國立交通大學

統計學學系

碩士論文

二項式分配線性組合之預測區間

Prediction Interval for a Linear Function of Binomial Random Variables

研究生： 張鐿瀞

指導教授： 王秀瑛 教授

中華民國 一〇一年 六月

二項式分配線性組合之預測區間

研究生： 張鐿瀞　　指導教授： 王秀瑛 教授

國立交通大學理學院

統計學研究所

摘要

在預測未來觀測值的研究方法裡，預測區間是一個非常實用的方法。不論是在工業上的應用或是醫學領域上的應用，預測區間都能夠實際的應用在這些領域中。由於現有的文獻研究著重於連續型的預測區間以及單一離散型變量的預測區間之應用，這些現有的方法，可能無法直接應用到多變量的狀況。因此，在這篇論文裡，我們所探討的預測區間是在多變量的伯努力分配變數之線性組合的應用。我們主要考量兩大方向：(1)在不同變數下，參數間具有某一相關，(2)在不同變數下，參數之間無特定相關。這個研究方法主要是延伸 Wang(2010)所提出的預測區間的方法。我們並以模擬結果來檢驗所提出的預測區間之優劣。

關鍵字：預測區間，覆蓋率，二項式分配

Prediction Interval for a Linear Function of Binomial Random Variables

Student: Yijing Chang
Advisor: Hsiuying Wang
Institute of Statistics
National Chiao Tung University
Hsinchu, Taiwan

Abstract

The prediction interval is a useful tool to predict the future observations. It can be widely used in industrial and medical applications. Although there are some previous studies focusing on the construction of prediction intervals for continuous distribution or some previous studies focusing on the construction of prediction intervals for discrete distribution of single variable, these methods cannot be directly applied to construct prediction interval for functions of multiple variables. In this thesis, we investigate prediction intervals for a linear function of binomial random variables. We consider two cases: (1) there is a relationship of parameters for different variables, and (2) there is no any relationship of parameters for different variables. The proposed method is an extension of Wang (2010). A simulation result shows the performance of the proposed method.

Key words: Prediction Interval, coverage probability, binomial distribution.

誌謝

很榮幸能成為王秀瑛 教授的指導學生之一，從碩士二年級開始與老師一起做研究。在這一年的相處之中，老師不只提供了我知識上的學習也教導了我許許多多處事的態度，給我精神上的支援。遇到了問題，要思考要多元，更要嘗試。很感謝秀瑛老師。 感謝口試委員：黃榮臣 教授、鄭少為 教授以及洪慧念教授，在口試中，針對我的論文，細心的提出更佳的建議以及修改的方向，使我在此篇研究結果能更佳的完整。

在碩士班兩年間，我要特別感謝研究室的同學，每當我在程式上遇到難題，研究室的同學都會熱心的、仔細的和我一起討論，協助我一起解決問題，才使得我的論文研究能夠如期完成。也謝謝以前碩士班的學長姊，在我論文研究期間供我許多的幫助。

最後，我要感謝最支持我的家人以及摯友們，在我忙碌、壓力大時，總會及時的給予我鼓勵與安慰。讓我更能專注在研究上。未來，離開了學校，我會更加努力，更加充實自己。也給自己一個期許，希望在未來，能夠成為一名更多元人才。

張鐿瀞 謹誌于

國立交通大學統計學研究所

中華民國 101 年六月

Contents

1 INTRODUCTION

The prediction interval (PI) is an important tool to predict the future observations. It is widely used in industrial applications to predict the number of defective units which will be produced during future production of a product (Wang 2008). In medical applications, it can be used in predicting disease count (Wang 2010).

Most of the previous study focus on the construction of prediction intervals for continuous distribution, see Basu, Ghosh and Mukerje (2003), Hall and Rieck (2001), Hamada, Johnson and Moore (2004) and Lawless and Fredette (2005). Nevertheless, compared with the continuous distributions, there are some studies for discrete distributions. Nelson (1982) proposed a widely used closed form prediction interval for a discrete random variable. Another useful prediction interval with a closed form for a discrete distribution was proposed by Bain and Patel (1993). Among these discrete distributions, the binomial distribution is a useful discrete distribution in many real application fields. Wang (2007a, 2008, 2009) proposed procedures to calculate the coverage probability of confidence intervals for a binomial proportion of a binomial distribution and the coverage probability of prediction intervals for a binomial random variable. Most of the previous studies focus on prediction interval for a single random variable. Although the above literatures provide methods for constructions of prediction intervals for discrete distributions, they cannot be applied to construct prediction intervals for functions of multiple variables. The multivariate binomial distribution can be applied in many real applications. For example, we consider

predicting the number of sickbeds arrangement between different departments in a hospital. Suppose the numbers of sickbeds follows a binomial distribution and we model the number of sickbeds needed for different departments with a multiple binomial distribution. Thus, in this thesis, we investigate prediction intervals for the binomial distribution with multiple variables.

In this thesis, we extend the prediction interval proposed by Wang (2010) from a single variable to multiple variables. Using the proposed method, we are able to obtain a suitable interval to predict a linear combination of multiple variables. In addition, we consider two cases that the parameters are related to the binomial distributions and the parameters are not related. The first case can be applied to predict the number of sickbed arrangement between different departments. A more accurate prediction interval can not only reduce the cost but also can increase the utilization rate of sickbeds. Such an interval would interest the hospital operator for the configuration management. For instance, a hospital operator may wish to construct a prediction interval to know how to arrange the number of sickbeds to different department such that they can be used sufficiently. In industrial applications, it can predict the number of defective units which need to be repaired and the number units should be supplied to clients. That can bring lots of effects, such as increasing the customer satisfaction and raising the company image.

The rest of the thesis is organized as follows. Section 2 reviews the related works and the prediction interval proposed by Wang (2010). In Section 3, we present the

proposed methods used by this study. Section 4 displays the simulation results of our study. A conclusion is summarized in Section 5.

2 PRELIMINARY

In this section, first, we introduce the definition of prediction interval, and the existing prediction intervals for the binomial distribution.

2.1 Prediction Interval

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 from the past sample $X_1, X_2, \ldots, X_n$. Let X be a function of $X_1, X_2, \ldots, X_n$ 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 samples. If $L(X)$ and $U(X)$ are determined as

$$P(L(X) \leq Y \leq U(X)) = 1 - \alpha, \tag{1}$$

then $[L(X), U(X)]$ is called a level 1-α prediction interval of Y.

2.2 Existing Methods

We present several existing prediction intervals for a single binomial random variable as follows:

(1). The Nelson prediction interval

Assume that the past data, X, follows a Binomial(n, p) distribution where $0 < p <$

1. Let Y be the future number following a Binomial(m, p) distribution. The Nelson prediction interval to predict Y is derived from

$$\frac{Y - m\hat{p}}{\sqrt{\hat{p}(1-\hat{p})m(m+n)/n}} \sim N(0,1), \tag{2}$$

where $\hat{p} = X/n$. Nelson (1982) [6] proposed an approximate level 1-α two-sided prediction interval which is

$$(\hat{Y} - z_{\alpha/2}\sqrt{m\hat{p}(1-\hat{p})(m+n)/n}\ ,\ \hat{Y} + z_{\alpha/2}\sqrt{m\hat{p}(1-\hat{p})(m+n)/n}\ ) \tag{3}$$

where z_α is the upper α cut off point of the standard normal distribution and $\hat{Y} = m\hat{p}$ when X, $n - X$, Y, and $m - y$ all of these are large.

(2). Bain and Patel prediction interval

Bain and Patel approximate prediction interval (1993) is based on the conditional distribution to exclude the unknown parameter and uses the conditional distribution to derive the predictive boundaries. The approximate level 1-α has the form

$$(L_X - X\ ,\ U_X - X\ ), \tag{4}$$

where

$$L_X = \frac{((2X-1)v + sw) - \sqrt{s^2w^2 + 4(X-1/2)w(n-X+1/2)}}{2(v^2+w)},$$
$$U_X = \frac{((2X+1)v + sw) - \sqrt{s^2w^2 + 4(X+1/2)w(n-X-1/2)}}{2(v^2+w)},$$

$s = m + n$, $v = n/s$, and $w = z^2_{(1-\alpha/2)}v(1-v)/(s-1)$.

In terms of these existing prediction intervals, Wang (2008) constructs an adjustable prediction interval and proposed procedures to calculate the minimum coverage probability and average coverage probability of a binomial prediction interval.

(3). Wang's prediction interval

Based on a similar argument as Wilson confidence interval construction (Wilson 1927), Wang (2010) used the fact that the random variable

$$\frac{Y - m\hat{p}}{\sqrt{\frac{X+Y}{n+m}(1-\frac{X+Y}{n+m})\frac{m(m+n)}{n}}}, \tag{5}$$

is approximating a standard normal distribution to construct a prediction interval. In the existing prediction interval, Wilson interval, there is an disadvantage that when the true parameter, p, is closed to the boundaries, the coverage probability is much less than the nominal level. Let $k = z^2_{1-\alpha/2}$, in order to prevent poor coverage provability, Wang (2010) inverts

$$\{y : y = m\hat{p} \pm \sqrt{kW(x,y)}\} \tag{6}$$

to derive the prediction limits, where

$$W(x,y) = \frac{(x+k/2+y)}{(n+k+m)} \times \left(1 - \frac{(x+k/2+y)}{(n+t^2+m)}\right) \times \left(\frac{m+n}{n}\right) \tag{7}$$

Based on this outcome, Wang (2010) proposed the following prediction interval

$$(\frac{A}{C} - \frac{B}{C}, \frac{A}{C} + \frac{B}{C}) \tag{8}$$

where

$$A = mn[2xk(n+k+m) + (2x+k)(m+n)^2],$$

$$B = (mn(m+n)k(m+n+k)^2 \times (2(n-x)[n^2(2x+k) + 4mnx + 2m^2x]$$

$$+nk[n(2x+k) + 3mn + m^2]))^{\frac{1}{2}},$$

$$C = 2n[(n+k)(m^2 + n(n+k)) + mn(2n+3k)]$$

and

$$k = z^2_{(1+\alpha)/2}.$$

The performance of the coverage probability of (5) is referred to Figure 1 in Wang(2010).

