Computational Statistics and Data Analysis 53 (2008) 17–26

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Coverage probability of prediction intervals for discrete random variables

Hsiuying Wang*

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

ARTICLE INFO

Article history:
Received 27 December 2007
Received in revised form 29 June 2008
Accepted 6 July 2008
Available online 18 July 2008

ABSTRACT

Prediction interval is a widely used tool in industrial applications to predict the distribution of future observations. The exact minimum coverage probability and the average coverage probability of the conventional prediction interval for a discrete random variable have not been accurately derived in the literature. In this paper, procedures to compute the exact minimum confidence levels and the average confidence levels of the prediction intervals for a discrete random variable are proposed. These procedures are illustrated with examples and real data applications. Based on these procedures, modified prediction intervals with the minimum coverage probability or the average coverage probability close to the nominal level can be constructed.

1. Introduction

Prediction interval (PI) is a very useful tool to predict the future observations. It is widely used in industrial applications to predict the number of defective units that will be produced during future production of a product. Such an interval would interest the purchaser of a single unit of a particular product or the manufacturer. For example, a manufacturer may wish to construct a prediction interval to bound the number of defective units of a product in the future production process. The construction of prediction intervals has been extensively studied, see Basu et al. (2003), Hall and Rieck (2001), Hamada et al. (2004) and Lawless and Fredette (2005). Compared with continuous distributions, there have been fewer investigations for discrete distributions.

Let $X_1, X_2, \ldots, X_n$ be an observed random sample of size n from a discrete distribution with a probability mass function $f(x;\theta)$, where θ is an unknown parameter. Let $Y_1, \ldots, Y_m$ be a future random sample of size m from the same distribution. Assume that the future sample $Y_1, \ldots, Y_m$ is drawn independently of the past sample $X_1, \ldots, X_n$. Let X be a function of $X_1, \ldots, X_n$ and have a probability mass function $f_n(x;\theta)$. Let Y be a function of $Y_1, \ldots, Y_m$ and have a probability mass function $f_m(y;\theta)$. Let $L(X)$ and $U(X)$ be two statistics based on the observed sample. If $L(X)$ and $U(X)$ are determined so that

$$P(L(X) \le Y \le U(X)) = \gamma, \tag{1}$$

then $[L(X), U(X)]$ is called a level γ prediction interval.

A level γ prediction interval is usually given by a large sample approximate method. When the sample size is large, the coverage probability of the prediction interval constructed by a large sample approximation method may be close to the nominal level. However, when the sample size is not large enough, the coverage probability of the prediction interval may be far away from the nominal coefficient. Inaccurate interval prediction may cause financial loss in real applications. Therefore, establishing an approach to derive the exact confidence level of a prediction interval for fixed sample size is an important issue in practice.

* Tel.: +886 3 5712121x56813.
E-mail address: wang@stat.nctu.edu.tw.

0167-9473/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2008.07.017

For a claimed γ level prediction interval, the true coverage probability depends on the true parameter. Since the true parameter value is unknown, we can not know the true coverage probability. However, we do know that it is greater than the minimum coverage probability. Hence, if the minimum coverage probability can be derived, it can provide, at least, a conservative estimation for the coverage probability. The exact value of the minimum coverage probability has not been accurately calculated before and can only be approximated by a simulation study.

This problem has not only arisen in prediction intervals, but also for confidence intervals. Wang (2007a, in press) proposed approaches to calculate the minimum coverage probabilities for confidence intervals of a binomial proportion and the unknown parameter for one-parameter discrete distributions. Using Wang's approach, the point in the parameter space in which the minimum coverage probability occurs as well as the minimum coverage probability can be exactly derived. Since the estimation goal is different between the confidence interval and the prediction interval, this approach can not be directly applied to the prediction interval.

In this article, procedures are proposed to calculate the minimum coverage probability and average coverage probability of prediction intervals for a discrete random variable. By using the proposed procedures, we are able to obtain an improved prediction interval with their true minimum coverage probability or average coverage probability very close to the nominal levels.

In addition, besides the results under the frequentist setup, we also propose a procedure for computing the coverage probability with respect to a prior distribution on the parameter space under the Bayesian framework. Although this paper mainly focuses on the inference from the frequentist point of view, the inference can be applied to the Bayesian setup.

The paper is organized as follows. The methodologies of computing the minimum coverage probability and the average coverage probability of a prediction interval are given in Section 2. In Section 3, the procedure for calculating the coverage probability with respect to a prior distribution under the Bayesian framework is provided. The most widely used prediction intervals for a binomial and Poisson random variables are introduced in Section 4. Proposed methodologies for applying these prediction intervals are illustrated by numerical studies in this section as well. In Section 5, modified prediction intervals with required minimum or average confidence levels are proposed. In Section 6, the modified prediction interval is compared with the conventional prediction interval in the context of an industrial application.

