Statistica Sinica **19** (2009), 905-923

TOLERANCE INTERVALS FOR DISCRETE DISTRIBUTIONS IN EXPONENTIAL FAMILIES

Tianwen Tony Cai and Hsiuying Wang

University of Pennsylvania and National Chiao Tung University

Abstract: Tolerance intervals are widely used in industrial applications. So far attention has been mainly focused on the construction of tolerance intervals for continuous distributions. In this paper we introduce a unified analytical approach to the construction of tolerance intervals for discrete distributions in exponential families with quadratic variance functions. These tolerance intervals are shown to have desirable probability matching properties and outperform existing tolerance intervals in the literature.

Key words and phrases: Coverage probability, Edgeworth expansion, exponential family, probability matching, tolerance interval.

1. Introduction

Statistical tolerance intervals are important in many industrial applications ranging from engineering to the pharmaceutical industry. See, for example, Hahn and Chandra (1981) and Hahn and Meeker (1991). The goal of a tolerance interval is to contain at least a specified proportion of the population, β, with a specified degree of confidence, $1-\alpha$. More specifically, let X be a random variable with cumulative distribution function F. An interval $(L(X), U(X))$ is said to be a β-content, $(1-\alpha)$-confidence tolerance interval for F (called a $(\beta, 1-\alpha)$ tolerance interval for short) if

$$P\{[F(U(X)) - F(L(X))] \geq \beta\} = 1 - \alpha. \tag{1.1}$$

One-sided tolerance bounds can be defined analogously. A bound $L(X)$ is said to be a $(\beta, 1-\alpha)$ lower tolerance bound if $P\{1-F(L(X)) \geq \beta\} = 1-\alpha$ and a bound $U(X)$ is said to be a $(\beta, 1-\alpha)$ upper tolerance bound if $P\{F(U(X)) \geq \beta\} = 1-\alpha$.

Ever since the pioneering work of Wilks (1941, 1942), construction of tolerance intervals for continuous distributions has been extensively studied. See, for example, Wald and Wolfowitz (1946), Easterling and Weeks (1970), Kocherlakota and Balakrishnan (1986), Vangel (1992), Mukerjee and Reid (2001), and Krishnamoorthy and Mathew (2004). Compared with the continuous distributions, literature on tolerance intervals for discrete distributions is sparse. This is mainly due to the difficulty in deriving explicit expression for the tolerance intervals in the discrete case. Zacks (1970) proposed a criterion to select tolerance limits for monotone likelihood ratio families of discrete distributions. The most widely used tolerance intervals to date for Poisson and Binomial distributions were proposed by Hahn and Chandra (1981). The intervals are constructed by a two-step procedure. See Hahn and Meeker (1991) for a survey of these intervals.

Although tolerance intervals are useful and important, their properties, such as their coverage probability, have not been studied as much as those of confidence intervals. As we shall see in Section 2, the tolerance intervals given in Hahn and Chandra (1981) tend to be very conservative in terms of their coverage probability. Techniques for the construction of tolerance intervals in the literature often vary from distribution to distribution.

In this paper, we introduce a unified analytical approach using the Edgeworth expansions for the construction of tolerance intervals for the discrete distributions in exponential families with quadratic variance functions. We show that these tolerance intervals enjoy desirable probability matching properties and outperform existing tolerance intervals in the literature. The most satisfactory aspects of our results are the constancy of the phenomena, and uniformity in the final resolutions of these problems. Edgeworth expansions have also been used very successfully for the construction of confidence intervals in discrete distributions. See Hall (1982), Brown, Cai and DasGupta (2002, 2003), and Cai (2005). Construction of tolerance interval is closely related to the construction of confidence interval for quantiles. A one-sided tolerance bound is equivalent to a one-sided confidence bound on a quantile of the distribution. See Hahn and Meeker (1991). Therefore, the proposed method can be also employed for the quantile estimation problem.

The paper is organized as follows. We begin in Section 2 by briefly reviewing the existing tolerance intervals for Binomial and Poisson distributions and showing that they have serious deficiencies in terms of coverage probability. The serious deficiency of these intervals calls for better alternatives. After Section 3.1, in which basic notations and definitions of natural exponential family are summarized, the first-order and second-order probability matching tolerance intervals are introduced. As in the case of confidence intervals, the coverage probability of the tolerance intervals for the lattice distributions such as Binomial and Poisson distributions contains two components: oscillation and systematic bias. The oscillation in the coverage probability, which is due to the lattice structure of the distributions, is unavoidable for any non-randomized procedures. The systematic bias, which is large for many existing tolerance intervals, can be nearly eliminated. We show that our new tolerance intervals have better coverage properties in the sense that they have nearly vanishing systematic bias in all the distributions under consideration.

In Section 4, two-sided tolerance intervals are constructed by using one-sided upper and lower probability matching tolerance bounds. In addition to the coverage properties, parsimony in expected length of the two-sided intervals is also discussed. Section Appendix is an appendix containing detailed technical derivations of the tolerance intervals. The derivations are based on the two-term Edgeworth and Cornish-Fisher expansions.

2. Tolerance Intervals: Existing Methods

As mentioned in the introduction, we construct tolerance intervals for discrete distributions in the exponential families. In this section we review the existing tolerance intervals for two important discrete distributions, the Binomial and Poisson distributions. These tolerance intervals will be used for comparison with the new intervals constructed in the present paper.

The most widely used method for constructing tolerance intervals for the Binomial and Poisson distributions was proposed by Hahn and Chandra (1981). Suppose x is the observed value of a random variable X having a Binomial distribution $B(n, \theta)$ or a Poisson distribution $Poi(n\theta)$, and that one wishes to construct a tolerance interval based on x. The method introduced by Hahn and Chandra (1981) for constructing a $(\beta, 1-\alpha)$ tolerance interval $(L(x), U(x))$ has two steps.

(i) Construct a two-sided $(1-\alpha)$-level confidence interval (l, u) for θ, where l and u depends on x.

(ii) Find the minimum number $U(x)$ and the maximum number $L(x)$ such that

$$p(X \leq U(x)|\theta = \mu) \geq \frac{1+\beta}{2} \quad \text{and} \quad p(X > L(x)|\theta = l) \geq \frac{1+\beta}{2}.$$

Similarly, a lower $(\beta, 1-\alpha)$ tolerance bound $L(x)$ can be constructed by finding a lower $(1-\alpha)$ confidence bound of θ, say l, and then deriving the maximum value $L(x)$ such that $p_l(X > L(x)) \geq \beta$.

For this two-step procedure, it is clear that the choice of the confidence interval used in Step 1 is important to the performance of the resulting tolerance interval. For any $0 < \gamma < 1$, let $z_\gamma = \Phi^{-1}(1-\gamma)$ be the $1-\gamma$ quantile of a standard normal distribution. Hahn and Meeker (1991) suggested $(1-\alpha)$ level confidence intervals for the Binomial case,

$$(l, u) = \hat{\theta} \pm z_{\alpha/2}\Big(\frac{\hat{\theta}(1-\hat{\theta})}{n}\Big)^{1/2}, \tag{2.1}$$

$$(l, u) = \bigg(\Big(1 + \frac{(n-x+1)F_{(\alpha/2;2n-2x+2,2x)}}{x}\Big)^{-1}, \Big(1 + \frac{n-x}{(x+1)F_{(\alpha/2;2x+2,2n-2x)}}\Big)^{-1}\bigg), \tag{2.2}$$

where $F_{(a;r_1,r_2)}$ denotes the $1-a$ quantile of the F distribution with r_1 and r_2 degrees of freedom. For the Poisson distribution, the suggested $(1-\alpha)$ confidence intervals in Hahn and Meeker (1991) are

$$(l,u) = \hat{\theta} \pm z_{\alpha/2}\Big(\frac{\hat{\theta}}{n}\Big)^{1/2}, \tag{2.3}$$

$$(l,u) = \left(0.5\frac{\chi^2_{(\alpha/2;2x)}}{n},\ 0.5\frac{\chi^2_{(1-\alpha/2;2x+2)}}{n}\right), \tag{2.4}$$

where $\chi^2_{(a;r_1)}$ is a quantile of the chi-square distribution with r_1 degrees of freedom. The $\hat{\theta}$ in (2.1) and (2.3) denotes the sample mean. The confidence bounds for one-sided tolerance intervals in both Binomial and Poisson distributions are given analogously.