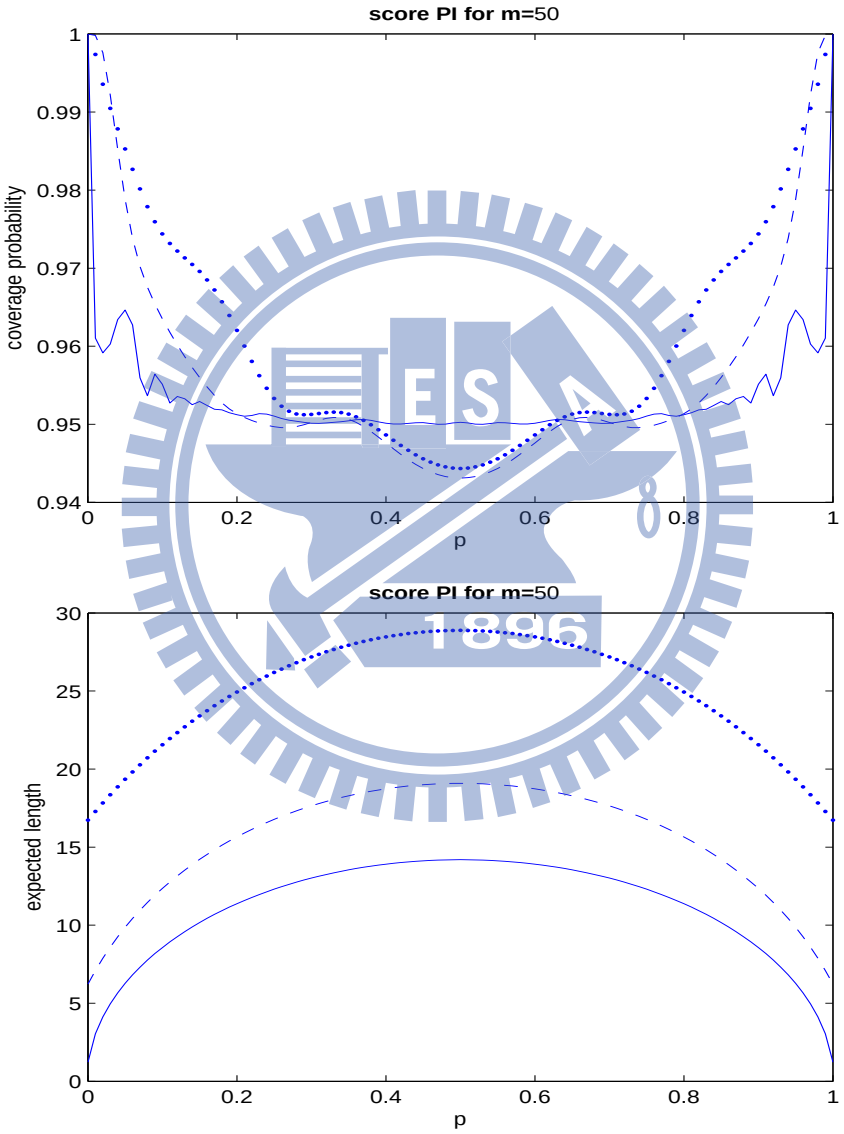

Figure 1: Coverage probability of the 95% level prediction intervals for the Binomial distributions with $n = 10$(dotted line), $n = 50$(dashed line) and $n = 1000$(solid line).

Furthermore, to prevent the poor coverage probability when the parameter, p, is closed to the boundaries, Wang (2010) improves the existing prediction interval (3) by substituting $\tilde{p}$ for $\hat{p}$, which contributes to the second proposed interval

$$[\, L_a(X)\ ,\ U_a(X)\,] = [\, \hat{Y} - z_{(1-\alpha/2)}\sqrt{m\tilde{p}(1-\tilde{p})(m+n)/n}\ ,\ \hat{Y} + z_{(-\alpha/2)}\sqrt{m\tilde{p}(1-\tilde{p})(m+n)/n}\,], \tag{9}$$

where $\tilde{p} = (X + (z^2_{(1+\alpha)/2})/2)/(n + z^2_{(1+\alpha)/2})$.

As mentioned in the introduction, in our study we principally focus on the prediction interval for a linear function of multiple variables.

3 METHOD

3.1 The prediction interval for a linear function of two variables without correlation

Let $\boldsymbol{X}_1$=$(X_{11}, X_{12}, \ldots, X_{1n_1})$ be an observed random sample of size n_1 from a discrete distribution with a probability mass function $f_{n_1}(x_1; p_1)$, and let $\boldsymbol{X}_2$=$(X_{21}$, X_{22}, $\ldots, X_{2n_2})$ be an observed random sample of size n_2 from a discrete distribution with a probability mass function $f_{n_2}(x_2; p_2)$, where p_1 and p_2 are unknown parameters. Let $Y_{11}, \ldots, Y_{1m_1}$ be a future random sample of size m_1 from the distribution, $f_{m_1}(y_1; p_1)$ and let $Y_{21}, \ldots, Y_{2m_2}$ be a future random sample of size m_2 from the distribution, $f_{m_2}(y_2; p_2)$.

Assume that the future sample $Y_{11}, \ldots, Y_{1m_1}$ and $Y_{21}, \ldots, Y_{2m_2}$ are drawn independently of the past sample $X_{11}, X_{12}, \ldots, X_{1n_1}$ and $X_{21}, X_{22}, \ldots, X_{2n_2}$. Let $L(X_1)$ and $U(X_1)$ be two statistics based on the observed sample, $X_{11}, X_{12}, \ldots, X_{1n_1}$ and let $L(X_2)$ and $U(X_2)$ be two statistics based on the observed sample, $X_{21}, X_{22}, \ldots, X_{2n_2}$. If $L(\boldsymbol{X}_1)$, $L(\boldsymbol{X}_2)$, $U(\boldsymbol{X}_1)$ and $U(\boldsymbol{X}_2)$ are determined so that

$$P(L(\boldsymbol{X}_1) + L(\boldsymbol{X}_2) \leq Y_1 + Y_2 \leq U(\boldsymbol{X}_1) + U(\boldsymbol{X}_2)) = 1 - \alpha, \tag{10}$$

where $L(\boldsymbol{X}_1)$, $U(\boldsymbol{X}_1)$, $L(\boldsymbol{X}_2)$ and $U(\boldsymbol{X}_2)$ are determined by (8), then

$$[L(\boldsymbol{X}_1) + L(\boldsymbol{X}_2), U(\boldsymbol{X}_1) + U(\boldsymbol{X}_2)], \tag{11}$$

is called a level 1-α prediction interval of $Y_1 + Y_2$.

In our study, we find a lower bound of the coverage probability of the prediction

interval (11). At first, let $Y_1 \in [L(\boldsymbol{X}_1), U(\boldsymbol{X}_1)]$ and $Y_2 \in [L(\boldsymbol{X}_2), U(\boldsymbol{X}_2)]$, where $L(\boldsymbol{X}_1)$, $U(\boldsymbol{X}_1)$, $L(\boldsymbol{X}_2)$, and $U(\boldsymbol{X}_2)$ are determined by (8). Using (8), we have

$$P(L(\boldsymbol{X}_1) \leq Y_1 \leq U(\boldsymbol{X}_1)) = 1 - \alpha$$

and

$$P(L(\boldsymbol{X}_2) \leq Y_2 \leq U(\boldsymbol{X}_2)) = 1 - \alpha.$$

We consider the case that the future observations, Y_1 will belong to $[L(\boldsymbol{X}_1), U(\boldsymbol{X}_1)]$ and Y_2 will belong to $[L(\boldsymbol{X}_2), U(\boldsymbol{X}_2)]$. We find a lower bound of coverage probability which is

$$P(L(\boldsymbol{X}_1) \leq Y_1 \leq U(\boldsymbol{X}_1)) \times P(L(\boldsymbol{X}_2) \leq Y_2 \leq U(\boldsymbol{X}_2)) = (1-\alpha)^2.$$

It is due to

$$P(L(\boldsymbol{X}_1) + L(\boldsymbol{X}_2) \leq Y_1 + Y_2 \leq U(\boldsymbol{X}_1) + U(\boldsymbol{X}_2))$$

$$\geq P(L(\boldsymbol{X}_1) \leq Y_1 \leq U(\boldsymbol{X}_1)) \times P(L(\boldsymbol{X}_2) \leq Y_2 \leq U(\boldsymbol{X}_2)).$$

Here, we can ensure that $[L(\boldsymbol{X}_1) + L(\boldsymbol{X}_2), U(\boldsymbol{X}_1) + U(\boldsymbol{X}_2)]$ has a lower approximate coverage probability $(1-\alpha)^2$.