2. Minimum coverage probability and average coverage probability

In a one-parameter discrete distribution, the coverage probability of a prediction interval is a variable function of the parameter θ. In this section, a procedure is developed to calculate the exact minimum coverage probability of a prediction interval $[L(X), U(X)]$ of a discrete random variable.

The proposed procedure to calculate the minimum coverage probability is as follows.

Procedure 1. *Computing minimum coverage probability.*

For a level γ prediction interval $[L(X), U(X)]$ of a discrete random variable Y with parameter θ, the minimum coverage probability can be obtained by the following steps. Assume that the parameter space is $\Omega = (a, b)$, where $-\infty < a < b < \infty$.

Step 1. Find the set $A = \{(X, Y) : L(X) \leq Y \leq U(X)\}$.

Step 2. Let

$$g(\theta) = \sum_{(X,Y)\in A} f_n(X;\theta) f_m(Y;\theta). \tag{2}$$

Numerically solve the equation

$$\frac{\partial}{\partial\theta} g(\theta) = 0. \tag{3}$$

Let $\{s_1, \ldots, s_k\}$ be the set of the solutions in (a, b).

Step 3. Compute the $k+2$ coverage probabilities $\sum_{(X,Y)\in A} f_n(X;\theta) f_m(Y;\theta)$ at $\theta = a$, b and s_i, $i = 1, \ldots, k$. The minimum coverage probability of the prediction interval is the smallest value of these coverage probabilities.

The first step in the procedure is to collect the observations of X and Y with positive probabilities used in calculating the coverage probability of the prediction interval. The second step is to derive the possible points in the parameter space at which the local minima of the coverage probability occur. The third step is to calculate the global minimum coverage probability.

Instead of reporting the minimum coverage probability, the average coverage probability, which is the average value of the coverage probability on the parameter space, is also an important index for evaluating the performance of a prediction interval. Compared with the minimum coverage probability, which is the behavior at a point or several points in the parameter space, the average coverage probability can provide a more objective evaluation for a prediction interval.

For a parameter space $\Omega = (a, b)$, where $-\infty < a < b < \infty$, the average coverage probability is defined by

$$\frac{1}{(b-a)} \int_{\Omega} P_\theta(L(X) \leq Y \leq U(X)) \mathrm{d}\theta. \tag{4}$$

This value takes the average of the coverage probability in the parameter space Ω.

A procedure to calculate the average coverage probability is as follows.

Procedure 2. *Computing the average coverage probability.*

For a level γ prediction interval $[L(X), U(X)]$, the average coverage probability can be derived by the following two steps:
Step 1: Follow step 1 in Procedure 1.
Step 2: Compute the summation

$$\frac{1}{(b-a)} \sum_{(x,y)\in A} \int_{\Omega} f_n(X;\theta) f_m(Y;\theta) \mathrm{d}\theta, \tag{5}$$

which is the exact average coverage probability of the prediction interval $[L(X), U(X)]$.

3. Coverage probability under the Bayesian framework

We mainly discuss calculation of the coverage probability under the frequentist viewpoint in Section 2. If it is known that there exists a prior for the parameter space, we need to consider the coverage probability of the prediction interval under the Bayesian framework. In fact, the average coverage probability discussed in Section 2 can be viewed as the coverage probability under the Bayesian framework with respect to the uniform prior. In the Bayesian setup, it is assumed that there exists a prior $\eta(\theta)$, which conveys information about the parameters. The coverage probability under this setup should be the coverage probability with respect to the prior, which is

$$\int_{\Omega} P_\theta(L(X) \le Y \le U(X)) \eta(\theta) \mathrm{d}\theta. \tag{6}$$

The coverage probability (6) reflects the overall performance of a prediction level under a given prior $\eta(\theta)$ on the parameter space.

By a similar argument as in Section 2, the procedure for calculating the coverage probability under the Bayesian setup is as follows.

Procedure 3. *Computing coverage probability under Bayesian framework.*

Assume that the parameter space is known to have a prior $\eta(\theta)$. For a prediction interval $[L(X), U(X)]$, the coverage probability with respect to the prior $\eta(\theta)$ can be derived by the following two steps:
Step 1: Follow step 1 in Procedure 1.
Step 2: Compute the summation

$$\sum_{(x,y)\in A} \int_{\Omega} f_n(X;\theta) f_m(Y;\theta) \eta(\theta) \mathrm{d}\theta, \tag{7}$$

which is the exact coverage probability of the prediction interval $[L(X), U(X)]$ with respect to the prior.

4. Existing prediction intervals

In this paper, we will use the proposed Procedures 1 and 2 to calculate the minimum and coverage probabilities of the widely-used prediction intervals for binomial and Poisson random variables.