Figure 1 presents the coverage probabilities of both the two-sided and one-sided tolerance intervals for the Binomial and Poisson distributions. It can easily be seen from the plots that these tolerance intervals are too conservative with higher or lower coverage probability than the nominal level for both distributions.

As in the case of confidence intervals, the coverage probability of the tolerance intervals contains two components: oscillation and systematic bias. The oscillation in the coverage probability, due to the lattice structure of the Binomial and Poisson distributions, is unavoidable for any non-randomized procedures. However, the systematic biases for the existing tolerance intervals are significantly larger than we expected. This is partly due to the poor behavior of the confidence intervals used in the construction of the tolerance intervals. Note that (2.1) and (2.2) are the Wald and Clopper-Pearson intervals for the binomial proportion. It is known that these confidence intervals have poor performances. See, for example, Agresti and Coull (1998) and Brown et al. (2002). It is thus possible to improve the performance of the tolerance intervals by using better confidence intervals, like those presented in Brown et al. (2002) for the binomial distribution.

However, we do not take this approach here. The goal of this paper is to provide a unified analytical approach to the construction of desirable tolerance intervals for exponential families with certain optimality properties. The results show that the Edgeworth expansion approach is a powerful tool for solving this problem.

In Section 3 we introduce new tolerance intervals using the Edgeworth expansion. These intervals have better coverage properties in the sense that they have nearly vanishing systematic bias. Figure 3 presents the coverage probabilities of the proposed two-sided tolerance intervals for the Binomial and Poisson

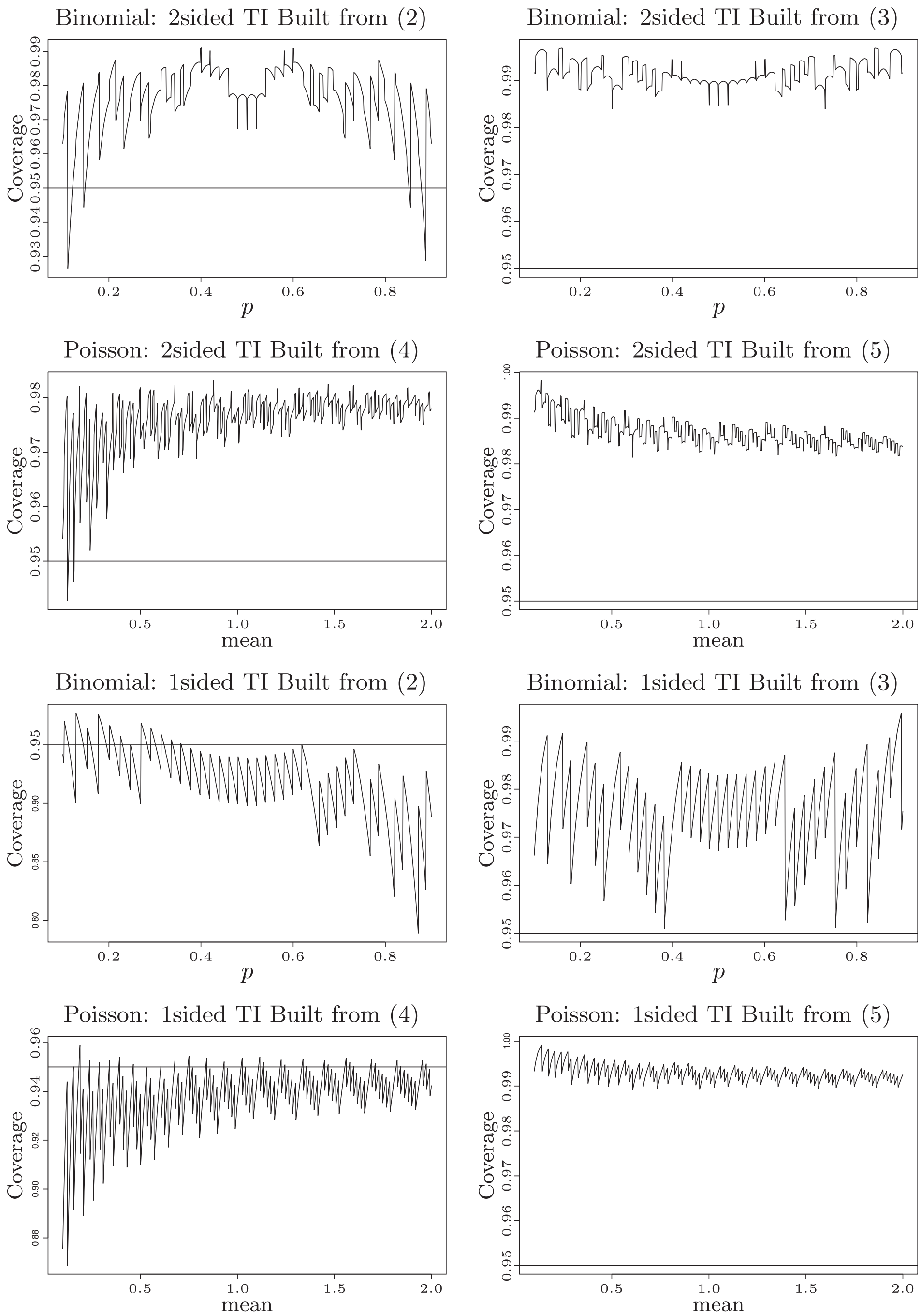

Figure 1. Coverage probabilities of the 90%-content, 95% level two-sided (top two rows) and one-sided (bottom two rows) tolerance intervals for the Binomial and Poisson distributions with $n = 50$, where p is the probability of success for the binomial distribution.

distributions. Compared with Figure 1, the tolerance intervals certainly have much better performance than the existing intervals in the sense that the actual coverage probability is much closer to the nominal level. The detailed derivation of our tolerance intervals is given in the next section.

3. Probability-Matching Tolerance Intervals

In this section we construct one-sided probability-matching tolerance intervals in the natural discrete exponential family (NEF) with quadratic variance functions (QVF) by using the Edgeworth expansion. After Section 3.1 in which basic notations and definitions of natural exponential family are given, we introduce the first-oder and second-order probability matching tolerance intervals in Section 3.2.

3.1. Natural exponential family

The NEF-QVF family contains three important discrete distributions: Binomial, Negative Binomial, and Poisson (see, e.g., Morris (1982) and Brown (1986).