3.2 The prediction interval for a linear function of multiple variables without correlation

Let $\boldsymbol{X}_i = (X_{i1}, X_{i2}, \ldots, X_{in_i})$ be an observed random sample of size n_i from a discrete distribution with a probability mass function $f_{n_i}(x_i; p_i)$, i=1, …, k, where p_i

is unknown parameters. Let $Y_{i1}, \ldots, Y_{im_i}$ be a future random sample of size m_i from the same distribution, $f_{m_i}(y_i; p_i)$.

Assume that the future sample $Y_{i1}, \ldots, Y_{im_i}$ are drawn independently of the past sample $\boldsymbol{X}_{i1}, \boldsymbol{X}_{i2}, \ldots, \boldsymbol{X}_{in_i}$, i=1, …, k. When $i = 3$, let $L(\boldsymbol{X}_1)$ and $U(\boldsymbol{X}_1)$ be two statistics based on the observed sample, $X_{11}, X_{12}, \ldots, X_{1n_1}$, let $L(\boldsymbol{X}_2)$ and $U(\boldsymbol{X}_2)$ be two statistics based on the observed sample, $X_{21}, X_{22}, \ldots, X_{2n_2}$ and let $L(\boldsymbol{X}_3)$ and $U(\boldsymbol{X}_3)$ be two statistics based on the observed sample, $X_{31}, X_{32}, \ldots, X_{3n_3}$. If $L(\boldsymbol{X}_1)$, $L(\boldsymbol{X}_2)$, $L(\boldsymbol{X}_3)$, $U(\boldsymbol{X}_1)$, $U(\boldsymbol{X}_2)$ and $U(\boldsymbol{X}_3)$ are determined so that

$$P(L(\boldsymbol{X}_1)+L(\boldsymbol{X}_2)+L(\boldsymbol{X}_3) \le Y_1+Y_2+Y_3 \le U(\boldsymbol{X}_1)+U(\boldsymbol{X}_2)+U(\boldsymbol{X}_3)) = 1-\alpha, \quad (12)$$

where $L(\boldsymbol{X}_1)$, $U(\boldsymbol{X}_1)$, $L(\boldsymbol{X}_2)$, $U(\boldsymbol{X}_2)$, $L(\boldsymbol{X}_3)$, and $U(\boldsymbol{X}_3)$ are determined by (8) then

$$[L(\boldsymbol{X}_1) + L(\boldsymbol{X}_2) + L(\boldsymbol{X}_3), U(\boldsymbol{X}_1) + U(\boldsymbol{X}_2) + U(\boldsymbol{X}_3)], \quad (13)$$

is called a level 1-α prediction interval of $Y_1 + Y_2 + Y_3$.

In our study, we find a lower bound of the coverage probability of the prediction interval of (13). At first, let $Y_1 \in [L(\boldsymbol{X}_1), U(\boldsymbol{X}_1)]$, $Y_2 \in [L(\boldsymbol{X}_2), U(\boldsymbol{X}_2)]$ and $Y_3 \in [L(\boldsymbol{X}_3), U(\boldsymbol{X}_3)]$ where $L(\boldsymbol{X}_1)$, $U(\boldsymbol{X}_1)$, $L(\boldsymbol{X}_2)$, $U(\boldsymbol{X}_2)$, $L(\boldsymbol{X}_3)$ and $U(\boldsymbol{X}_3)$ are determined by (8). Using (8), we have

$$P(L(\boldsymbol{X}_1) \le Y_1 \le U(\boldsymbol{X}_1)) = 1 - \alpha,$$

$$P(L(\boldsymbol{X}_2) \le Y_2 \le U(\boldsymbol{X}_2)) = 1 - \alpha$$

and

$$P(L(\boldsymbol{X}_3) \leq Y_3 \leq U(\boldsymbol{X}_3)) = 1 - \alpha.$$

We consider the case that the future observations, Y_1 will belong to $[L(\boldsymbol{X}_1), U(\boldsymbol{X}_1)]$, Y_2 will belong to $[L(\boldsymbol{X}_2), U(\boldsymbol{X}_2)]$ and Y_3 will belong in $[L(\boldsymbol{X}_3), U(\boldsymbol{X}_3)]$. We find a lower bound of coverage probability which is

$$P(L(\boldsymbol{X}_1) \leq Y_1 \leq U(\boldsymbol{X}_1)) \times P(L(\boldsymbol{X}_2) \leq Y_2 \leq U(\boldsymbol{X}_2)) \times P(L(\boldsymbol{X}_3) \leq Y_3 \leq U(\boldsymbol{X}_3)) = (1-\alpha)^3.$$

It is due to

$$P(L(\boldsymbol{X}_1) + L(\boldsymbol{X}_2) + L(\boldsymbol{X}_3) \leq Y_1 + Y_2 + Y_3 \leq U(\boldsymbol{X}_1) + U(\boldsymbol{X}_2) + U(\boldsymbol{X}_3))$$

$$\geq P(L(\boldsymbol{X}_1) \leq Y_1 \leq U(\boldsymbol{X}_1)) \times P(L(\boldsymbol{X}_2) \leq Y_2 \leq U(\boldsymbol{X}_2)) \times P(L(\boldsymbol{X}_3) \leq Y_3 \leq U(\boldsymbol{X}_3)).$$

Here, we can ensure that $[L(\boldsymbol{X}_1) + L(\boldsymbol{X}_2) + L(\boldsymbol{X}_3), U(\boldsymbol{X}_1) + U(\boldsymbol{X}_2) + U(\boldsymbol{X}_3)]$ has a lower approximate coverage probability $(1-\alpha)^3$.

3.3 The prediction interval for a linear function of two variables with correlation

Let $\boldsymbol{X}_1$=$(X_{11}, X_{12}, \ldots, X_{1n_1})$ be an observed random sample of size n_1 from a discrete distribution with a probability mass function $f_{n_1}(x_1; p_1)$, and let $\boldsymbol{X}_2$=$(X_{21}, X_{22}, \ldots, X_{2n_2})$ be an observed random sample of size n_2 from a discrete distribution with a probability mass function $f_{n_2}(x_2; p_2)$, where p_1 and p_2 are unknown parameters. Let $Y_{11}, \ldots, Y_{1m_1}$ be a future random sample of size m_1 from the same distribution, $f_{m_1}(y_1; p_1)$ and let $Y_{21}, \ldots, Y_{2m_2}$ be a future random sample of size m_2 from the same

distribution, $f_{m_2}(y_2;p_2)$. Let $\boldsymbol{X}_3$=$(X_{31}, X_{32}, \ldots, X_{3n_3})$ be an observed random sample of size n_3 from a discrete distribution with probability mass function $f_{n_3}(x_3;p_3)$, where p_3 is a function of p_1 and p_2. The variable $\boldsymbol{X}_3$ is dependent on variables $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$. Let $Y_{31}, \ldots, Y_{3m_3}$ be a future random sample of size m_3 from the distribution, $f_{m_3}(y_3;p_3)$.

Assume that the future sample $Y_{11}, \ldots, Y_{1m_1}$, $Y_{21}, \ldots, Y_{2m_2}$ and $Y_{31}, \ldots, Y_{3m_3}$ are drawn independently of the past sample $X_{11}, X_{12}, \ldots, X_{1n_1}$, X_{21}, X_{22} , $\ldots, X_{2n_2}$ and X_{31}, X_{32} , $\ldots, X_{3n_2}$. Let $\boldsymbol{X}_1$ be a function of $X_{11}, X_{12}, \ldots, X_{1n_1}$, $\boldsymbol{X}_2$ be a function of $X_{21}, X_{22}, \ldots, X_{2n_2}$ and $\boldsymbol{X}_3$ be a function of $X_{31}, X_{32}, \ldots, X_{3n_3}$, with a probability mass function $f_{n_1}(x_1;p_1)$, $f_{n_2}(x_2;p_2)$ and $f_{n_3}(x_3;p_3)$, respectively. Let Y_1 be a function of $Y_{11}, \ldots, Y_{1m_1}$, Y_2 be a function of $Y_{21}, \ldots, Y_{2m_2}$ with probability mass functions $f_{m_1}(y_1;p_1)$ and $f_{m_2}(y_2;p_2)$, and Y_3 be a function of $Y_{31}, \ldots, Y_{3m_3}$ with a probability mass function $f_{m_3}(y_3;p_3)$, respectively. Let $L(\boldsymbol{X}_1)$ and $U(\boldsymbol{X}_1)$ be two statistics based on the observed sample, $X_{11}, X_{12}, \ldots, X_{1n_1}$ and let $L(\boldsymbol{X}_2)$ and $U(\boldsymbol{X}_2)$ be two statistics based on the observed sample, $X_{21}, X_{22}, \ldots, X_{2n_2}$. If there are some sample recounted between the observed samples, $X_{11}, X_{12}, \ldots, X_{1n_1}$ and $X_{21}, X_{22}, \ldots, X_{2n_2}$, let the recounted item be the $X_{31}, X_{32}, \ldots, X_{3n_3}$. The $L(\boldsymbol{X}_1)$, $L(\boldsymbol{X}_2)$, $L(\boldsymbol{X}_3)$, $U(\boldsymbol{X}_1)$, $U(\boldsymbol{X}_2)$ and $U(\boldsymbol{X}_3)$ are determined under the condition that $\boldsymbol{X}_3$ is dependent on $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ such that

$$P(L(\boldsymbol{X}_1)+L(\boldsymbol{X}_2)-U(\boldsymbol{X}_3) \leq Y_1+Y_2-Y_3 \leq U(\boldsymbol{X}_1)+U(\boldsymbol{X}_2)-L(\boldsymbol{X}_3)) = 1-\alpha. \quad (14)$$

Then

$$[L(\boldsymbol{X}_1) + L(\boldsymbol{X}_2) - U(\boldsymbol{X}_3), U(\boldsymbol{X}_1) + U(\boldsymbol{X}_2) - L(\boldsymbol{X}_3)], \tag{15}$$

is called a level 1-α prediction interval of $Y_1 + Y_2 - Y_3$. By taking the correlation into consideration, the prediction interval can be applied to more realistic situations.