4.1. Binomial example

The widely used prediction interval for a binomial random variable is constructed by Nelson (1982) and reviewed in Hahn and Meeker (1991). Suppose the past data consist of X out of n trials from a $B(n, p)$ distribution. Let Y be the future number of successes out of m trials from a $B(m, p)$ distribution. A large-sample approximate γ two-sided prediction interval $[L(X), U(X)]$ for the future number Y of occurrences based on the observed value of the number X of past occurrences for the binomial distribution constructed by Nelson (1982) is

$$\hat{Y} \pm z_{(1+\gamma)/2} (m\hat{p}(1-\hat{p})(m+n)/n)^{1/2}, \tag{8}$$

where $\hat{Y} = m\hat{p} = m(x/n)$ when X, $n-X$, Y and $m-Y$ all are large.

Fig. 1 shows the performance of the coverage probabilities for the prediction interval (8) for $n = 50$ and different m. The performances for other n have similar patterns as the case of $n = 50$. The coverage probability is worse when p is in a domain close to the boundaries. For the cases of $m = 5$ and $m = 10$, there are three local minima. In the case of $m = 100$, the coverage probability is less than 0.95 for the entire range of the parameter space.

By applying the procedures in Section 2, the minimum coverage probability and the average coverage probability of the prediction interval can be derived. The procedures for calculating the minimum coverage probability and the average coverage probability are illustrated by examples.

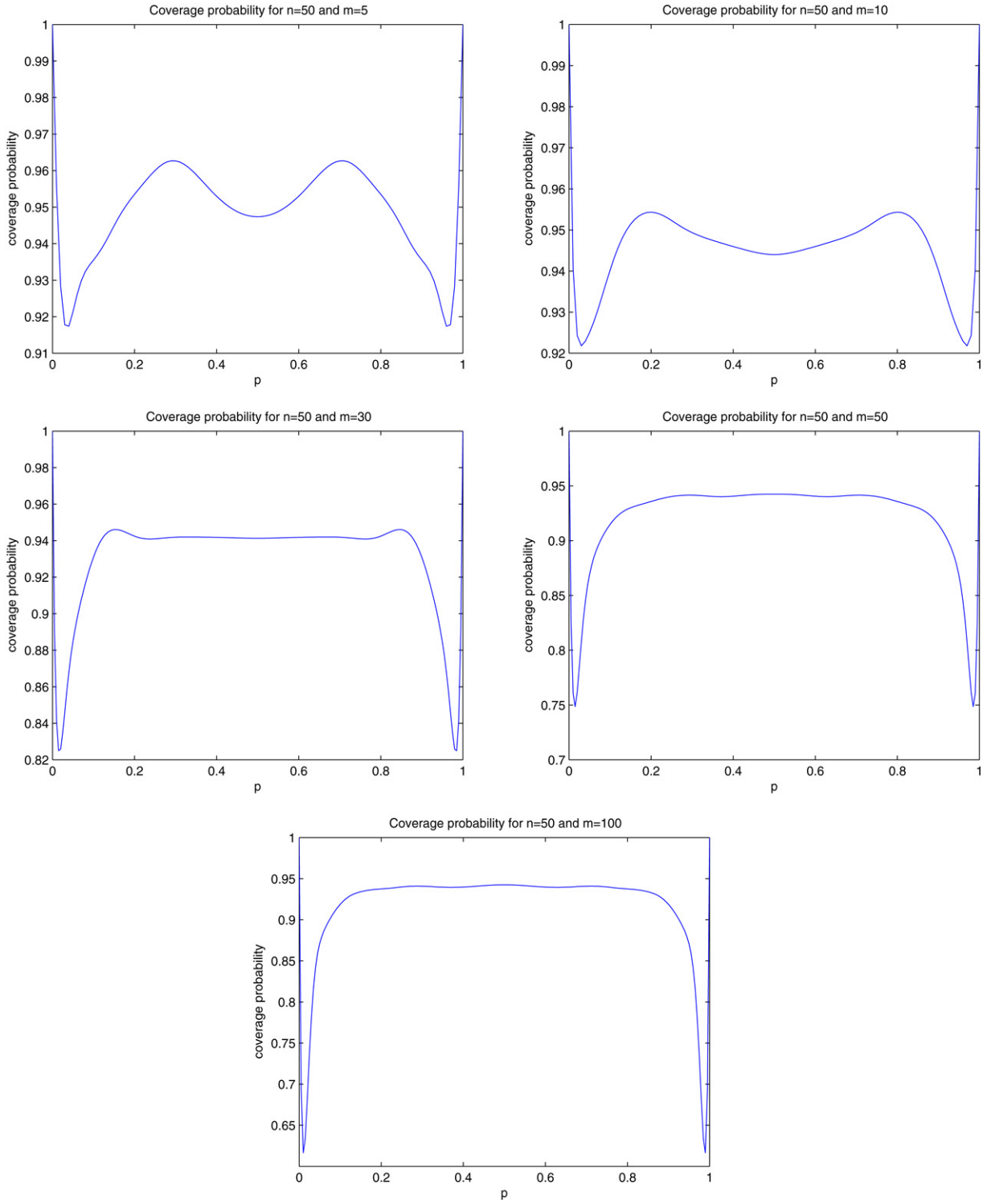

Fig. 1. Coverage probabilities of the 95% level two-sided prediction intervals for the binomial distributions with $n = 50$ and different m.