We first state some basic facts about the NEF-QVF families. The distributions in a natural exponential family have the form

$$f(x|\xi) = e^{\xi x - \psi(\xi)} h(x),$$

where ξ is called the natural parameter. The mean μ, variance σ^2 and cumulant generating function ϕ_ξ are, respectively,

$$\mu = \psi'(\xi), \quad \sigma^2 = \psi''(\xi), \quad \text{and} \quad \phi_\xi(t) = \psi(t+\xi) - \psi(\xi).$$

The cumulants are given as $K_r = \psi^{(r)}(\xi)$. Let β_3 and β_4 denote the skewness and kurtosis. In the subclass with a quadratic variance function (QVF), the variance $\psi''(\xi)$ depends on ξ only through the mean μ and, indeed,

$$\sigma^2 \equiv V(\mu) = d_0 + d_1\mu + d_2\mu^2 \tag{3.1}$$

for suitable constants d_0, d_1, and d_2. We denote the discriminant by

$$\Delta = d_1^2 - 4d_0d_2. \tag{3.2}$$

The notation Δ is used in the statements of theorems for both the discrete and the continuous cases, although for all the discrete cases Δ happens to be equal to 1. Note that $d\mu/d\xi = \psi''(\xi) = \sigma^2$, so

$$K_3 = \psi^{(3)}(\xi) = \frac{dV}{d\mu}\cdot\frac{d\mu}{d\xi} = (d_1 + 2d_2\mu)\sigma^2 \text{ and } K_4 = \psi^{(4)}(\xi) = \frac{dK_3}{d\mu}\cdot\frac{d\mu}{d\xi} = \Delta\sigma^2 + 6d_2\sigma^4.$$

Hence,

$$\beta_3 = \frac{K_3}{\sigma^3} = (d_1 + 2d_2\mu)\sigma^{-1} \quad \text{and} \quad \beta_4 = \frac{K_4}{\sigma^4} = \Delta\sigma^{-2} + 6d_2. \tag{3.3}$$

Here are the important facts about the Binomial, Negative Binomial, and Poisson distributions.

- Binomial, $\mathrm{B}(1,p)$: $\xi = \log(p/q)$, $\psi(\xi) = \log(1+e^{\xi})$, and $h(x) = 1$. Also $\mu = p$, $V(\mu) = pq = \mu - \mu^2$. Thus $d_0 = 0$, $d_1 = 1$, $d_2 = -1$,

$$\beta_3 = \frac{1-2\mu}{(\mu(1-\mu))^{1/2}}, \quad \text{and} \quad \beta_4 = \frac{1-6\mu+6\mu^2}{\mu(1-\mu)}.$$

- Negative Binomial, $\mathrm{NB}(1,p)$, the number of successes before the first failure: $\xi = \log p$, $\psi(\xi) = -\log(1-e^{\xi})$, and $h(x) = 1$, where p is the probability of success. Here $\mu = p/q$, and $V(\mu) = p/q^2 = \mu + \mu^2$, so $d_0 = 0$, $d_1 = 1$, $d_2 = 1$,

$$\beta_3 = \frac{1+2\mu}{(\mu(1+\mu))^{1/2}}, \quad \text{and} \quad \beta_4 = \frac{1+6\mu+6\mu^2}{\mu(1+\mu)}.$$

- Poisson, $\mathrm{Poi}(\lambda)$: $\xi = \log\lambda$, $\psi(\xi) = e^{\xi}$, and $h(x) = 1/x!$. Then $\mu = \lambda$, $V(\mu) = \mu$, and here $d_0 = 0$, $d_1 = 1$, $d_2 = 0$,

$$\beta_3 = \frac{1}{\mu^{1/2}}, \quad \text{and} \quad \beta_4 = \frac{1}{\mu}.$$

3.2. One-sided tolerance interval

We now introduce the first-order and second-order probability matching one-sided tolerance intervals. Let $X = \sum_{i=1}^{n} X_i$, where X_i are iid observations from one of the three distributions discussed in Section 3.1. We denote the distribution of X by $F_{n,\mu}$ and focus our discussion on the lower tolerance intervals. The upper tolerance intervals can be constructed analogously. Two-sided tolerance intervals will be discussed in Section 4.

Similar to confidence intervals, the coverage probability of a lower $(\beta, 1-\alpha)$ tolerance interval admits a two-term Edgeworth expansion of the general form

$$\begin{aligned} P(1 - F_{n,\mu}(L(X)) \geq \beta) = 1 - \alpha + S_1 \cdot n^{-1/2} + Osc_1 \cdot n^{-1/2} + S_2 \cdot n^{-1} \\ + Osc_2 \cdot n^{-1} + O(n^{-3/2}), \end{aligned} \tag{3.4}$$

where the first $O(n^{-1/2})$ term, $S_1 n^{-1/2}$, and the first $O(n^{-1})$ term, $S_2 n^{-1}$, are the first and second order smooth terms, respectively, and $Osc_1 \cdot n^{-1/2}$ and $Osc_2 \cdot n^{-1}$ are the oscillatory terms. (The oscillatory terms vanish in the case of continuous distributions.) The smooth terms capture the systematic bias in the coverage probability. See Bhattacharya and Rao (1976) and Hall (1982) for details on Edgeworth expansions.

We call a tolerance interval *first-order probability matching* if the first order smooth term $S_1 n^{-1/2}$ vanishes, and call the interval *second-order probability matching* if both $S_1 n^{-1/2}$ and $S_2 n^{-1}$ vanish. Note that the oscillatory terms are unavoidable for any nonrandomized procedures in the case of lattice distributions. See Ghosh (1994) and Ghosh (2001) for general discussions on probability matching confidence sets.

Motivated by the discussion given at the end of this section, we consider an approximate β-content, $(1-\alpha)$-confidence lower tolerance bound of the form

$$L(X) = X + a - b\sqrt{n(d_0 + \frac{d_1 X}{n} + \frac{d_2 X^2}{n^2}) + c}, \tag{3.5}$$

where d_0, d_1 and d_2 are the constants in (3.1), and a, b, and c are constants depending on α and β such that

$$L(X) < L(Y) \text{ if } X < Y. \tag{3.6}$$

Remark. The quantity a in (3.5) “re-centers” the tolerance interval and, we see later, a is important to the performance of the tolerance interval. The quantity c in (3.5) plays the role of a “boundary correction”. The parameter c is to be adjusted such that the coefficients S_1 and S_2 of the smooth terms $n^{-1/2}$ and n^{-1} in (3.4) can be zero. The effect of c can be significant when μ is near the boundaries.

We use the Edgeworth expansion to choose the constants a, b, and c so that the resulting tolerance intervals are first-order and second-order probability matching. The first step in the derivation is to invert the constraint $1 - F(L(X)) \geq \beta$ to a constraint on X of the form $X \leq u(\mu, \beta)$. Then the coverage probability of the tolerance interval can be expanded using the Edgeworth expansion. The optimal choice of the values a, b, and c can then be solved by setting the smooth terms in the expansion to zero. The algebra involved here is more tedious than for deriving the probability matching confidence interval. The detailed proof is given in the Appendix.