4 SIMULATION

In this section, we conduct a simulation study to evaluate the performance of the proposed prediction intervals. First, we present the simulation for two variables without correlation case. Second, we consider multiple variables without correlation case. Finally, the simulation is presented for the two variables with correlation case.

4.1 The simulation in two variables without correlation

Using the prediction interval (11), we consider two points of view. First, we observe the coverage probability corresponding to different α when p_1 or p_2 be fixed. Second, we observe the coverage probability corresponding to different p_1 and p_2 when α be fixed.

(1). Coverage probability corresponding to different p_1 and p_2 at $\alpha = 0.05$

Here, we present some cases about the coverage probability corresponding to different p_1 and p_2 when α=0.05. In this paragraph, we consider three cases about the sample size, which are $n_1 = n_2$, $n_1 = 2n_2$ and $n_1 = 3n_2$ when $\alpha = 0.05$.

Figure 2, 3 and 4 show that the relationship between coverage probability and different (p_1, p_2) in different cases of the sample size. In Figure 2, 3 and 4, we observe that the maximum coverage probability occurs at the four top of corners, and the minimum coverage probability occurs at the center of figure.

- **The case for X_1 and X_2 with the same sample size.**

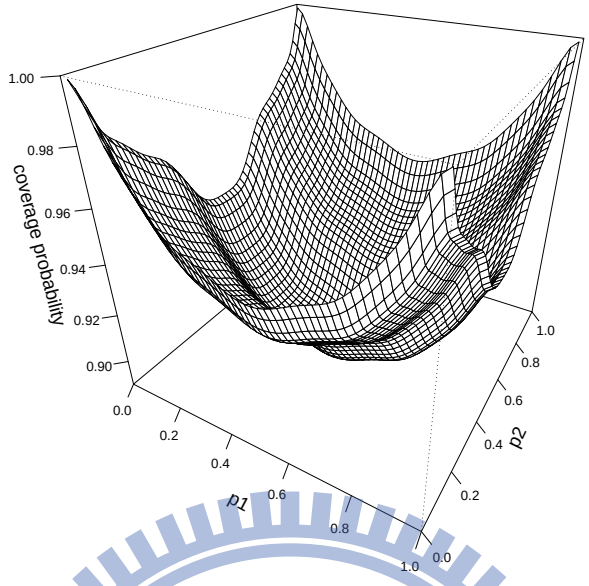

Figure 2: Coverage probability of the prediction interval corresponding to (n_1, n_2, m_1, m_2)= (30,30,10,5) with the maximum value=0.9989452 occurring at (p_1, p_2) = (0.01,0.01), (0.01,0.99), (0.99,0.01), (0.99,0.99) and the minimum value=0.8893743 occurring at (p_1, p_2) = (0.49,0.49), (0.49,0.51), (0.51,0.49), (0.51,0.51).

- **The case for X_1 and X_2 with the different sample size when $n_1 = 2n_2$.**

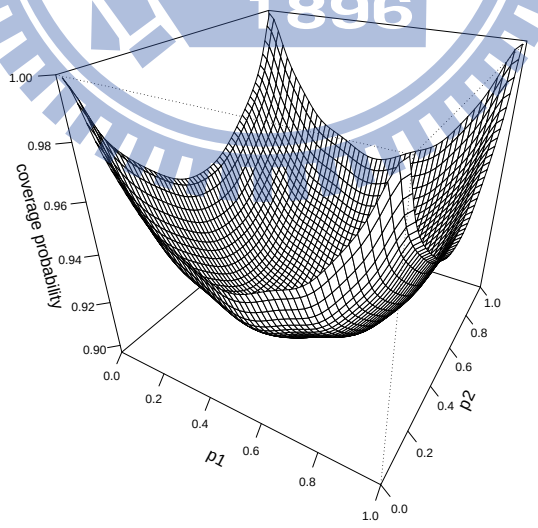

Figure 3: Coverage probability of the prediction interval corresponding to (n_1, n_2, m_1, m_2)= (30,15,10,5) with the maximum value=0.9990451 occurring at (p_1, p_2) = (0.01,0.01), (0.01,0.99), (0.99,0.01), (0.99,0.99) and the minimum value=0.8957184 occurring at (p_1, p_2) = (0.49,0.49), (0.49,0.51), (0.51,0.49), (0.51,0.51).

- **The case for X_1 and X_2 with the different sample size when $n_1 = 3n_2$.**

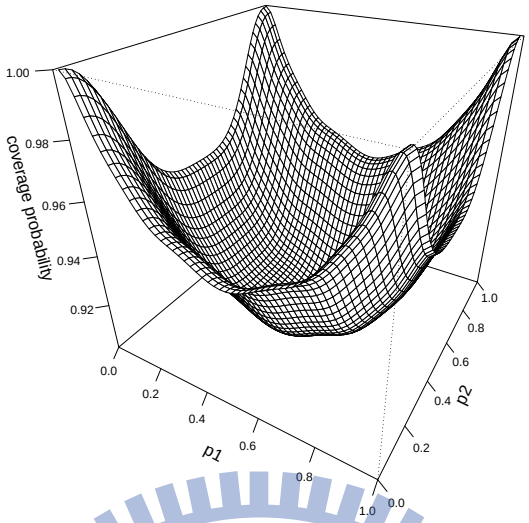

Figure 4: Coverage probability of the prediction interval corresponding to (n_1, n_2, m_1, m_2)= (30,10,10,5) with the maximum value=0.9998804 occurring at (p_1, p_2) = (0.01,0.01), (0.01,0.99), (0.99,0.01), (0.99,0.99) and the minimum value=0.9010947 occurring at (p_1, p_2) = (0.49,0.49), (0.49,0.51), (0.51,0.49), (0.51,0.51).

(2). The α trend

Figure 5 shows the coverage probability corresponding to different α and p_2 when p_1=0.9. Figure 6 shows the coverage probability corresponding to different α and p_1 when p_2=0.9. According to Figure 5 and Figure 6, we observe a trend that the coverage probability decreases when α increases.

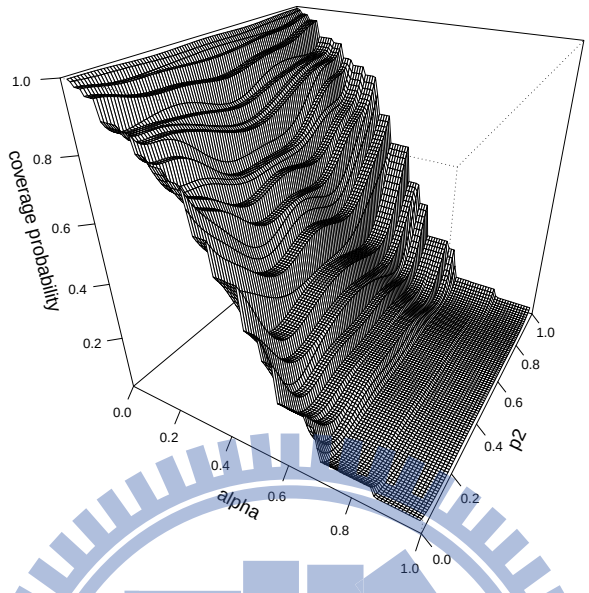

Figure 5: The coverage probability corresponding to different α and p_2 when p_1=0.9.

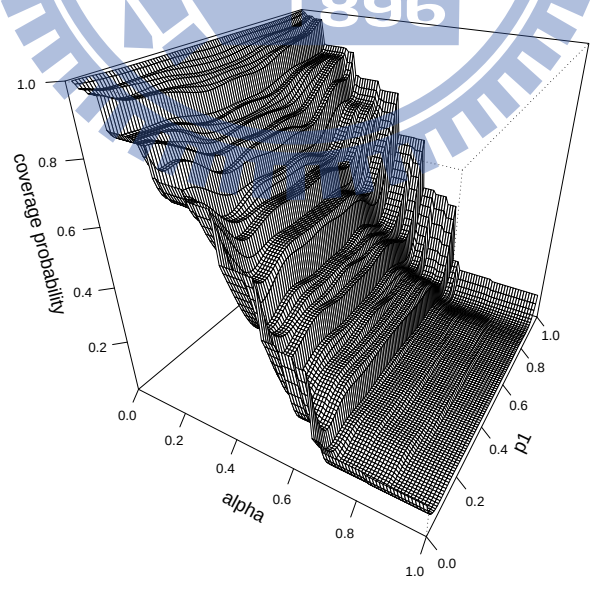

Figure 6: The coverage probability corresponding to different α and p_1 when p_2=0.9.