Example 1. For $n = 4$ and $m = 2$ and the 0.95 two-sided prediction interval (8), the set A is $\{(0, 0), (1, 0), (1, 1), (2, 0), (2, 1), (2, 2), (3, 1), (3, 2), (4, 2)\}$. There is only one solution for $\frac{\partial}{\partial p}\sum_{(x,y)\in A}\binom{n}{x}p^x(1-p)^{n-x}\binom{m}{y}p^y(1-p)^{m-y} = 0$ by a numerical calculation, which occurs at $p = 0.5$. Calculate the probabilities $\sum_{(x,y)\in A}\binom{n}{x}p^x(1-p)^{n-x}\binom{m}{y}p^y(1-p)^{m-y}$ at

Table 1
The minimum coverage probability and the average coverage probability for the 0.95 two-sided prediction intervals (8) for different sample sizes n when $m = 20$

n	Minimum coverage probability of PI (8)	Average coverage probability of PI (8)
10	0.6147	0.8350
20	0.7484	0.8958
30	0.8092	0.9210
40	0.8506	0.9319
50	0.8544	0.9333
60	0.8777	0.9405
70	0.8958	0.9382
80	0.9082	0.9450
90	0.9105	0.9426
100	0.9214	0.9462

Table 2
The minimum coverage probability and the average coverage probability for the 0.95 two-sided prediction interval (8) for different sample sizes n when $m = 30$

n	Minimum coverage probability of PI (8)	Average coverage probability of PI (8)
10	0.5292	0.8176
20	0.6776	0.8883
30	0.7487	0.9106
40	0.7923	0.9191
50	0.8259	0.9306
60	0.8489	0.9369
70	0.8420	0.9324
80	0.8645	0.9382
90	0.8778	0.9417
100	0.8908	0.9430

$p = 0.5$, 0 and 1, which are 0.7813, 1 and 1. Thus, the minimum coverage probability is 0.7813. By applying Procedure 2, the average coverage probability is 0.8286.

Example 2. For $n = 5$ and $m = 3$ and the 0.95 two-sided prediction interval (8), the set A is $\{(0, 0), (1, 0), (1, 1), (1, 2), (2, 0), (2, 1), (2, 2), (2, 3), (3, 0), (3, 1), (3, 2), (3, 3), (4, 1), (4, 2), (4, 3), (5, 3)\}$. The solutions of $\frac{\partial}{\partial p}\sum_{(x,y)\in A}\binom{n}{x}p^x(1-p)^{n-x}\binom{m}{y}p^y(1-p)^{m-y} = 0$ are $p = 0.152, 0.5$ and 0.848 by a numerical calculation. Then calculate the probability $\sum_{(x,y)\in A}\binom{n}{x}p^x(1-p)^{n-x}\binom{m}{y}p^y(1-p)^{m-y}$ at $p = 0$, 0.152 and 0.5, which are 1, 0.826 and 0.9062, respectively. Note that since the coverage probability is symmetric to $p = 0.5$, we only need to calculate the coverage probability when p is less than or equal to 0.5.

By step 3 in Procedure 1, the minimum coverage probability is 0.826. By applying Procedure 2, the average coverage probability is 0.8730.

Tables 1 and 2 list the minimum coverage probabilities and the average coverage probabilities for different sample sizes n corresponding to $m = 20$ and $m = 30$, respectively, for the prediction interval (8).

In these two tables, the minimum coverage probability is smaller than the nominal level and increases with n. The minimum coverage probability may be very close to the nominal level when n goes to infinity. Compared with the minimum coverage probability, the average coverage probability is closer to the nominal level, although it is still smaller than the nominal level. In the comparison of the two cases of $m = 20$ and $m = 30$, the minimum coverage probability and the average coverage probability for $m = 20$ are always larger than those for $m = 30$. This may be due to the greater inaccuracy in predicting the distribution of observations farther in the future compared to observations in the near future using this prediction interval.

From both tables, the average coverage probabilities may be above 0.94 when n is greater than 100 and the minimum coverage probability is below 0.93 for the n listed in the tables. This indicates that the performance of the prediction interval (8) is not good unless the sample size n is large. In practical applications, the sample size may not be large, which will lead to inaccurate estimation when using (8) to predict the distribution of the future observations. Based on the above results, it would be helpful for practical users if we could provide an improved prediction interval that can attain the required nominal level for any sample size.

The proposed improved prediction interval, in which the minimum coverage probability or the average coverage probability is close to the nominal level based on Procedures 1 and 2, is given in Section 4.

