國 立 交 通 大 學

統計學研究所

碩 士 論 文

過度離散資料的容忍區間

The Tolerance Bounds for Overdispersion Data

研 究 生：詹千慧

指導教授：王秀瑛 博士

中 華 民 國 九 十 九 年 六 月

過度離散資料的容忍區間

The Tolerance Bounds for Overdispersion Data

研 究 生：詹千慧　　Student: Chien-hui Chan

指導教授：王秀瑛　　Advisor: Dr. Hsiuying Wang

國 立 交 通 大 學

統計學研究所

碩 士 論 文

A Thesis Submitted to
institute of Statistics
College of Science
National Chiao Tung University
in Partial Fulfillment of the Requirements
for the Degree of
Master
in
Statistics
June 2010

Hsinchu, Taiwan, Republic of China

中華民國九十九年六月

過度分散資料的容忍區間
The Tolerance Bounds for Overdispersion Data

研究生：詹千慧

指導教授：王秀瑛 博士

國立交通大學統計學研究所

摘要

我們都知道容忍區間被廣泛的使用於工業應用，例如醫藥工程、製程穩定研究等等，都是為了要偵測不良品的數量。不良品的數量通常假設服從二項式分配；然而，資料過度分散的現象卻很普遍的存在於二項式分配。在這篇文章中，我們使用 beta 二項式模型去適配過度分散的資料且提出方法去建立過度分散的資料的容忍區間。

關鍵字：過度離散、容忍區間、beta 二項式分配

The Tolerance Bounds for Overdispersion Data

Student: Chien-hui Chan Advisor: Dr. Hsiuying Wang

Institute of Statistics

National Chico Tung University

Abstract

It is well known that tolerance intervals are widely used in industrial applications, such as pharmaceutical engineering, process reliability studies, etc, to control the number of defective units. The number of defective units is usually assumed to follow a binomial distribution; however, it is common that overdispersion phenomenon exists for binomially distributed data. In this study, we use the beta-binomial model to fit the overdispersion data and propose an approach to construct tolerance intervals for overdispersion data.

Keywords: overdispersion, tolerance interval, beta-binomial distribution.

誌謝

一開始當然是要先感謝我的指導教授王秀瑛老師，身為所長，他不僅要忙所上事務，還要協助我們完成論文，真的很感謝老師，因為有你我才能這麼順利畢業。另外要感謝的就是為我們付出甚多的郭姊，郭姊就像是大家的媽媽，什麼事情都會細心叮嚀我們，大家煩心時，也可以找他聊聊天，是個紓解壓力的地方，有任何事情都可以找郭老師諮詢中心，棒透了。

雖然說兩年時間並不長，但在 408 的快樂時光卻是無可取代的，那些瘋狂美好的記憶是值得回憶的；大家都知道，心情悶的時候來 408 就對了，我們永遠有歡樂的笑聲，只要想的到的事情，我們都可以陪你一起完成唷。現在要各自畢業了，心中很是不捨，也謝謝你們陪我這兩年。過了碩士這階段，大家都要從新開始自己以後的人生，雖然不在彼此身邊了，但還是可以一起加油!

詹千慧　筆于風城

2010.06

Contents

List of Tables

1 Introduction

A tolerance interval is an important statistical tool and widely used in various practical applications, such as plant or animal inbreeding, and environmental monitoring, etc. A tolerance interval is constructed such that a certain proportion b of the population will be contained with a stated confidence level $(1-a)$ that gives an idea of what range each individual measurement should fall within. In other words, it is a statistical interval within which, with some confidence, a specified proportion of a population falls. Thus there are two different proportions associated with the tolerance interval: a degree of confidence and a percent coverage.

Let F denote the cumulative distribution for a random variable, X. An interval, $(L(X), U(X))$, is said to be a tolerance interval(**TI**), that is $b\%$ of the population will fall within with $1-a$ confident, denoted as a $(b, 1-a)$ **TI** satisfies the equation:

$$Pr_\theta\{[F(U(X)) - F(L(X))] \geq b\} = 1 - a. \tag{1}$$

But for discrete distributions, the value of the left hand side of (1) depends on the parameter θ. For a fixed β, the left side of (1) is not a constant. In this situation, it is reasonable to modify the definition of (1) for a discrete distribution as follows:

$$Pr_\theta\{[F(U(X)) - F(L(X))] \geq b\} \geq 1 - a \tag{2}$$

and there exists a θ such that the equality holds. For instance, we may be 95% confident that 90% of the population will fall within the range specified for a (0.9,0.95) TI.

Construction of tolerance intervals for univariate distributions has been extensively studied for continuous distributions or discrete distributions (Wang and Tung 2009, Perishability and Mathew 2004, Collani and Karl 2002, Mukerjee and

Reid 2001, Easterling and Weeks 1970, Vangel 1992). However, tolerance intervals have not been explored in detail for more complex situations, such as the data with overdispersion.

Because the tolerance intervals for some discrete variables such as binomial and Poisson distributions are widely used in industrial quality control, some applications require one-sided lower or upper tolerance bounds for the distribution of the number of defective units in future samples for the purpose of achieving the desirable quality control.

However, in industrial applications, a problem has been raised that count data exhibit variation greater than that predicted by the stochastic component of a model. In these cases, adopting a binomial model for modeling the data may lead to an unsatisfactory result. Such data are referred to as overdispersion that exits the presence of greater variability in a data set than would be expected based on a given simple statistical model. If the phenomenon of overdispersion occurs for binomially distributed data, the data is said to have extra-binomial variation. This reflects a lack of independence or heterogeneity, which is an adjective used to describe an object or system consisting of multiple items having a large number of structural variations among individuals.