4.2 The simulation in three variables without correlation

Using the prediction interval (12), we consider two points of view. First, we observe the coverage probability corresponding to different α when p_1 or p_2 or p_3 be fixed. Second, we observe the coverage probability corresponding to different p_1 and p_2 when α and p_3 be fixed, or p_1 and p_3 when α and p_2 be fixed, or p_2 and p_3 when α and p_1 be fixed.

(1). Coverage probability corresponding to different p_1, p_2 and p_3 at α=0.05

- **The case for X_1, X_2 and X_3 with the same sample size.**

Figure 7 shows that the relationship between coverage probability and different (p_1, p_2) under fixed p_3 and equal sample size. Figure 8 shows that the relationship between coverage probability and different (p_2, p_3) under fixed p_1 and equal sample size. Figure 9 shows that the relationship between coverage probability and different (p_1, p_3) under fixed p_2 and equal sample size.

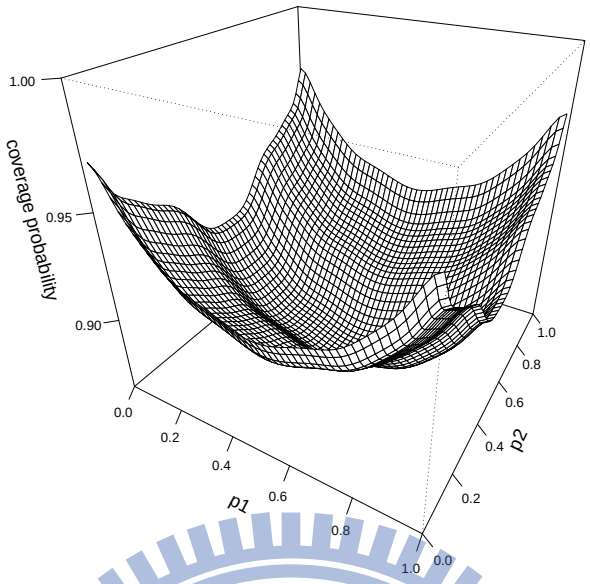

Figure 7: Coverage probability of the prediction interval corresponding to (n_1, n_2, n_3, m_1, m_2, m_3)= (30,30,30,10,5,3) with the maximum value=0.969829 occurring at (p_1, p_2, p_3) = (0.01,0.01,0.75), (0.01,0.99,0.75), (0.99,0.01,0.75), (0.99,0.99,0.75) and the minimum value=0.8634517 occurring at (p_1, p_2, p_3) = (0.49,0.49,0.75), (0.49,0.51,0.75), (0.49,0.51,0.75), (0.51,0.51,0.75).

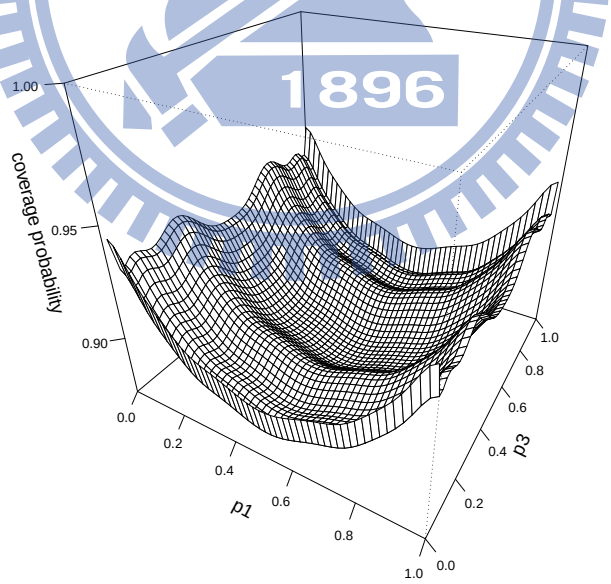

Figure 8: Coverage probability of the prediction interval corresponding to (n_1, n_2, n_3, m_1, m_2, m_3)= (30,30,30,10,5,3) with the maximum value=0.9442639 occurring at (p_1, p_2, p_3) = (0.01,0.75,0.01), (0.01,0.75,0.99), (0.99,0.75,0.01), (0.99,0.75,0.99) and the minimum value=0.8721178 occurring at (p_1, p_2, p_3) = (0.49,0.75,0.49), (0.49,0.75,0.51), (0.51,0.75,0.49, (0.51,0.75,0.51).

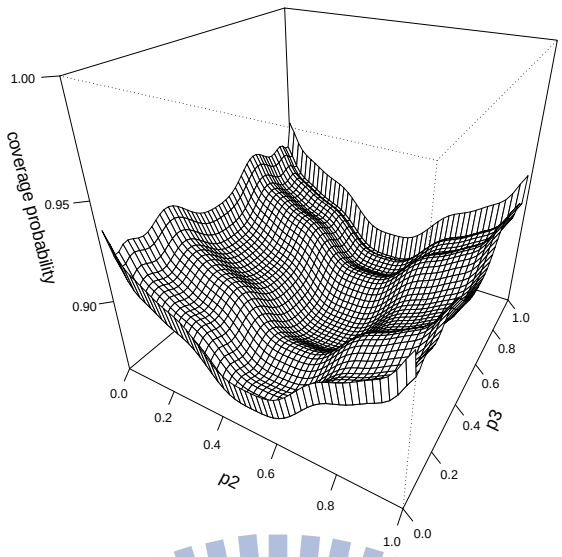

Figure 9: Coverage probability of the prediction interval corresponding to (n_1, n_2, n_3, m_1, m_2, m_3)= (30,30,30,10,5,3) with the maximum value=0.9360158 occurring at (p_1, p_2, p_3) = (0.75,0.01,0.01), (0.75,0.01,0.99), (0.75,0.99,0.01), (0.75,0.99,0.99) and the minimum value=0.8600604 occurring at (p_1, p_2, p_3) = (0.75,0.49,0.49), (0.75,0.49,0.51), (0.75,0.51,0.49), (0.75,0.51,0.51).

- **The case for X_1, X_2 and X_3 with the different sample size when $n_1 = 3n_2 = 2n_3$.**

Figure 10 shows that the relationship between coverage probability and different (p_1, p_2) under fixed p_3 and equal sample size. Figure 11 shows that the relationship between coverage probability and different (p_2, p_3) under fixed p_1 and equal sample size. Figure 12 shows that the relationship between coverage probability and different (p_1, p_3) under fixed p_2 and different sample size while $n_1 = 3n_2 = 2n_3$.

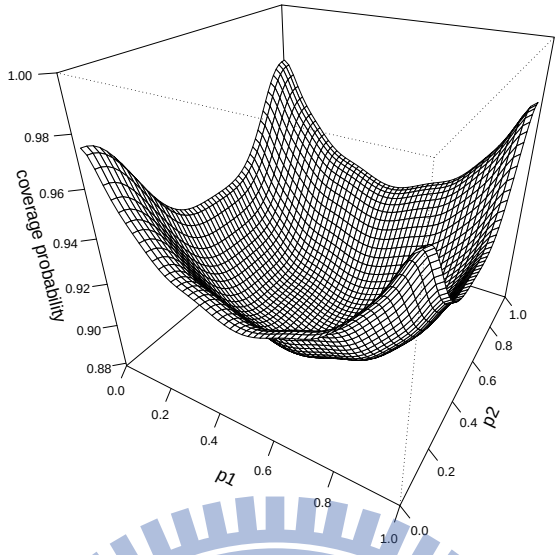

Figure 10: Coverage probability of the prediction interval corresponding to (n_1, n_2, n_3, m_1, m_2, m_3)= (30,10,15,10,5,3) with the maximum value=0.9749658 occurring at (p_1, p_2, p_3) = (0.01,0.01,0.75), (0.01,0.99,0.75), (0.99,0.01,0.75), (0.99,0.99,0.75) and the minimum value=0.8786416 occurring at (p_1, p_2, p_3) = (0.49,0.49,0.75), (0.49,0.51,0.75), (0.51,0.49,0.75), (0.51,0.51,0.75).

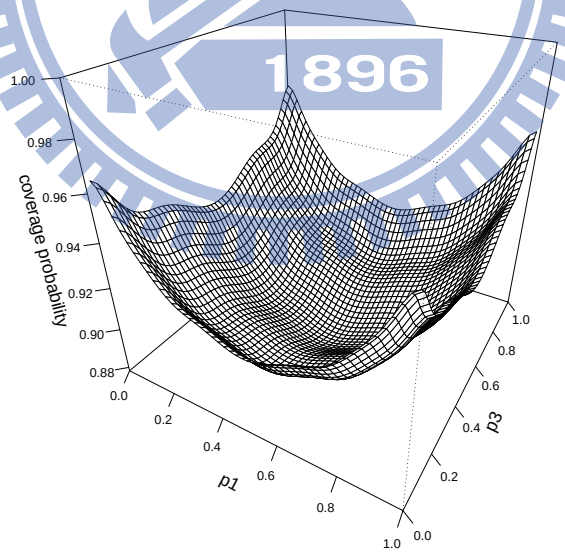

Figure 11: Coverage probability of the prediction interval corresponding to (n_1, n_2, n_3, m_1, m_2, m_3)= (30,10,15,10,5,3) with the maximum value=0.9647086 occurring at (p_1, p_2, p_3) = (0.01,0.75,0.01), (0.01,0.75,0.99), (0.99,0.75,0.01), (0.99,0.75,0.99) and the minimum value=0.878146 occurring at (p_1, p_2, p_3) = (0.49,0.75,0.49), (0.49,0.75,0.51), (0.51,0.75,0.49), (0.51,0.75,0.51).