Besides evaluating the performance of the existing intervals under the frequentist framework, we also investigate their performance under the Bayesian setup. Suppose that the parameter space has a conjugate prior following a $Beta(\alpha, \beta)$ distribution. The coverage probabilities calculated by Procedure 3 with respect to different α and β are listed in Table 3.

Table 3
The coverage probability for the 0.95 two-sided prediction interval (8) with respect to the prior $beta(\alpha, \beta)$ for different m when $n = 30$

(α, β)	(5, 5)	(5,10)	(5, 20)	(5, 50)	(10,10)	(10, 20)	(10, 50)
$m = 10$	0.9414	0.9401	0.9336	0.9058	0.9421	0.9408	0.9327
$m = 20$	0.9409	0.9386	0.9282	0.8836	0.9420	0.9399	0.9271
$m = 30$	0.9384	0.9350	0.9192	0.8648	0.9399	0.9374	0.9150
$m = 40$	0.9374	0.9331	0.9166	0.8605	0.9396	0.9354	0.9136

From the calculation results in Table 3 and additional results that we do not present in this paper, the coverage probability for the prediction interval (8) is closer to the nominal level when α and β in the prior distribution $beta(\alpha, \beta)$ are chosen to be the same than the case when α and β are not the same. When the absolute difference of α and β, $|\alpha - \beta|$, increases, the coverage probability decreases from the calculated results. The phenomenon can be explained from Fig. 1. The beta distribution has the property that the density function only has a mode and is symmetric about the mode when α and β are the same, and the density function is skewed when α and β are not the same. By this fact, the skewness increases as the absolute value of the difference between α and β increases. Note that the coverage probability drops when p is in a neighborhood near the boundaries in Fig. 1. It then follows that the coverage probability with respect to the beta distribution decreases as the absolute value of the difference between α and β increases.

4.2. Poisson example

The widely used prediction interval for a Poisson random variable is constructed by Nelson (1982). Suppose the past data consist of X observed occurrences in an observation of length n with rate λ. Let Y be the number of occurrences in a future observation of length m with the same rate λ. Assume that X and Y are independent Poisson random variables. A large-sample approximate γ two-sided prediction interval $[L(X), U(X)]$ for the future number Y is

$$m\hat{\lambda} \pm z_{(1+\gamma)/2} m \left(\hat{\lambda}\left(\frac{1}{n} + \frac{1}{m}\right)\right)^{1/2}, \tag{9}$$

where $\hat{\lambda} = X/n$.

For the Poisson random variable, the support of the possible observations is infinite and the natural parameter space is $(0, \infty)$. Therefore, for the Poisson distribution, if we consider the natural parameter space, the possible observation is infinite and set A in Procedure 1 is not bounded. In this case, since set A is unbounded, we can not apply Procedure 1 to derive the minimum coverage probability because $g(\theta)$ in Step 2 cannot be derived. However, in real applications, the parameter space can be assumed to be bounded. And the bounds can be estimated from the empirical knowledge and data information. In Examples 3 and 4, the parameter space is assumed to be restricted to (0, 1) and (0.5, 2), respectively. Although the support of the Poisson random variable is infinite, we can use the approximation approach by considering a finite support of the random variable because the probability of the random variable being greater than some constant is small enough to be neglected. In this case, since the parameter space is restricted to (0, 1) and (0.5, 2), we can approximate the coverage probability by considering the possible observations from 1 to 100. The probability of the observations being greater than 100 is very small if $\lambda \in (0, 1)$ or $\lambda \in (0.5, 2)$ when n and m are not large. The derivations of the minimum coverage probability and the average coverage probability of the prediction interval are illustrated by Examples 3 and 4.

Example 3. Assume that the parameter space of λ is restricted to (0, 1). For $n = 4$ and $m = 2$ and the 0.95 two-sided prediction interval (9), we calculate set A and the probability function $\sum_{(x,y)\in A}(e^{-n\lambda}(n\lambda)^x/(x!))(e^{-m\lambda}(m\lambda)^y/(y!))$. There is only one solution for $\frac{\partial}{\partial\lambda}\sum_{(x,y)\in A}(e^{-n\lambda}(n\lambda)^x/(x!))(e^{-m\lambda}(m\lambda)^y/(y!)) = 0$ by a numerical calculation, which occurs at $\lambda = 0.2$. There is a minimum coverage probability of 0.8489 occurring at this point. The coverage probabilities are 1 and 0.925 corresponding to $\lambda = 0$ and 1. Therefore, by Procedure 1, the minimum coverage probability is the minimum value of these three coverage probabilities, which is 0.8489. By applying Procedure 2, the average coverage probability is 0.8932 with respect to the prior $\eta(\lambda) = 1$ for $\lambda \in (0, 1)$.