Theorem 1. *The tolerance interval given in* (3.5) *is first-order probability matching for the three discrete distributions in the NEF-QVF if*

$$a = \frac{1}{6}[(z_{1-\beta}^2 - 1)(1 + 2d_2\hat{\mu}) + (1 + 3z_\alpha z_{1-\beta} + 2z_\alpha^2)(d_1 + 2d_2\hat{\mu})], \tag{3.7}$$

$$b = z_\alpha + z_{1-\beta}, \tag{3.8}$$

and $c = 0$, where $\hat{\mu} = X/n$ and $\hat{\sigma} = \sqrt{d_0 + d_1\hat{\mu} + d_2\hat{\mu}^2}$. The tolerance interval (3.5) *is second-order probability matching with a and b given as in* (3.7) *and (3.8), and c given by*

$$c = \frac{1}{36(z_\alpha + z_{1-\beta})}\left\{(-1 + 18d_0d_2 + 2(-8 + 9d_1)d_2\hat{\mu} + 2d_2^2\hat{\mu}^2)z_{1-\beta}^3 \right. \\ +24d_2(d_0 + \hat{\mu}(d_1 + d_2\hat{\mu}))z_{1-\beta}^2 z_\alpha + z_{1-\beta}[1 + 2d_2\hat{\mu}(20 + 5d_2\hat{\mu} + 24d_2\hat{\mu}z_\alpha^2) \\ +3d_1^2(2 + z_\alpha^2) + 18d_0d_2(-3 + 2z_\alpha^2) + 6d_1d_2\hat{\mu}(-5 + 8z_\alpha^2)] + z_\alpha[d_1^2(7 + 2z_\alpha^2) \\ \left. +2d_1d_2\hat{\mu}(5 + 13z_\alpha^2) + 2d_2(9d_0(-1 + z_\alpha^2) + d_2\hat{\mu}^2(5 + 13z_\alpha^2))]\right\}. \qquad (3.9)$$

Remark. We have focused above on the construction of lower tolerance intervals. The first order and second order β-content, $(1-\alpha)$-confidence upper tolerance intervals can be constructed analogously as

$$X + a + b\sqrt{n(d_0 + \frac{d_1X}{n} + \frac{d_2X^2}{n^2})}, \qquad (3.10)$$

$$X + a + b\sqrt{n(d_0 + \frac{d_1X}{n} + \frac{d_2X^2}{n^2}) + c}, \qquad (3.11)$$

respectively, with the same a, b, and c as the lower tolerance intervals.

For all three distributions, $b = z_\alpha + z_{1-\beta}$. It is useful to give the expressions of the constants a and c individually for each of the three distributions.

1. Binomial: $a = \frac{1}{6}(1 - 2\hat{\mu})(z_\alpha + z_{1-\beta})(2z_\alpha + z_{1-\beta})$, and

$$c = -\frac{1}{18}(13z_\alpha^2 + 11z_\alpha z_{1-\beta} + z_{1-\beta}^2 + 5)(\hat{\mu} - \hat{\mu}^2) + \frac{1}{36}(2z_\alpha^2 + z_\alpha z_{1-\beta} - z_{1-\beta}^2 + 7).$$

2. Poisson: $a = \frac{1}{6}(z_\alpha + z_{1-\beta})(2z_\alpha + z_{1-\beta})$ and $c = \frac{1}{36}(7 - z_{1-\beta}^2 + z_\alpha z_{1-\beta} + 2z_\alpha^2)$.
3. Negative Binomial: $a = \frac{1}{6}(1 + 2\hat{\mu})(z_\alpha + z_{1-\beta})(2z_\alpha + z_{1-\beta})$ and $c = \frac{1}{18}(13z_\alpha^2 + 11z_\alpha z_{1-\beta} + z_{1-\beta}^2 + 5)(\hat{\mu} + \hat{\mu}^2) + \frac{1}{36}(2z_\alpha^2 + z_\alpha z_{1-\beta} - z_{1-\beta}^2 + 7)$.

We consider the discrete NEF-QVF family because the most important discrete distributions are there. Our results can be generalized to the discrete NEF family with variance functions of the form $V(\mu) = d_0 + d_1\mu + d_2\mu^2 + d_3\mu^3 + d_4\mu^4$, because there exists a general solution for $V(\mu) = 0$.

Figure 2 plots the coverage probabilities of the first-order and second-order probability matching $(0.9, 0.95)$ lower tolerance intervals for $n = 50$. It is clear from Figure 2 that for the three discrete distributions, the first and second order probability matching tolerance intervals have nearly vanishing systematic bias,

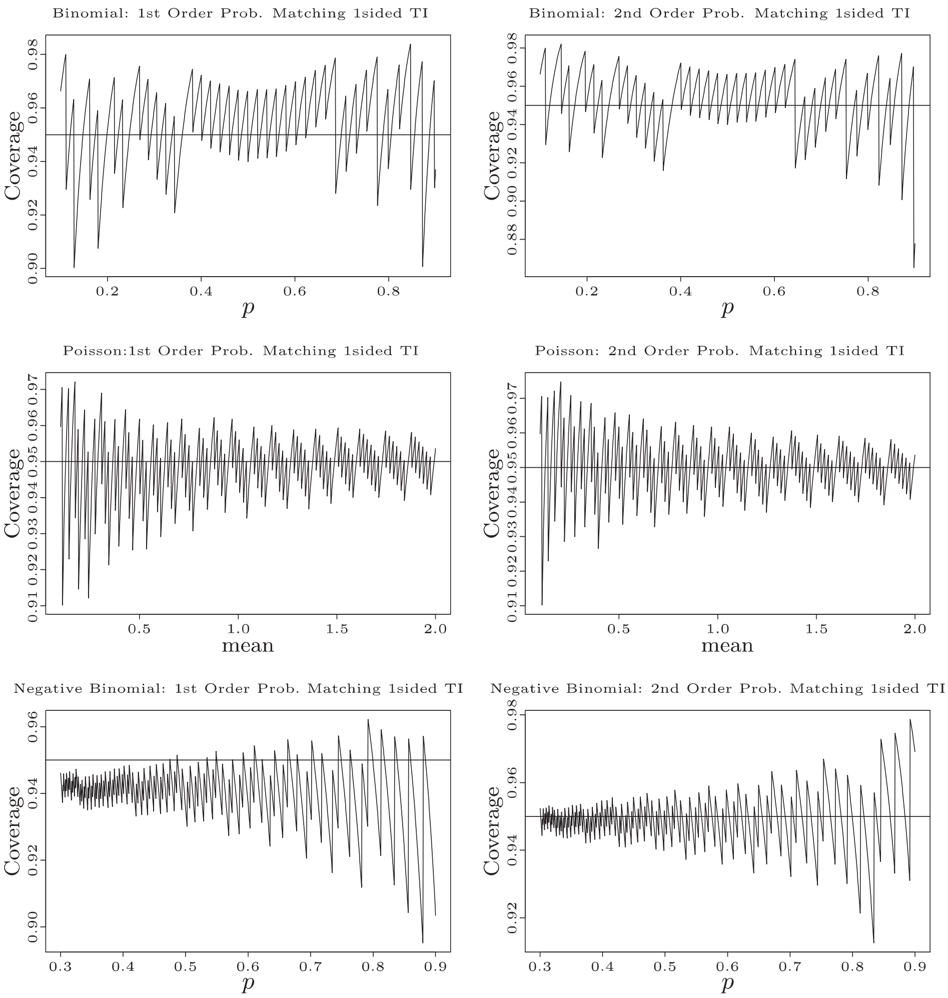

Figure 2. Coverage probabilities of the 90%-content, 95% level of the first order and second order probability matching lower tolerance bounds for the Binomial, Poisson and Negative Binomial distributions, with $n = 50$.

and the second order probability matching interval has noticeably smaller systematic bias than the first order probability matching interval for the Negative Binomial distribution. Since there are oscillatory terms for the discrete distributions in the Edgeworth expansion, we mainly evaluate the performances of the tolerance intervals in terms of the smooth terms. Here, we evaluate the performance of a tolerance interval by checking if their coverage probabilities can approximate the nominal level well. The coverage probability of the proposed one-sided tolerance interval can be closer to the nominal level than that of the existing tolerance intervals, comparing Figure 2 with Figure 1.