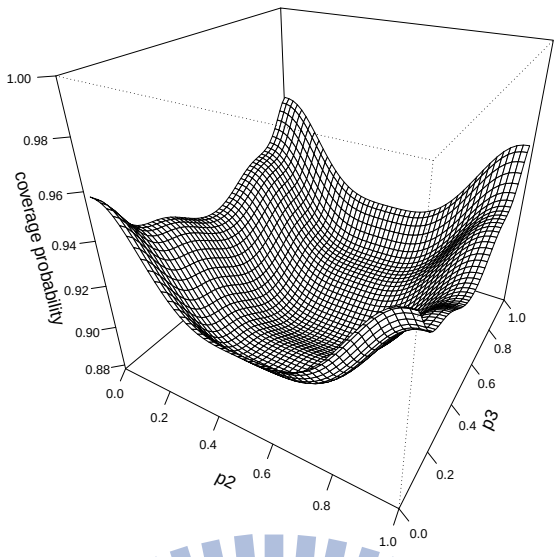

Figure 12: Coverage probability of the prediction interval corresponding to (n_1, n_2, n_3, m_1, m_2, m_3)= (30,10,15,10,5,3) with the maximum value =0.9577311 occurring at (p_1, p_2, p_3) = (0.75,0.01,0.01), (0.75,0.01,0.99), (0.75,0.99,0.01), (0.75,0.99,0.99) and the minimum value =0.8779254 occurring at (p_1, p_2, p_3) = (0.75,0.49,0.49), (0.75,0.49,0.51), (0.75,0.51,0.49), (0.75,0.51,0.51).

(2). The α trend

Figure 13 shows the coverage probability corresponding to different α and p_1 when p_2=0.9 and p_3=0.9. Figure 14 shows the coverage probability corresponding to different α and p_3 when p_1=0.9 and p_2=0.9. According to Figure 13 and 14, it shows a trend that the coverage probability increases by α decreasing. Both Figure 13 and 14 show that the coverage probability starts to closed to zero while α approaches to 0.6.

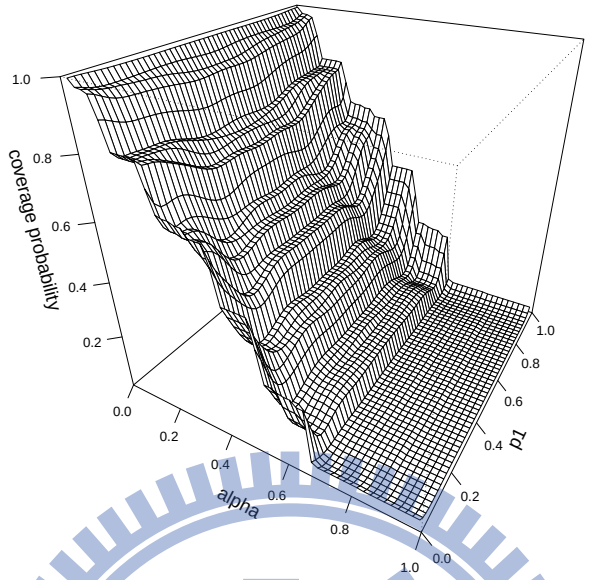

Figure 13: The coverage probability corresponding to different α and p_1 when p_2=0.9 and $p_3 = 0.9$.

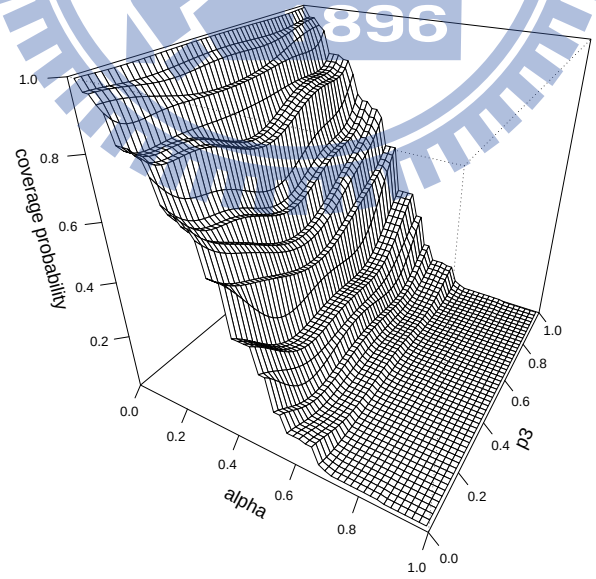

Figure 14: The coverage probability corresponding to different α and p_3 when p_1=0.9 and $p_2 = 0.9$.

4.3 The simulation in two variables with correlation

Using the prediction interval (14), we consider two points of view. First, we observe the coverage probability corresponding to different α when p_1 or p_2 be fixed with $p_3 = p_1 \times p_2$ and $p_3 = p_1^2 \times p_2$. Second, we observe the coverage probability corresponding to different p_1 and p_2 when α be fixed.

(1). Coverage probability corresponding to different p_1 and p_2 at $\alpha = 0.05$

Here, we present some cases about the coverage probability corresponding to different α and the coverage probability corresponding to different p_1 and p_2 when $\alpha = 0.05$.

• Coverage probability corresponding to different p_1 and p_2 under $p_3 = p_1 \times p_2$ with same sample size

Figure 15 shows that the relationship between coverage probability and different (p_1, p_2) under $p_3 = p_1 \times p_2$ and the equal sample size. Figure 16 shows that the relationship between coverage probability and different p_3 under the equal sample size.

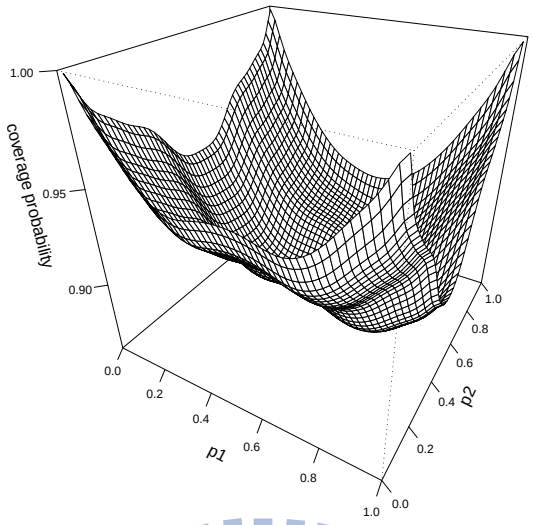

Figure 15: The coverage probability corresponding to different p_1 and p_2 under $\alpha = 0.05$ and $p_3 = p_1 \times p_2$. Coverage probability of the prediction interval corresponding to (n_1, n_2, n_3, m_1, m_2, m_3)= (30,30,10,15,5,3) with the maximum value =0.9987448 occurring at (p_1, p_2) = (0.01,0.01) and the minimum value =0.860743 occurring at (p_1, p_2) = (0.71,0.57) while $p_3 = p_1 \times p_2$.

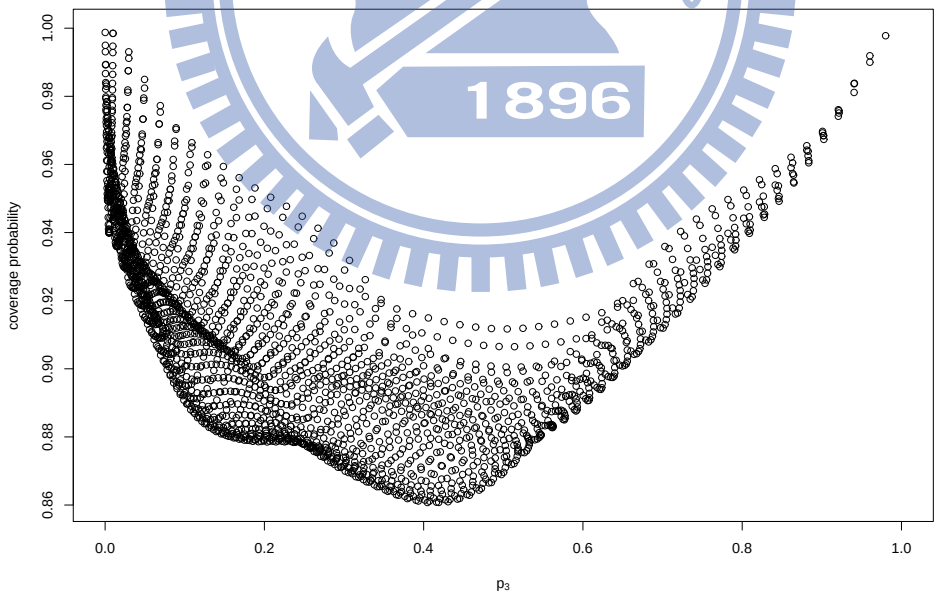

Figure 16: The coverage probability corresponding to different p_1 and p_2 under $\alpha = 0.05$ and $p_3 = p_1 \times p_2$. Coverage probability of the prediction interval corresponding to (n_1, n_2, n_3, m_1, m_2, m_3)= (30,30,10,15,5,3) with the maximum value =0.9987448 occurring at $p_3 = 0.0001$ where (p_1, p_2) = (0.01,0.01) and the minimum value =0.860743 occurring at $p_3 = 0.4047$ while (p_1, p_2) = (0.71,0.57).