Example 4. Assume that the parameter space of λ is restricted to (0.5, 2). For $n = 5$ and $m = 3$ and the 0.95 two-sided prediction interval (9), we calculate set A and the probability function $\sum_{(x,y)\in A}(e^{-n\lambda}(n\lambda)^x/(x!))(e^{-m\lambda}(m\lambda)^y/(y!))$. The probability function is an increasing function of λ. The minimum coverage probability occurs at $\lambda = 0.5$, which is 0.8791. By applying Procedure 2, the average coverage probability is 0.9267 with respect to $\eta(\lambda) = 2/3$ for $\lambda \in (0.5, 2)$.

Fig. 2 shows the performances of the coverage probabilities for the prediction interval (9) for some n and m when the parameter space is restricted to (0, 1). The coverage probability is worse when p is in a domain close to 0 and it is increasing when p is greater than a constant.

Tables 4 and 5 list the minimum coverage probabilities and the average coverage probabilities for some sample sizes n and m, respectively, for the prediction interval (9).

The performances of the minimum coverage probability and average coverage probability for the Poisson prediction interval are similar to the binomial prediction interval.

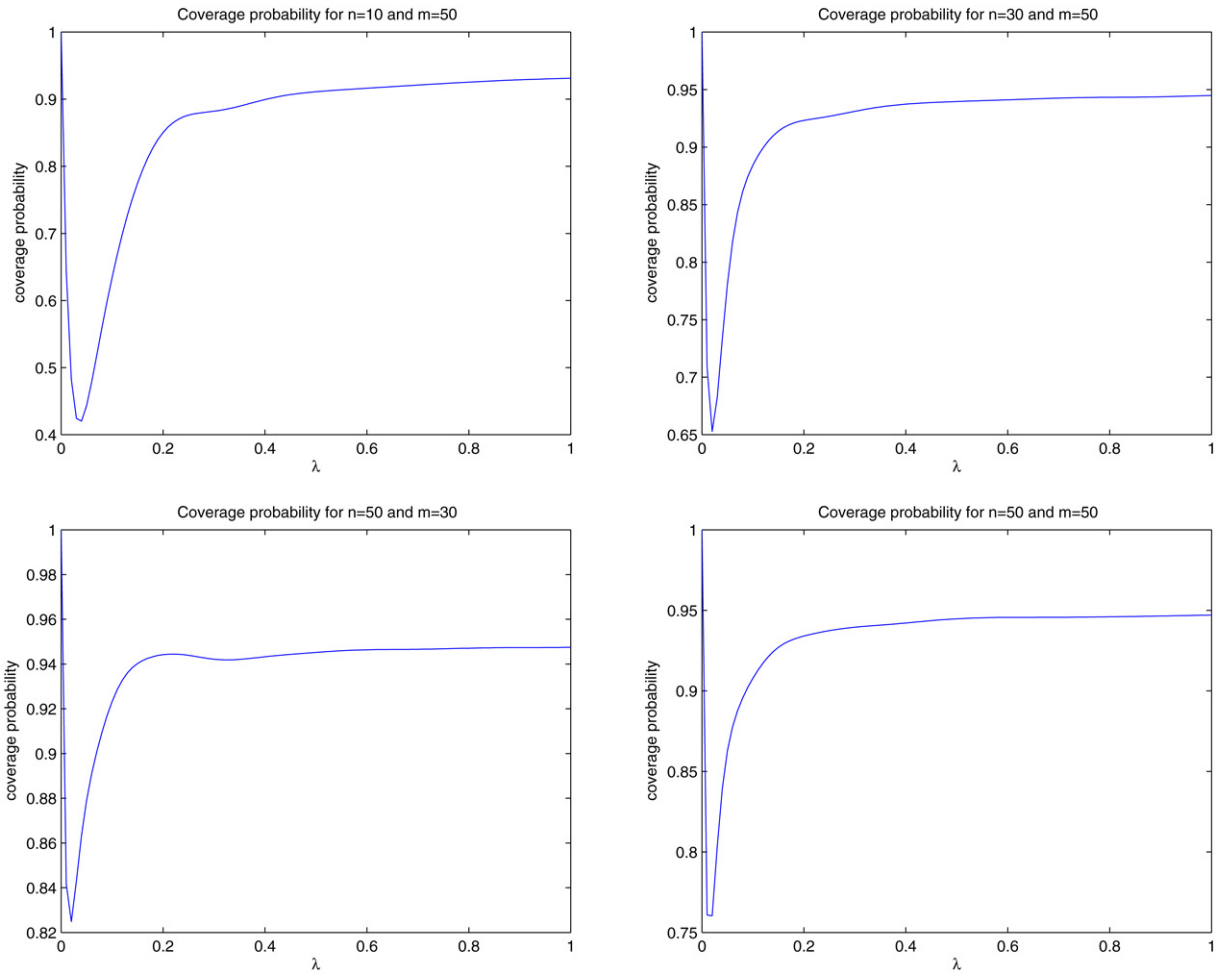

Fig. 2. Coverage probabilities of the 95% level two-sided prediction intervals for the Poisson distributions for different n and m.