The motivation for considering the form (3.5) is briefly described as follows. Let $X_1, \ldots, X_n$ be a sample from a normal distribution $N(\mu, \sigma^2)$. Wald and Wolfowitz (1946) introduced the β-content, $(1-\alpha)$-confidence tolerance interval

$$[\bar{X} - \sqrt{\frac{n-1}{\chi^2_{n-1,\alpha}}}tS, \bar{X} + \sqrt{\frac{n-1}{\chi^2_{n-1,\alpha}}}tS], \tag{3.12}$$

where $\bar{X}$ and S are the sample mean and sample standard deviation, respectively, $\chi^2_{n-1,\alpha}$ is the $\alpha-$quantile of the chi-squared distribution with $n-1$ degrees of freedom, and t is the solution of the equation

$$\int_{\frac{1}{\sqrt{n}}-t}^{\frac{1}{\sqrt{n}}+t} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx = \beta.$$

To make a better analogy between the NEF-QVF families and normal cases, we first attempt to rewrite the tolerance interval in (3.12) in terms of $X = \sum_{i=1}^n X_i$ under $N(n\mu, n\sigma^2)$. Note that (3.12) implies

$$1 - \alpha \approx P(\Phi_{\mu,\sigma}(\bar{X} + \sqrt{\frac{n-1}{\chi^2_{n-1,\alpha}}}tS) - \Phi_{\mu,\sigma}(\bar{X} - \sqrt{\frac{n-1}{\chi^2_{n-1,\alpha}}}tS) \geq \beta),$$

where $\Phi_{\mu,\sigma}$ denotes the cdf of the $N(\mu, \sigma^2)$ distribution. Since

$$\Phi_{\mu,\sigma}(\bar{X} \pm \sqrt{\frac{n-1}{\chi^2_{n-1,\alpha}}}tS) = \Phi_{n\mu,\sqrt{n}\sigma}(n\mu + \sqrt{n}(\bar{X} - \mu) \pm \sqrt{\frac{n-1}{\chi^2_{n-1,\alpha}}}t\sqrt{n}S), \tag{3.13}$$

and replacing μ by the lower or upper limits of a β-confidence confidence interval $(\bar{X} - z_{(1-\beta)/2}S/\sqrt{n}, \bar{X} + z_{(1-\beta)/2}S/\sqrt{n})$ for μ, we have the tolerance interval

$$[X - (\sqrt{\frac{n-1}{\chi^2_{n-1,\alpha}}}t + (1 - \frac{1}{\sqrt{n}})z_{(1-\beta)/2})\sqrt{n}s, X + (\sqrt{\frac{n-1}{\chi^2_{n-1,\alpha}}}t + (1 - \frac{1}{\sqrt{n}})z_{(1-\beta)/2})\sqrt{n}s]$$

under the $N(n\mu, n\sigma^2)$ distribution.

For the NEF-QVF families, by the Central Limit Theorem, and adopting the method in the normal case by identifying $d_0 + d_1X/n + d_2X^2/(n^2)$ as s^2, an approximate β-content, $(1-\alpha)$-confidence tolerance lower bound and upper bound are $X - A$ and $X + A$, respectively, where $A = (\sqrt{n-1/\chi^2_{n-1,\alpha}}t + (1 - 1/\sqrt{n})z_{(1-\beta)/2})\sqrt{n(d_0 + d_1X/n + d_2X^2/n^2)}$. More generally, we consider tolerance bounds of the form

$$L(X) = X + a - b\sqrt{n(d_0 + \frac{d_1X}{n} + \frac{d_2X^2}{n^2}) + c}$$
$$U(X) = X + a + b\sqrt{n(d_0 + \frac{d_1X}{n} + \frac{d_2X^2}{n^2}) + c}$$

with suitably chosen constants a, b, and c.

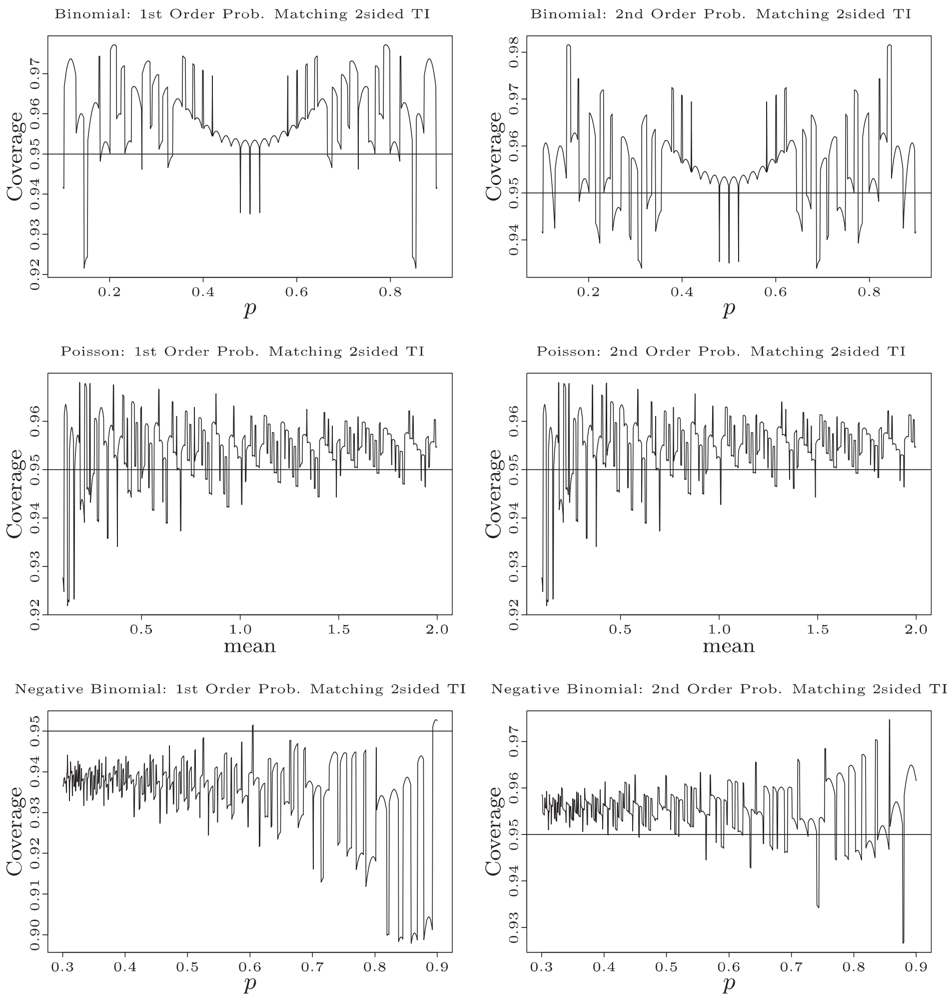

Figure 3. Coverage probabilities of the 90%-content, 95% level of the first order and second order two-sided probability matching tolerance intervals for the Binomial, Poisson and Negative Binomial distributions, with $n = 50$.