• **Coverage probability corresponding to different p_1 and p_2 under $p_3 = p_1 \times p_2$ with different sample size ($n_1 = 2n_2$)**

Figure 17 shows that the relationship between coverage probability and different (p_1, p_2) under $p_3 = p_1 \times p_2$ and different sample size while $n_1 = 2n_2$. Figure 18 shows that the relationship between coverage probability and different p_3 under different sample size while $n_1 = 2n_2$.

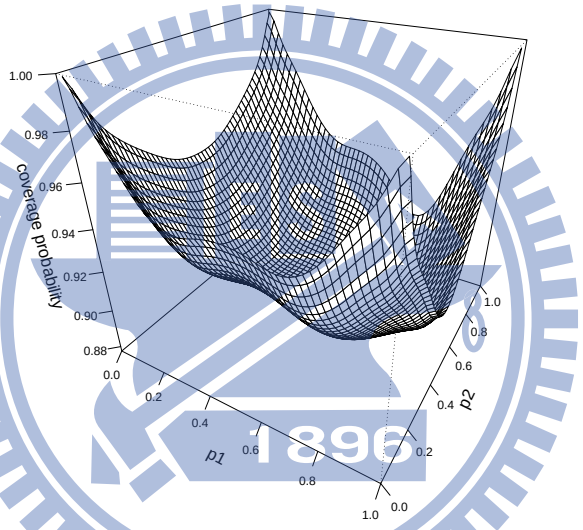

Figure 17: The coverage probability corresponding to different p_1 and p_2 under $\alpha = 0.05$ and $p_3 = p_1 \times p_2$. Coverage probability of the prediction interval corresponding to (n_1, n_2, n_3, m_1, m_2, m_3)= (30,15,10,15,5,3) with the maximum value = 0.9988455 occurring at (p_1, p_2) = (0.01,0.01) and the minimum value = 0.8665772 occurring at (p_1, p_2) = (0.71,0.61) while $p_3 = p_1 \times p_2$.

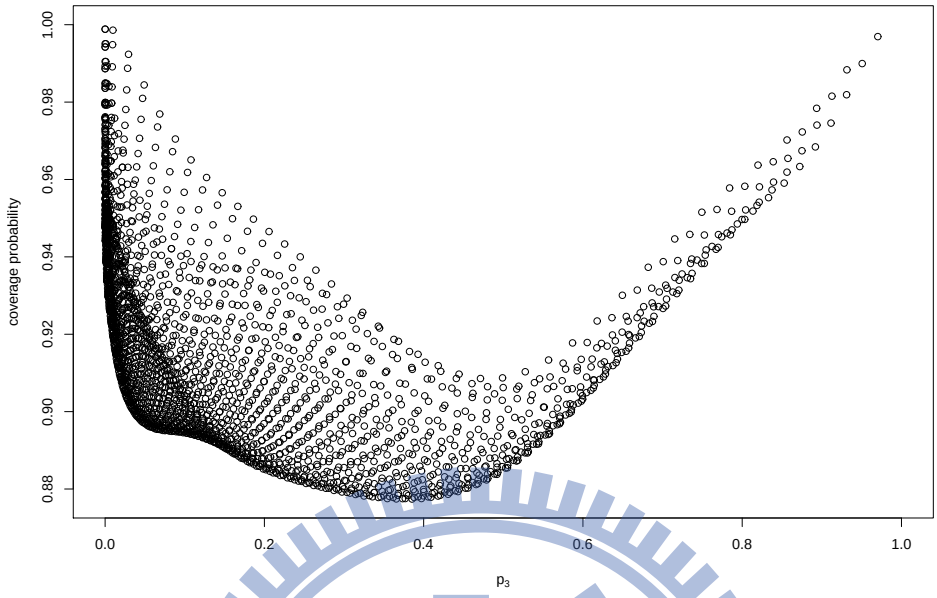

Figure 18: The coverage probability corresponding to different p_1 and p_2 under $\alpha = 0.05$ and $p_3 = p_1 \times p_2$. Coverage probability of the prediction interval corresponding to (n_1, n_2, n_3, m_1, m_2, m_3)= (30,15,10,15,5,3) with the maximum value = 0.9988455 occurring at $p_3 = 0.0001$ where $(p_1, p_2) = (0.01,0.01)$ and the minimum value = 0.8665772 occurring at $p_3 = 0.4331$ while $(p_1, p_2) = (0.71,0.61)$.

• Coverage probability corresponding to different p_1, p_2 and with correlation $p_3 = p_1^2 \times p_2$

Here, we present some cases about the coverage probability corresponding to different α, the coverage probability corresponding to different p_1 and p_2 when $\alpha = 0.05$.

• Coverage probability corresponding to different p_1 and p_2 under $p_3 = p_1^2 \times p_2$ with same sample size

Figure 19 shows that the relationship between coverage probability and different (p_1, p_2) under $p_3 = p_1^2 \times p_2$ and the equal sample size. Figure 20 shows that the relationship between coverage probability and different p_3 with equal sample size.

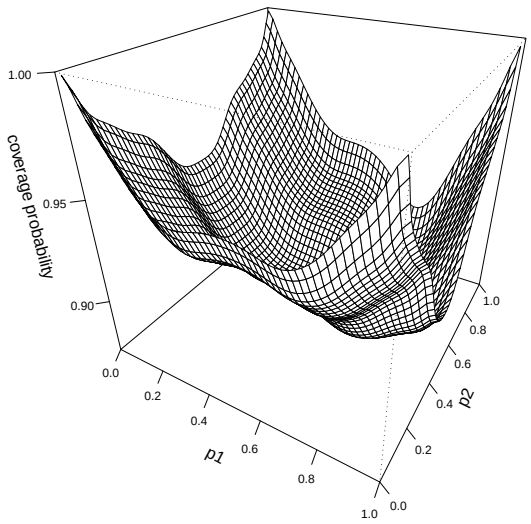

Figure 19: The coverage probability corresponding to different p_1 and p_2 under $\alpha = 0.05$ and $p_3 = p_1^2 \times p_2$. Coverage probability of the prediction interval corresponding to (n_1, n_2, n_3, m_1, m_2, m_3)= (30,30,10,15,5,3) with the maximum value = 0.9987445 occurring at (p_1, p_2) = (0.01,0.01) and the minimum value = 0.8720496 occurring at (p_1, p_2) = (0.75,0.57) while $p_3 = p_1^2 \times p_2$.

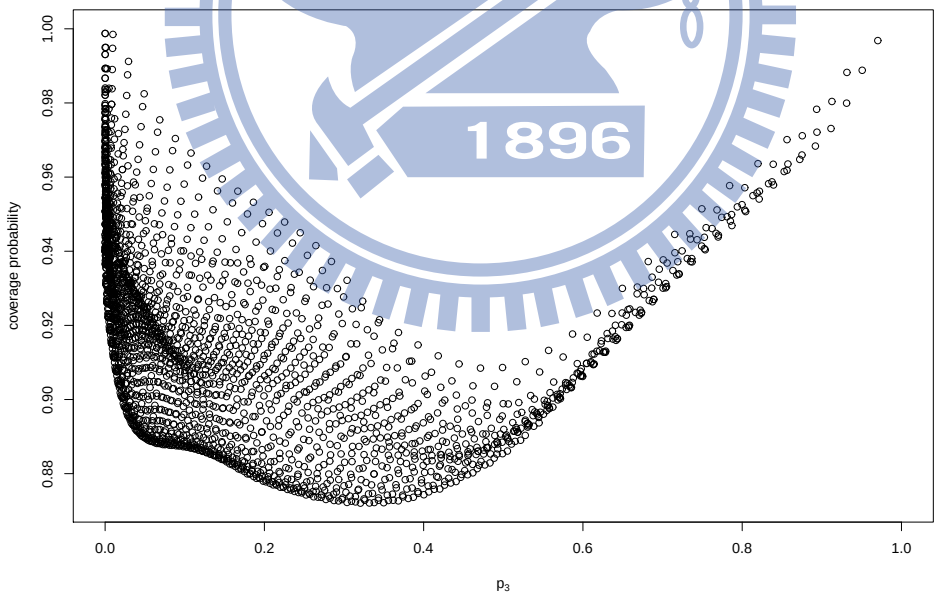

Figure 20: The coverage probability corresponding to different p_1 and p_2 under $\alpha = 0.05$ and $p_3 = p_1^2 \times p_2$. Coverage probability of the prediction interval corresponding to (n_1, n_2, n_3, m_1, m_2, m_3)= (30,30,10,15,5,3) with the maximum value = 0.9987445 occurring at p_3 = 0.0001 where (p_1, p_2) = (0.01,0.01) and the minimum value = 0.8720496 occurring at p_3 = 0.320625 while (p_1, p_2) = (0.75,0.57).

• **Coverage probability corresponding to different p_1 and p_2 under $p_3 = p_1^2 \times p_2$ with different sample size ($n_1 = 2n_2$)**

Figure 21 shows that the relationship between coverage probability and different (p_1, p_2) under $p_3 = p_1^2 \times p_2$ and different sample size while $n_1 = 2n_2$. Figure 22 shows that the relationship between coverage probability and different p_3 under different sample size while $n_1 = 2n_2$.