Table 4
The minimum coverage probability and the average coverage probability for the 0.95 two-sided prediction intervals (9) for different sample sizes n when $m = 20$ for $\lambda \in (0, 1)$

n	Minimum coverage probability of PI (9)	Average coverage probability of PI (9)
10	0.6164	0.8727
20	0.7493	0.9135
30	0.808	0.9313
40	0.8496	0.9374
50	0.852	0.9383
60	0.882	0.9430

Table 5
The minimum coverage probability and the average coverage probability for the 0.95 two-sided prediction intervals (9) for different sample sizes n when $m = 30$ for $\lambda \in (0.5, 2)$

n	Minimum coverage probability of PI (9)	Average coverage probability of PI (9)
5	0.8731	0.9135
10	0.9105	0.9321
20	0.9341	0.9431
30	0.9395	0.9464
40	0.9409	0.9465

Table 6
The suggested k value in (10) such that the average coverage probability is close to the nominal level 0.95 as well as their corresponding minimum coverage probability and average coverage probability for different sample sizes n when $m = 20$

n	k	Minimum coverage probability	Average coverage probability
20	3.66	0.7534	0.9505
30	2.36	0.816	0.9509
40	2.26	0.8568	0.9521
50	2.16	0.8762	0.9543
60	2.11	0.8777	0.9508
70	2.06	0.8958	0.9516
80	2.01	0.9117	0.9502
90	2.01	0.9113	0.9488
100	2.01	0.9214	0.9508

Table 7
The suggested k value in (10) such that the average coverage probability is close to the nominal level 0.95 as well as their corresponding minimum coverage probability and average coverage probability for different sample sizes n when $m = 30$

n	k	Minimum coverage probability	Average coverage probability
20	3.96	0.6879	0.9412
30	2.66	0.7591	0.9526
40	2.26	0.7964	0.9494
50	2.16	0.8264	0.9480
60	2.16	0.8562	0.9532
70	2.06	0.8812	0.9514
80	2.06	0.8645	0.9485
90	2.06	0.8839	0.9508
100	2.06	0.9004	0.9525

5. Modified prediction intervals

Based on the numerical results in Section 4, the coverage probabilities for the existing prediction intervals (8) and (9) are lower than the nominal level 0.95. In this section, improved prediction intervals are constructed based on Procedure 1 or Procedure 2 such that their minimum or average confidence levels can be close to the nominal level.

First, we rewrite (8) as

$$\hat{Y} \pm k(m\hat{p}(1-\hat{p})(m+n)/n)^{1/2}. \tag{10}$$

The $z_{(1+\gamma)/2}$ in (8) is replaced by k in (10).

The approach is to choose an appropriate k such that the minimum or average coverage probability of the constructed prediction intervals are equal to the nominal level. For fixed sample sizes n and m, if we adopt the minimum coverage probability criterion, by Procedure 1, we can calculate the minimum coverage probability for different k and choose an appropriate k such that its minimum coverage probability is close to the nominal level. If we adopt the average coverage probability criterion, by Procedure 2, we can calculate the average coverage probability for different k and choose an appropriate k such that its average coverage probability is close to the nominal level. The adjusted k value does not only depend on the criterion we adopt, but also depends on the sample sizes n and m. Therefore, for any given sample size (n, m), we can find the improved prediction interval with the exact minimum coverage probability or the exact average coverage probability close to a given nominal coverage probability.

Since both minimum and average coverage probability criterions are given and two sets of k are suggested, we may be interested in which criterion we should employ. If we need a conservative prediction interval that guarantees the coverage probability is greater than the nominal level in any circumstances, then we can use the modified prediction interval fitted to the minimum coverage probability criterion. If we only require a prediction interval attaining the nominal level on average with respect to a prior which can have a smaller expected length than the prediction interval fitted to the minimum coverage probability, then we can use the modified prediction interval fitted to the average coverage probability criterion.

For the two cases of $m = 20$ and $m = 30$ in Section 3, the appropriate k such that the minimum coverage probability is close to the nominal level 0.95 need to be very large. Thus, in these two cases, we do not suggest adopting Procedure 1 to construct the modified prediction intervals. The appropriate k for different sample sizes, chosen by adopting Procedure 2 such that the average coverage probability is close to 0.95, are presented in Tables 6–8.

For the Poisson distribution, since the selection of modified intervals depends on the restricted parameter space, we do not list the suggested modified intervals.

For a fixed sample size, the selection of a modified interval satisfying the average coverage probability close to the nominal level needs to rely on numerical computation because it involves complicated calculations, such as integrations. The programs for computing the minimum coverage probability, average coverage probability and selection of k for the modified interval are available from the author upon request.