4. Two-sided Tolerance Interval

We have derived the optimal one-sided first-order and second-order probability matching tolerance intervals in Section 3, and is natural to consider two-sided tolerance intervals. However, it is difficult to obtain optimal choices for the values of a, b, and c for a two-sided tolerance interval using the same approach. A key step in the derivation of the one-sided intervals given in Section 3 is the inversion of the constraint $1 - F(L(X)) \geq \beta$ to $X \leq u(\mu, \beta)$. Similarly, for a two-sided tolerance interval, it is necessary to invert the constraint

$$F_{n,\mu}(U(X)) - F_{n,\mu}(L(X)) \geq \beta \tag{4.1}$$

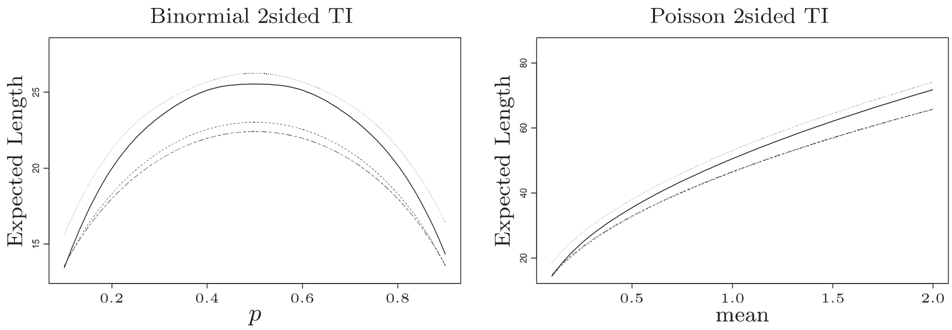

Figure 4. Expected lengths of the 90%-content, 95% level of the two-sided tolerance interval based on (2)(solid, binomial) and (3)(dotted, binomial), the tolerance interval based on (4)(solid, Poisson) and (5)(dotted, Poisson), the first order probability matching two-sided tolerance interval (dashed) and the second order probability matching two-sided tolerance interval (long-dashed) for Binomial (left panel) and Poisson (right panel) distributions, with $n = 50$. For the Poisson distribution, the dashed and long-dashed lines almost overlap.

in terms of X. This is theoretically difficult.

We thus take the alternative approach of using one-sided upper and lower tolerance bounds. Let $U_{(1+\beta)/2}(X)$ and $L_{(1+\beta)/2}(X)$ be the upper and lower probability matching $((1+\beta)/2, 1-\alpha)$ tolerance bounds, respectively. We propose to use the interval

$$(L_{(1+\beta)/2}(X),\ U_{(1+\beta)/2}(X)) \tag{4.2}$$

as a β-content, $(1-\alpha)$-confidence two-sided tolerance interval.

Figure 3 plots the coverage probabilities of two-sided $(0.9, 0.95)$ tolerance intervals built from the first-order and second-order probability matching tolerance bounds. The coverage probabilities for the two-sided tolerance intervals are calculated exactly for the three discrete distributions. By comparing Figure 3 with Figure 1, it is clear that the performance of these two-sided intervals is better than that of existing two-sided tolerance intervals in the case of Binomial and Poisson distributions. The coverage probability of the proposed two-sided tolerance intervals oscillates in the center from 0.95 to 0.96 with a systematic bias less than 0.01. In contrast, the coverage probability of the two-sided tolerance intervals in Figure 1 oscillates in the center from 0.975 to 0.99 with a systematic bias greater than 0.025.

In addition to coverage probability, parsimony in length is also an important issue. Figure 4 compares the expected length of the two new tolerance intervals with that of the two intervals discussed in Section 2. It is clear that the expected length of the proposed tolerance intervals is less than that of the existing tolerance intervals. Thus, based on both coverage probability and expected length, the tolerance intervals derived from our analytical approach outperform existing tolerance intervals.

Acknowledgement

The research of Tony Cai was supported in part by NSF Grant DMS-0604954.

Appendix. Proof of Theorem 1

We begin by introducing notation and a technical lemma. All three discrete distributions under consideration are lattice distributions with the maximal span of one. Lemma 1 below gives the Edgeworth expansion and Cornish-Fisher expansion for these distributions. The first part is from Brown et al. (2003). For details on the Edgeworth expansion and Cornish-Fisher expansion, see Esseen (1945), Petrov (1975), Bhattacharya and Rao (1976), and Hall (1982).

Let $X_1, \ldots, X_n$ be iid observations from a discrete distribution in the NEF-QVF family. Denote the mean of X_1 by μ and the standard deviation by σ. Let $\beta_3 = K_3/\sigma^3$ and $\beta_4 = K_4/\sigma^4$ be the skewness and kurtosis of X_1, respectively. Set $X = \sum_1^n X_i$ and $Z_n = n^{1/2}(\bar{X} - \mu)/\sigma$, where $\bar{X} = X/n$. Let $F_n(z) = P(Z_n \le z)$ be the cdf of Z_n and let $f_{n,\mu,\beta} = \inf\{x : P(X \le x) \ge 1-\beta\}$ be the $1-\beta$ quantile of the distribution of X.

Lemma 1. *Suppose $z = z_0 + c_1 n^{-1/2} + c_2 n^{-1} + O(n^{-3/2})$, where z_0, c_1 and c_2 are constants. Then the two-term Edgeworth expansion for $F_n(z)$ is*

$$F_n(z) = \Phi(z_0) + p_1(z)\phi(z_0)n^{-1/2} + p_2(z)\phi(z_0)n^{-1} + Osc_1 \cdot n^{-1/2} + Osc_2 \cdot n^{-1} + O(n^{-3/2}), \tag{A.1}$$

where Osc_1 and Osc_2 are bounded oscillatory functions of μ and z, and

$$p_1(z) = c_1 + \frac{1}{6}\beta_3(1 - z_0^2), \tag{A.2}$$

$$p_2(z) = c_2 - \frac{1}{2}z_0 c_1^2 + \frac{1}{6}(z_0^3 - 3z_0)\beta_3 c_1 - \frac{1}{24}\beta_4(z_0^3 - 3z_0) - \frac{1}{72}\beta_3^2(z_0^5 - 10z_0^3 + 15z_0), \tag{A.3}$$

$$p_3(z) = -c_1 + \frac{1}{6}\beta_3(z_0^2 - 3). \tag{A.4}$$

The two-term Cornish-Fisher expansion for $f_{n,\mu,\beta}$ is

$$\begin{aligned} f_{n,\mu,\beta} = & n\mu - z_{1-\beta}(n\sigma^2)^{1/2} + \frac{1}{6}(1+2d_2\mu)(z_{1-\beta}^2 - 1) \\ & + \Big[\frac{1}{72}(z_{1-\beta}^3 - z_{1-\beta}) + \frac{1}{9}(d_2\mu + d_2^2\mu^2)(2z_{1-\beta}^3 - 5z_{1-\beta}) \\ & - \frac{\sigma^2}{4}d_2(z_{1-\beta}^3 - 3z_{1-\beta})\Big](n\sigma^2)^{-1/2} + Osc_3 + Osc_4 \cdot n^{-1/2} + O(n^{-1}), \end{aligned} \quad (A.5)$$

where Osc_3 and Osc_4 are bounded oscillatory functions of μ and β.

We focus on the smooth terms and ignore the oscillatory terms in (A.1) and (A.5) in the following calculations.

Proof of Theorem 1. It follows from (3.6) that $1 - F(L(X)) \geq \beta$ is equivalent to $L(X) \leq f_{n,\mu,\beta}$ and to $X \leq u(\mu, \beta)$, where

$$u(\mu,\beta) = \frac{1}{(1 - b^2 d_2 n^{-1})}\left\{-a + \frac{1}{2}b^2 d_1 + f_{n,\mu,\beta} + bD_n\right\} \quad (A.6)$$

with

$$D_n = \Big\{d_0 n + n^{-1} f_{n,\mu,\beta}(nd_1 + d_2 f_{n,\mu,\beta}) - ad_1 + \frac{1}{4}b^2 d_1^2 + c - 2ad_2 f_{n,\mu,\beta} n^{-1} + (a^2 - b^2 c)d_2 n^{-1} - b^2 d_0 d_2\Big\}^{1/2}. \quad (A.7)$$

The coverage of the tolerance interval is then

$$P(1 - F_{n,\mu}(L(X)) \geq \beta) = P(X \leq u(\mu,\beta)) = P(Z_n \leq z_n), \quad (A.8)$$

where $Z_n = (X - n\mu)/\sqrt{n\sigma^2}$ and $z_n = (u(\mu,\beta) - n\mu)/\sqrt{n\sigma^2}$.