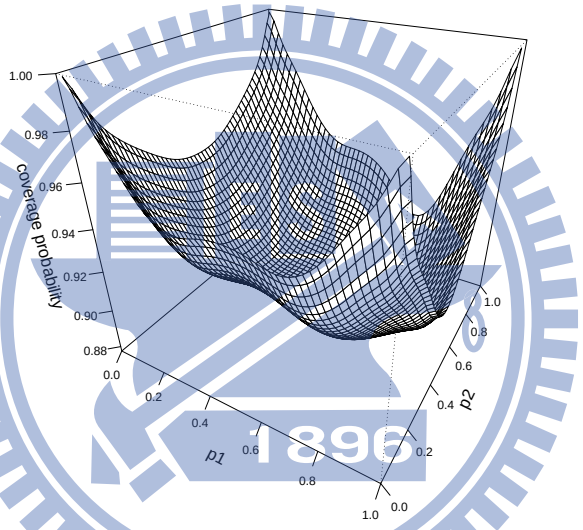

Figure 21: The coverage probability corresponding to different p_1 and p_2 under $\alpha = 0.05$ and $p_3 = p_1^2 \times p_2$. Coverage probability of the prediction interval corresponding to (n_1, n_2, n_3, m_1, m_2, m_3)= (30,15,10,15,5,3) with the maximum value = 0.9988455 occurring at (p_1, p_2) = (0.01,0.01) and the minimum value = 0.8774135 occurring at (p_1, p_2) = (0.77,0.65) while $p_3 = p_1^2 \times p_2$.

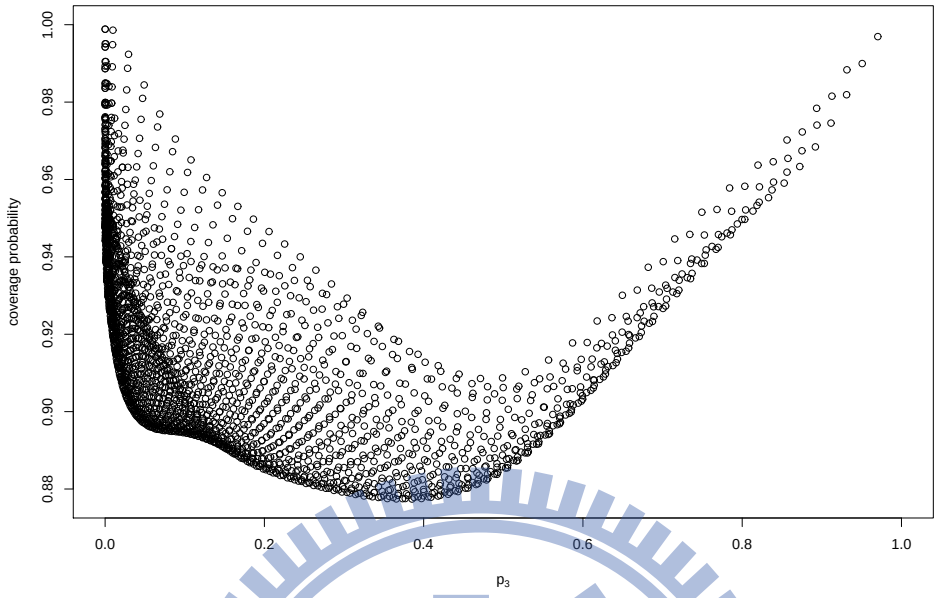

Figure 22: The coverage probability corresponding to different p_1 and p_2 under $\alpha = 0.05$ and $p_3 = p_1^2 \times p_2$. Coverage probability of the prediction interval corresponding to (n_1, n_2, n_3, m_1, m_2, m_3)= (30,15,10,15,5,3) with the maximum value = 0.9988455 occurring at p_3 = 0.000001 where (p_1, p_2) = (0.01,0.01) and the minimum value = 0.8774135 occurring at p_3 = 0.385385 while (p_1, p_2) = (0.77,0.65)

(2). The α trend

Figure 23 shows the coverage probability corresponding to different α and p_1 when p_2=0.5 and $p_3 = p_1 \times p_2$. Figure 24 shows the coverage probability corresponding to different α and p_1 when p_2=0.5 and $p_3 = p_1^2 \times p_2$. According to Figure 23 and 24, it shows a trend that the coverage probability increases by α decreasing.

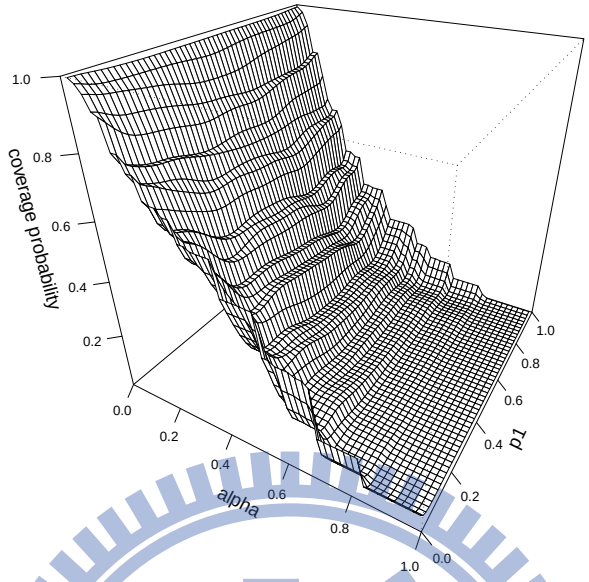

Figure 23: The coverage probability corresponding to different α and p_1 when $p_2 = 0.5$ and $p_3 = p_1 \times p_2$.

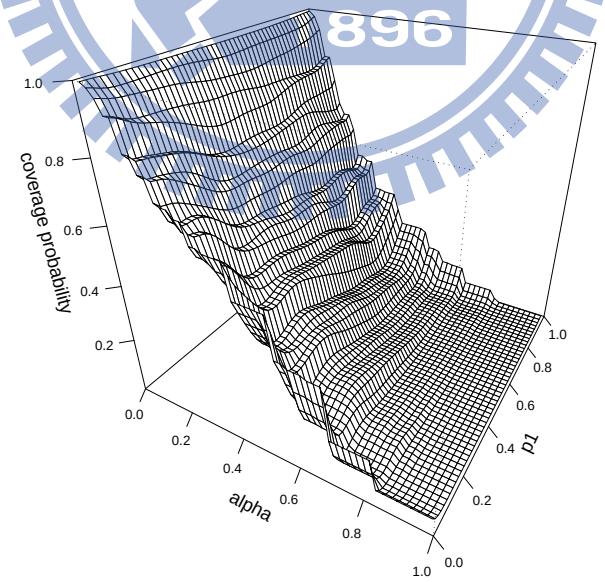

Figure 24: The coverage probability corresponding to different α and p_1 when $p_2 = 0.5$ and $p_3 = p_1^2 \times p_2$.

5 CONCLUSION

The prediction interval is widely used in industrial and medical applications. Literatures have provided methods for constructions of prediction intervals for discrete distributions. However, those existing methods cannot be applied to construct prediction interval for functions of multiple variables.

In this article, we have reviewed the prediction intervals of single binomial random variable. We extend the prediction interval which were proposed by Wang (2010) to the multiple binomial variables case. Our method proposes prediction interval for a linear function of multiple binomial random variables. The prediction interval method can be extended to construct prediction intervals for more models. We simulate some cases in which the two variables are independent or dependent and in which three variables are independent. In our simulations, we observe that the coverage probability of the prediction interval at least 0.88 and 0.86 while we consider two and three binomial random variables, respectively. The case in which variables are independent, we observe that the prediction interval has the characteristic that it is symmetric at p=0.5. In addition, comparing to the cases without relationship of variables, we observe the coverage probability of the prediction interval shift away the center.

The thesis can be generalized to additive models, such as there are relationship of parameters, p_i, for $i = 1, \ldots, k$. We can use the regression model to find the relationship between the variables and then using the prediction interval to predict the future observations.

References

[1] Bain, L. J., and Patel, J. K. (1993). Prediction Intervals Based on Partial Observations for Some Discrete Distributions, IEEE Transactions on Reliability, 42, 459-463.

[2] Basu, R., Ghosh, J.K., Mukerjee, R. (2003). Empirical Bayes Prediction Intervals in Normal Regression Model: Higher Order Asymptotics, Statistics and Probability Letters, 63, 197-203.

[3] Hall, P., and Rieck, A. (2001), Improving Coverage Accuracy of Nonparametric Prediction Intervals. Journal of the Royal Statistical Society, Ser. B, 63, 717-725.

[4] Hamada, M., Johnson, V., Moore, L. M., and Wendelberger, J. (2004). Bayesian Prediction Intervals and Their Relationship to Tolerance Intervals, Technometrics, 46, 452-459.

[5] Lawless, J. F., and Fredette, M. (2005). Frequentist Prediction Intervals and Predictive Distributions, Biometrika, 92, 529-542.

[6] Nelson, W. (1982). Applied Life Data Analysis. Wiley, NY.

[7] Wang, H. (2007). Exact Confidence Coefficients of Confidence Intervals for a Binomial Proportion, Statistica Sinica, 17, 361-368.

[8] Wang, H. (2008). Coverage probability of prediction intervals for discrete random variables, Computational Statistics and Data Analysis, 53, 17-26.

[9] Wang, H. (2009). Exact average coverage probabilities and confidence coefficients of confidence intervals for discrete distributions, Statistics and Computing, 19, 139-148.

[10] Wang, H. (2010).Closed form prediction intervals applied for disease counts, The American Statistician, 64, 250-256.

[11] Wilson, E. B. (1927).Probable Inference, the Low of Succession, and Statistical Inference, Journal of the American Statistical Association, 22, 209-212.