Table 8
The suggested k value in (10) such that the average coverage probability is close to the nominal level 0.95 for different sample sizes n and m

n	m									
	10	20	30	40	50	60	70	80	90	100
10	3.96	3.96	3.96	3.96	3.96	3.96	3.96	3.96	3.96	3.96
20	2.66	3.66	3.96	3.96	3.96	3.96	3.96	3.96	3.96	3.96
30	2.26	2.36	2.66	2.76	2.96	3.16	3.36	3.56	3.76	3.96
40	2.06	2.26	2.26	2.36	2.46	2.46	2.56	2.56	2.66	2.66
50	2.06	2.16	2.16	2.26	2.26	2.36	2.36	2.36	2.36	2.46
60	2.06	2.16	2.11	2.16	2.16	2.16	2.26	2.26	2.26	2.26
70	2.06	2.06	2.06	2.06	2.16	2.16	2.16	2.16	2.26	2.26
80	2.06	2.01	2.06	2.06	2.16	2.16	2.16	2.16	2.16	2.16
90	1.96	2.01	2.06	2.06	2.06	2.06	2.06	2.06	2.16	2.16
100	1.96	2.01	2.06	2.06	2.06	2.06	2.06	2.06	2.06	2.06

6. An illustrative example

The proposed prediction intervals are illustrated by a real-data example and are compared with the existing prediction interval using this example.

The example is taken from a semiconductor manufacturing process. The location of chips on a wafer is measured on 30 wafers. On each wafer 50 chips are measured and the number of defective chips is recorded. The defective number follows a binomial distribution, $B(50, p)$. The data can be obtained from NIST/SEMATECH e-Handbook of Statistical Methods: http://www.itl.nist.gov/div898/handbook/pmc/section3/pmc332.htm

Sample Number	Fraction Defective	Sample Number	Fraction Defective	Sample Number	Fraction Defective
1	.24	11	.10	21	.40
2	.30	12	.12	22	.36
3	.16	13	.34	23	.48
4	.20	14	.24	24	.30
5	.08	15	.44	25	.18
6	.14	16	.16	26	.24
7	.32	17	.20	27	.14
8	.18	18	.10	28	.26
9	.28	19	.26	29	.18
10	.20	20	.22	30	.12

In this example, we use the first 20 defective rates to calculate the two intervals, the existing prediction and modified prediction intervals, in order to predict the average defective rate of the last 10 defective rates. First, there are about 214 defective chips in the 1000 chips for the first 20 wafers. By using this, we need to derive a prediction interval for the number of defective chips in the 500 chips for the last 10 wafers. By employing the prediction interval (8), let $n = 1000$, $m = 500$ and $X = 214$, then the prediction interval based on (8) is (84.9861, 129.0139). According to the real data, the number of the defective chips in the last 10 wafers is $(0.4 + \cdots + 0.12) \times 50 = 133$, which does not fall into the prediction interval based on (8). We can consider the modified prediction interval such that the minimum coverage probability is close to 0.95. In this case, using Procedure 1, k in this modified prediction interval (10) can be chosen as 2.43, and its corresponding minimum coverage probability is 0.9491. And the modified prediction interval is (79.68, 134.32), which contains the true defective number 133 for the future 500 chips.

This example shows that the modified prediction interval, which is adjusted according to the sample sizes, can outperform the conventional prediction interval in this real data example. These modified intervals can obtain the required coverage probability regardless of sample size, because we modify k depending on the sample size when constructing the prediction interval. The proposed methodology is certainly of great help in providing a more precise prediction interval.

Acknowledgements

The author thanks Professor T. Tony Cai at the University of Pennsylvania for helpful discussions and Professor Fugee Tsung at Hong Kong University of Science and Technology for providing the data example. The author also thanks the referee for helpful comments.

References

Basu, R., Ghosh, J.K., Mukerjee, R., 2003. Empirical Bayes prediction intervals in a normal regression model: Higher order asymptotics. Statist. Probab. Lett. 63, 197–203.

Hahn, G.J., Meeker, W.Q., 1991. Statistical Intervals: A Guide for Practitioners. Wiley Series.
Hall, P., Rieck, A., 2001. Improving coverage accuracy of nonparametric prediction intervals. J. R. Stat. Soc. Ser. B Stat. Methodol. 63, 717–725.
Hamada, M., Johnson, V., Moore, L.M., 2004. Wendelberger, Joanne Bayesian prediction intervals and their relationship to tolerance intervals. Technometrics 46, 452–459.
Lawless, J.F., Fredette, M., 2005. Frequentist prediction intervals and predictive distributions. Biometrika 92, 529–542.
Nelson, W., 1982. Applied Life Data Analysis. Wiley, NY.
Wang, H., 2007a. Exact confidence coefficients of confidence intervals for a binomial proportion. Statist. Sinica 17, 361–368.
Wang, H., 2007b. Exact average coverage probabilities and confidence coefficients of confidence intervals for discrete distributions. Stat. Comput. (in press).