discrete distributions and may be based on a small number of samples n and/or extreme numbers of events x (e.g., x small or near n). The "integration of the Bayesian posterior distribution" (IBP) based confidence interval (CI) Estimators (CIEs) are substantially more accurate than conventional methods under these conditions. Both the IBP and conventional CIEs described in this paper are available as MATLAB code [9], which may be used directly or as pseudo-code for other languages. The IBP methods (1–4) search for CI limits (a, b) such that the integral of the posterior distribution from a to b is $1-\alpha$ while satisfying the other criteria (1 or 2-sided, balanced tail/width, minimal-length). The posterior distribution used for proportions is a beta distribution with parameters $\alpha = x+1$ and $\beta = n-x+1$, i.e., $f_{p|x}(p\,|\,x) = (n+1)\binom{n}{x}p^x(1-p)^{(n-x)}$, $0 < p < 1$. The posterior for rates is a gamma distribution with parameters $\alpha = x+1$ and $\beta = 1$, i.e., $f_{\lambda|x}(\lambda\,|\,x) = \lambda^x e^{-\lambda}/x!$ $0 < \lambda$, $x = 0, 1, 2, \ldots, n$. The IBP methods are easy to implement with fourth-generation mathematical programming languages, such as MATLAB, especially if they provide cdf functions for the beta and gamma distributions. Tools are also provided for computing the significance of differences between proportions and rates and confidence intervals for such differences. These tools allow precise statements about the relationship between two estimates.

Accuracy of the CIEs is assessed in terms of the error in α, i.e., the difference between the desired α and the true α for the returned CI. The errors in α for the IBP CIEs (methods 1–4), for both proportions and rates, are less than 5×10^{-5} (0.1% for $\alpha_{\text{desired}} = 0.05$) across all tested conditions. Method 5 (Clopper–Pearson, sometimes considered to be "exact") has an error in α that is on the order of 20% for difficult conditions and is greater than 60% for some cases. Method

6 (normal approximation, perhaps the most commonly applied CIE in classifier testing) had errors in α that were several times the desired α. The Method 6 errors tend to be negative for small or large x's, meaning the CIs are understated, so confidence may be assumed when it is not appropriate. That method 6 has regions of poor accuracy is well known, although the validation tests here help identify the extent of those regions. Method 5's accuracy limitations are preferable to method 6's, but are still quite substantial in our region of interest. The IBP methods are accurate throughout.

CIs may be desired with additional properties, such as minimal-length, balanced-tail, or balanced-width. The IBP CIEs provide CIs consistent with the appropriate additional criteria (method 2—minimal-length, method 3—balanced-width, method 4—balanced-tail), except method 3 compromises width balance when limits would be unreasonable (e.g., less than zero). The different criteria result in the same CIs for large n and moderate x values. Where the CIs differ, the other methods have CIs 10% or so longer than Method 2's. Another way to look at the benefits of method 2 (the minimal-length IBP CIE) is by considering the number of additional samples required for a given CI width. Under conditions common in classifier performance assessment, method 2 requires about 10% fewer samples than method 5 for a given proportion CI width. Alternatively, under those conditions, method 2's CIs run about 10% shorter than method 5's for a given desired α. Method 2 would require about 15% less test area than method 5 for a given CI width at $x = 1$ for rate estimation. Method 2's CI widths are about 15% less than that of method 5's at $x = 1$. These differences decrease with increasing x; at $x = 50$, both are about 3%.

The IBP CIEs are relatively slow. Proportion or rate CIEs may take a second or two to run on a PC with interpreted MATLAB code. The difference CIEs may take tens of seconds. Methods 5 and 6 are much faster. In many classifier applications, even with several seconds per CI, the time required to compute CIs is small compared to that of generating the direct test data. For a given desired α, it may be faster to do fewer trials with IBP CIEs than to do the greater number of trials dictated by the inaccurate methods. The IBP CIEs provide advantages for classifier performance assessment from accuracy, flexibility, and overall time consumption perspectives. Considering the cost of samples or of inconclusive results, the implementation investment necessary for accurate IBP CIEs can be justified. For those situations where IBP CIEs are not available, the test results of Sections 5.2.2 and 5.3 help characterize the performance of conventional approaches.

Although accurate statistical characterization of uncertainty is a useful first step, dealing with the non-statistical uncertainty due to nonrandom sampling, dependent training and test sets, or non-representative populations remains an important problem in many classifier evaluation efforts. In such cases, statistical CIs provide a lower bound on the overall uncertainties. Since statistical uncertainty may only be a significant contributor to the overall uncertainty when it is large, we are most interested in statistical uncertainty for small sample sizes. So accuracy in statistical CIs is especially important for small sample sizes—exactly where conventional approaches are inaccurate.

Tools are provided here that accurately compute confidence intervals for estimates of proportions, rates, and their differences as they arise in classifier performance assessment; however, future work in this area could allow for informative priors, larger input values (i.e., larger than the current limits of 10^5), smaller α values (i.e., smaller then 10^{-4}), implementations in other languages, faster runtimes, inclusion of a minimal-length CIE for differences, or inclusion of other types of estimates beyond proportions, rates and their differences.