To derive the optimal choices for a, b, and c, we need the Edgeworth expansion of $P(Z_n \leq z_n)$ as well as the expansion of the quantile $f_{n,\mu,\beta}$ given in Lemma 1. By (A.5), the term $d_0 n + n^{-1} f_{n,\mu,\beta}(nd_1 + d_2 f_{n,\mu,\beta})$ in (A.7) is equal to

$$\begin{aligned} & n\sigma^2 - (n\sigma^2)^{1/2}(d_1 + 2d_2\mu)z_{1-\beta} + \frac{1}{6}(d_1 + 2d_2\mu)(1 + 2d_2\mu)(z_{1-\beta}^2 - 1) \\ & + \sigma^2 d_2 z_{1-\beta}^2 + O(n^{-1/2}). \end{aligned} \quad (A.9)$$

It then follows from (A.5), (A.7) and (A.9), and the Taylor expansion

$$(x + \epsilon)^{1/2} = x^{1/2} + \frac{1}{2}x^{-1/2}\epsilon - \frac{1}{8}x^{-3/2}\epsilon^2 + O(x^{-5/2}\epsilon^3)$$

for large x and small ϵ, that

$$D_n = (n\sigma^2)^{1/2} - \frac{1}{2}(d_1 + 2d_2\mu)z_{1-\beta} + \Big\{ -\frac{1}{2}(d_1 + 2d_2\mu)a + \frac{1}{8}b^2 d_1^2 + \frac{1}{2}c + \frac{1}{12}(1+2d_2\mu)(d_1+2d_2\mu)(z_{1-\beta}^2 - 1) - \frac{1}{2}b^2 d_0 d_2 - \frac{1}{8}(d_1^2 - 4d_0 d_2)z_{1-\beta}^2 \Big\}(n\sigma^2)^{-1/2} + O(n^{-1}).$$

Note that $(1-b^2 d_2 n^{-1})^{-1} = 1 + b^2 d_2 n^{-1} + O(n^{-2})$. Using this and the above expansion for D_n, we have

$$\begin{aligned} z_n &= (b - z_{1-\beta}) + \Big\{\frac{1}{6}(1 + 2d_2\mu)(z_{1-\beta}^2 - 1) - \frac{1}{2}(d_1 + 2d_2\mu)z_{1-\beta}b \\ &\quad + (\frac{1}{2}d_1 + d_2\mu)b^2 - a\Big\}\sigma^{-1}n^{-1/2} \\ &\quad + \Big\{\frac{1}{4}d_2(3z_{1-\beta} - z_{1-\beta}^3)\sigma^2 + \frac{1}{9}(d_2\mu + d_2^2\mu^2)(2z_{1-\beta}^3 - 5z_{1-\beta}) \\ &\quad + \frac{1}{72}(z_{1-\beta}^3 - z_{1-\beta}) + (b - z_{1-\beta})b^2 d_2\sigma^2 + [-\frac{1}{2}a(d_1 + 2d_2\mu) + \frac{1}{8}b^2 d_1^2 \\ &\quad - \frac{1}{2}b^2 d_0 d_2 + \frac{1}{2}c + \frac{1}{12}(1 + 2d_2\mu)(d_1 + 2d_2\mu)(z_{1-\beta}^2 - 1) \\ &\quad - \frac{1}{8}(d_1^2 - 4d_0 d_2)z_{1-\beta}^2]b\Big\}\sigma^{-2}n^{-1} + O(n^{-3/2}) \\ &\equiv (b - z_{1-\beta}) + c_1 n^{-1/2} + c_2 n^{-1} + O(n^{-3/2}). \end{aligned} \quad \text{(A.10)}$$

It then follows from the Edgeworth expansion (A.1) for $P(Z_n \le z_n)$ given in Lemma 1 that b needs to be chosen as $b = z_\alpha + z_{1-\beta}$ in order for the coverage probability of the tolerance interval to be close to the nominal level $1 - \alpha$. With this choice of b, and using the notation in (3.4) for the Edgeworth expansion of $P(Z_n \le z_n)$, the coefficients for the smooth terms are

$$S_1 = [c_1 + \frac{1}{6}\beta_3(1 - z_\alpha^2)]\phi(z_\alpha), \quad \text{(A.11)}$$

$$S_2 = \Big\{c_2 - \frac{1}{2}z_\alpha c_1^2 + \frac{1}{6}(z_\alpha^3 - 3z_\alpha)\beta_3 c_1 - \frac{1}{24}\beta_4(z_\alpha^3 - 3z_\alpha) - \frac{1}{72}\beta_3^2(z_\alpha^5 - 10z_\alpha^3 + 15z_\alpha)\Big\}\phi(z_\alpha). \quad \text{(A.12)}$$

First-order probability matching interval: To make the tolerance interval first-order probability matching, we need $S_1 \equiv 0$, or equivalently $c_1 = \frac{1}{6}\beta_3(z_\alpha^2 - 1)$.

This leads to

$$\begin{aligned} a &= \frac{1}{6}[(z_{1-\beta}^2 - 1)(1 + 2d_2\mu) + 3z_\alpha(z_\alpha + z_{1-\beta})(d_1 + 2d_2\mu) + \sigma\beta_3(1 - z_\alpha^2)] \\ &= \frac{1}{6}[(z_{1-\beta}^2 - 1)(1 + 2d_2\mu) + (1 + 3z_\alpha z_{1-\beta} + 2z_\alpha^2)(d_1 + 2d_2\mu)]. \qquad (A.13) \end{aligned}$$

However, μ is unknown. We replace μ by $\hat{\mu}$ in a and set $c = 0$. It is straightforward to verify that there is no first-order effect by replacing μ with $\hat{\mu}$ in (A.13), and that

$$X + a - b\sqrt{n(d_0 + \frac{d_1 X}{n} + \frac{d_2 X^2}{n^2})}, \qquad (A.14)$$

with a and b given in (3.7) and (3.8), is first-order probability matching lower bound.

Second-order probability matching interval: To make the interval second-order probability matching, we need both $S_1 \equiv 0$ and $S_2 \equiv 0$. We can find the value of c from (A.10), (A.11) and (A.12). However, a was assumed to be a constant not depending on X in the original derivation of z_n. While in (A.14), a is a function of X and this has a second-order effect. We thus need to consider tolerance bound of the form (3.5) with a given in (3.7) and b given in (3.8), and to redo the analysis to find the optimal c. Set $h_1 = (1/6)(d_2((z_{1-\beta}^2 - 1) + 3(z_\alpha + z_{1-\beta})z_\alpha)) + 2d_2/2(1 - z_\alpha^2)$ and $h_2 = (1/6)[(z_{1-\beta}^2 - 1) + 3d_1 z_\alpha(z_\alpha + z_{1-\beta}) + d_1(1 - z_\alpha^2)]$. Then (3.5) can be rewritten as

$$L(X) = X[1 + 2h_1 n^{-1}] + h_2 - (z_\alpha + z_{1-\beta})\sqrt{n(d_0 + \frac{d_1 X}{n} + \frac{d_2 X^2}{n^2}) + c}.$$

It follows from (3.6) that $1 - F(L(X)) \geq \beta$ if and only if $X \leq u^*(\mu, \beta)$, where

$$u^*(\mu, \beta) = \frac{f_{n,\mu,\beta} + \frac{1}{2}d_1(z_\alpha + z_{1-\beta})^2 - h_2 + 2h_1 n^{-1} f_{n,\mu,\beta} - 2h_1 h_2 n^{-1} + (z_\alpha + z_{1-\beta})D_n^*}{(1 + 2h_1 n^{-1})^2 - d_2(z_\alpha + z_{1-\beta})^2 n^{-1}}, \qquad (A.15)$$

$$\begin{aligned} D_n^* = \Big\{ & d_0 n + n^{-1} f_{n,\mu,\beta}(nd_1 + d_2 f_{n,\mu,\beta}) + \frac{d_1^2}{4}(z_\alpha + z_{1-\beta})^2 - h_2 d_1 + c \\ & - d_0 d_2 (z_\alpha + z_{1-\beta})^2 + 4h_1 d_0 + [4d_0 h_1^2 + 2f_{n,\mu,\beta}(h_1 d_1 - h_2 d_2) \\ & + (h_2^2 d_2 - 2h_1 d_1 h_2) + (4h_1 - (z_\alpha + z_{1-\beta})^2 d_2)c_n]n^{-1} + 4h_1^2 c n^{-2} \Big\}^{1/2}. \qquad (A.16) \end{aligned}$$

It then follows from (A.5) that

$$D_n^* = (n\sigma^2)^{1/2} + \frac{1}{2}(d_1 + 2d_2\mu)z_{1-\beta} + \Big\{\frac{1}{12}(1 + 2d_2\mu)(d_1 + 2d_2\mu)(z_{1-\beta}^2 - 1)$$
$$+z_\alpha(z_\alpha - 2z_{1-\beta})(\frac{1}{8}d_1^2 - \frac{1}{2}d_0d_2) + \frac{1}{2}c - \frac{1}{2}d_1h_2 + 2d_0h_1 - d_2\mu h_2$$
$$+d_1h_1\mu\Big\}(n\sigma^2)^{-1/2} + O(n^{-1}).$$

Note that $[(1+2h_1n^{-1})^2 - d_2(z_\alpha + z_{1-\beta})^2 n^{-1}]^{-1} = 1 - [4h_1 - (z_\alpha + z_{1-\beta})^2 d_2]n^{-1} + O(n^{-2})$. Set $z_n^* = (u^*(\mu, \beta) - n\mu)/\sqrt{n\sigma^2}$. It then follows from (A.15), after some algebra, that

$$z_n^* = z_\alpha + \frac{1}{6}(d_1 + 2d_2\mu)(z_\alpha^2 - 1)\sigma^{-1}n^{-1/2}$$
$$+\Big\{ - \frac{\sigma^2 d_2}{12}[z_{1-\beta}^2(3z_{1-\beta} + 4z_\alpha) + 2z_\alpha^3 + z_{1-\beta}(-9 + 6z_\alpha^2)]$$
$$+\frac{1}{72}[(1 + 16d_2\mu(1 + d_2\mu))z_{1-\beta}^3 - z_{1-\beta}(1 - 36c + 4d_2\mu(10 + 16d_2\mu + 3d_2\mu z_\alpha^2)$$
$$+3d_1(2 + z_\alpha^2)(d_1 + 4d_2\mu)) + 3z_\alpha(12c + 12d_1d_2\mu z_\alpha^2 + d_1^2(2 - 5z_\alpha^2)$$
$$-4d_2^2\mu^2(2 + z_\alpha^2))]\Big\}(n\sigma^2)^{-1} + O(n^{-3/2}).$$

The Edgeworth expansion in Lemma 1 then leads to the choice of c given at (3.9) when $\hat{\mu}$ is replaced by μ. Since μ is unknown, μ is replaced by $\hat{\mu}$ in (3.9). It can be verified directly that resulting tolerance interval is second-order probability matching.

References

Agresti, A. and Coull, B. (1998). Approximate is better than ‘exact’ for interval estimation of binomial proportions. *Amer. Statist.* **52**, 119-126.

Bhattacharya, R. N. and Rao, R. R. (1976). *Normal Approximation And Asymptotic Expansions.* Wiley, New York.

Brown, L. D. (1986). *Fundamentals of Statistical Exponential Families with Applications in Statistical Decision Theory.* Lecture Notes-Monograph Series, Institute of Mathematical Statistics, Hayward.

Brown, L. D., Cai, T. and DasGupta, A. (2002). Confidence intervals for a binomial proportion and Edgeworth expansions. *Ann. Statist.* **30**, 160-201.

Brown, L. D., Cai, T. and DasGupta, A. (2003). Interval estimation in exponential families. *Statistica Sinica* **13**, 19-49.

Cai, T. (2005). One-sided confidence intervals in discrete distributions. *J. Statist. Plann. Inference* **131**, 63-88.

Easterling, R. G. and Weeks, D. L. (1970). An accuracy criterion for Bayesian tolerance intervals. *J. Roy. Statist. Soc. Ser. B* **32**, 236-240.

Esseen, C. G. (1945). Fourier analysis of distribution functions: a mathematical study of the Laplace-Gaussian law. *Acta Math.* **77**, 1-125.

Ghosh, J. K. (1994). *Higher Order Asymptotics.* NSF-CBMS Regional Conference Series, Institute of Mathematical Statistics, Hayward.

Ghosh, M. (2001). Comment on "Interval estimation for a binomial proportion" by L. Brown, T. Cai and A. DasGupta. *Statist. Sci.* **16**, 124-125.

Hahn, G. J. and Chandra, R. (1981). Tolerance intervals for Poisson and binomial variables. *J. Quality Tech.* **13**, 100-110.

Hahn, G. J. and Meeker, W. Q. (1991). *Statistical Intervals: A Guide for Practitioners.* Wiley Series.

Hall, P. (1982). Improving the normal approximation when constructing one-sided confidence intervals for binomial or Poisson parameters. *Biometrika* **69**, 647-52.

Kocherlakota, S. and Balakrishnan, N. (1986). Tolerance limits which control percentages in both tails: sampling from mixtures of normal distributions. *Biometrical J.* **28**, 209-217.

Krishnamoorthy, K. and Mathew, T.(2004). One-sided tolerance limits in balanced and unbalanced one-way random models based on generalized confidence intervals. *Technometrics* **46**, 44-52.

Morris, C. N. (1982). Natural exponential families with quadratic variance functions. *Ann. Statist.* **10**, 65-80.

Mukerjee, R. and Reid, N. (2001). Second-order probability matching priors for a parametric function with application to Bayesian tolerance limits. *Biometrika* **8**, 587-592.

Petrov, V. V. (1975). *Sums of Independent Random Variables.* Springer, New York.

Vangel, M. G. (1992). New methods for one-sided tolerance limits for a one-way balanced random-effects ANOVA model. *Technometrics* **34**, 176-185.

Wald, A. and Wolfowitz, J. (1946). Tolerance limits for normal distribution. *Ann. Math. Statist.* **17**, 208-215.

Wilks, S. S. (1941). Determination of sample sizes for setting tolerance limits. *Ann. Math. Statist.* **12**, 91-96.

Wilks, S. S. (1942). Statistical prediction with special reference to the problem of tolerance limits. *Ann. Math. Statist.* **13**, 400-409.

Zacks, S. (1970). Uniformly most accurate upper tolerance limits for monotone likelihood ratio families of discrete distributions. *J. Amer. Statist. Assoc.* **65**, 307-316.

Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104, U.S.A.

E-mail: tcai@wharton.upenn.edu

Institute of Statistics, National Chiao Tung University, Hsinchu, Taiwan.

E-mail: wang@stat.nctu.edu.tw

(Received October 2007; accepted April 2008)