8. Summary

("Accurate Confidence Intervals for Binomial Proportion and Poisson Rate Estimation", Timothy D. Ross)

This paper describes confidence interval estimators for the measures used in classifier evaluation. For the discrete distributions, small sample sizes, and extreme outcomes encountered within classifier testing, the commonly used confidence intervals have limited accuracy. This paper makes computational tools available for confidence intervals that are accurate over the full range of conditions of interest. The approach is to search for intervals using an integration of the Bayesian posterior to measure α (chance of the CI not containing the true value). Confidence intervals so computed are accurate in both the classical and Bayesian (assuming diffuse priors) settings. The programs provided include proportion estimates based on binomial distributions, rate estimates based on Poisson distributions, and their differences. One or two-sided CIs may be selected. For two-sided CIEs, either minimal-length, balanced-tail probabilities, or balanced-width may be selected. The CIEs' accuracies are reported based on a Monte Carlo validated integration of the posterior probability distribution and compared to the normal approximation and Clopper–Pearson methods. While the IBP methods are accurate throughout, the conventional methods may realize α's with substantial error (up to 50%). This translates to 10–15% error in the interval widths or to requiring 10–15% more samples for a given confidence level. Tools are also provided for computing the significance of differences between proportions and rates, e.g., $\Pr(p_1 - p_2 \geqslant \delta)$ for proportions, and confidence intervals for estimates of differences. Such tools allow more precise statements about the relationship between two performance measures. The tools (prop_ci, rate_ci, prop_diff, rate_diff, prop_diff_ci, and rate_diff_ci) are available on the MATLAB Central File Exchange web site. Although accurate statistical characterization of uncertainty is a useful first step, dealing with the non-statistical uncertainty due to nonrandom sampling, dependent training and test sets, or non-representative populations remains an important open problem in many classifier evaluation efforts. In such cases, statistical CIs provide a lower bound on the overall uncertainty. Since the contribution of statistical uncertainty may only be significant in the overall uncertainty when it is large, we are most interested in statistical uncertainty for small sample sizes. So accuracy in statistical CIs is especially important exactly where conventional approaches are inaccurate. Tools are provided here that accurately compute confidence intervals for estimates of proportions, rates, and their differences as they arise in classifier performance assessment.

References

[1] T.D. Ross, L.A. Westerkamp, R.L. Dilsavor, J.C. Mossing, Performance measures for summarizing confusion matrices—The AFRL COMPASE approach, Proceedings of SPIE, Vol. 4727, Algorithms for Synthetic Aperture Radar Imagery IX, April 2002.

[2] P.L. Meyer, Introductory Probability and Statistical Applications, Addison-Wesley, Reading, MA, 1970.

[3] R.L. Winkler, W.L. Hays, Statistics: Probability, Inference and Decision, Holt, Rinehart and Winston, New York, 1975.

[4] N.L. Johnson, S. Kotz, A.W. Kemp, Univariate Discrete Distributions, 2nd Edition, Wiley, New York, 1992.

[5] T. Fagan, Quick basic program for exact and mid-P-confidence intervals for binomial proportion, Comput. Biol. Med. 26 (1996) 263–267.

[6] J.C. Pezzullo, JavaStat web site: http://members.aol.com/johnp71/javastat.html, 2001.

[7] L. Daly, Simple SAS macros for the calculation of exact binomial and Poisson confidence limits, Comput. Biol. Med. 22 (1992) 351–361.
[8] D.J. Brenner, H. Quan, Exact confidence limits for binomial proportions—Pearson and Hartley revisited, The Statistician 39 (1990) 391–397.
[9] MATLAB Central File Exchange web site: http://www.mathworks.com/matlabcentral/fileexchange, 2003.
[10] J.A. Woodward, W.C. Liu, D.G. Bonett, Shortest two-tailed confidence intervals, Appl. Math. Comput. 84 (1) (1997) 65–76.
[11] C.J. Clopper, E.S. Pearson, The use of confidence intervals or fiducial limits illustrated in the case of the binomial, Biometrika 26 (1934) 404–413.
[12] S.L. Wilks, Mathematical Statistics, Wiley, New York, 1962.
[13] T. Fagan, Exact 95% confidence intervals for differences in binomial proportions, Comput. Biol. Med. 29 (1999) 83–87.

Timothy D. Ross attended Wright State University and the Air Force Institute of Technology, both in the USA, where he received B.S., M.S., and Ph.D. degrees in Systems and Electrical Engineering. As a longtime employee of the United States Air Force Research Laboratory (AFRL), he has supported the development, test, and transition of technologies for computer exploitation of sensed data. He has recently renewed ties with Wright State as an adjunct faculty member, teaching in these same areas. The work reported here was in support of AFRL's Comprehensive Performance Assessment of Sensor Exploitation (COMPASE) Center.