In this study, we develop procedures for constructing a one-sided tolerance bound for discrete overdispersion data. The proposed methods are derived from the concept of generalized compound probability distributions.

This thesis is organized as follows. The methods of testing to check the data with overdispersion phenomenon are discussed in Section 2. The parameter estimation we use in this study are proposed in Section 3. In Section 4, we use the distribution-free interval to find the most appropriate quantile, and we demonstrate the simulation results with three different quantiles. An example from a semiconductor manufacturing process data is given in Section 5. Our conclusions and discussions are mentioned in Section 6.

1.1 Overdispersion with binomial data

Let $X_1, \ldots, X_n$ be a sample following a binomial distribution $bin(n, p)$, we have

$$X_i \sim bin(n, p) \tag{3}$$

and the variance is

$$var(X_i) = np(1-p). \tag{4}$$

When there does not exit a constant p such that the sample mean and sample variance of X_i simultaneously satisfy (3) and (4), the data is concluded to violate the binomial assumption. In this case, we conclude that there are overdispersion phenomenon for the data and we can assume the variance is

$$var(x_i) = \varphi \cdot np(1-p) \tag{5}$$

which is multiplying the variance with a constant φ and φ is defined as a dispersion parameter (Liang and McCullagh 1993 and Richards 2008).

If $\varphi = 1$, the data follows a binomial distribution. We say the data is overdispersion or underdispersion if $\varphi > 1$ or $\varphi < 1$, respectively. It is common that data is overdispersion. Richards (2008) points out the following two reasons to lead a overdispersion phenomenon.

- The model is not accounting for important covariates.
- There are interaction between binary responses (each treatments is not independent).

When the variability in the data exceeds what the binomial model can accommodate, it will cause standard error's underestimate and increase of the probability of Type I error if we still used the binomial model to fit the data. Thus, we should test whether the data is overdispersion first.

In general, there are three methods to test whether the data is overdispersion or not. If $\varphi > 1$, we think the data is overdispersion, so we need test whether φ is 1. The testing hypothesis is:

$$H_0 : \varphi = 1$$

$$H_1 : \varphi > 1$$

The test procedures are introduced in Section 2. If the p-value associated with a test is not too small, we do not reject H_0. In other words, it means that φ is close to 1, then there is no significant of overdispersion.

1.2 Method Dealing With Overdispersion Data

The first approach involves modeling the causes of overdispersion implicitly using compound probability distributions (beta-binomial distribution). For example, if all subjects of a treatment have the same probability of exhibiting a positive response (or success), then the number of successes among replicated treatments will follow a binomial distribution. However, in many cases the variation observed among replicates is often greater than that predicted by a binomial distribution. In this situation, we can model overdispersion by using compound probability distributions. Let $B(x; n, p)$ be the cumulative binomial distribution of X which follows a binomial distribution. We have

$$B(x; n, p) = Pr(X \leq x) = \sum_{i=0}^{x} \binom{n}{i} p^i (1-p)^{n-i} \tag{6}$$

Assume $X_1, ..., X_n$ is a sample following a $bin(n, p)$ with unknown p. Suppose that p varies randomly due to unknown covariates and the variation in p can be described by the probability density function $f(p)$. In this case, the probability x of the n subjects will exhibit a positive response:

$$Pr(x) = \int_0^1 f(p) \binom{n}{x} p^x (1-x)^{n-x}\, dp \tag{7}$$

The second approach ignores the causes of overdispersion and uses Akaikes information criterion (AIC) as a measure of the goodness of fit of an estimated statistical model. This criterion has been used in various fields of statistics, engineering and numerical analysis, and has a clear interpretation in model fitting. However, AIC is not a test of the model in the sense of hypothesis testing, rather it is a test between models - a tool for model selection. The best approximating model is the one which achieves the minimum AIC value compared to all models. Let $\mathbf{g_i}$ be the probability that if the study was performed, then the outcome indexed by i would be observed($\sum_i g_i = 1$) .In this study, the probability distribution defined by the $\mathbf{g_i}$, which we denote $\mathbf{g}$, is referred to as the truth. Suppose a stochastic model is proposed which predicts that outcome i will be observed with probability q_i. Let the distribution of predicted probabilities be denoted $\mathbf{q}$. An information-theoretic measure of the difference between the truth and the approximating model is the Kullback−Leibler distance (KLD):

$$I(\mathbf{g}, \mathbf{q}) = \sum_i^n g_i ln(\frac{g_i}{q_i}). \tag{8}$$

I is often interpreted as the information lost when the truth $\mathbf{g}$ is approximated by $\mathbf{q}$. The smaller the value of I, the better the model approximates the truth.

Let θ denote the set of model parameters (i.e. the g_i depend on θ). The best parameter values for a model according to information theory are those that minimize I. Unfortunately, because the truth, $\mathbf{g}$, is unknown in realistic cases, it is not possible to apply (8) directly to find the best θ. However, parameter values can be estimated readily by fitting the model to the data using maximum likelihood.

Suppose the study was repeated an infinite number of times and the processes generating the data did not change from one study to the next. If the model's parameters were reestimated each time using maximum likelihood, then the EKLD of the model would be:

$$E_g\{I(\mathbf{g}, \mathbf{q})\} = \sum_j g_j I(\mathbf{g}, \mathbf{q}(\widehat{\theta}_\mathbf{j})) \tag{9}$$

where θ_j is the set of maximum likelihood parameter estimates. Akaike (1973) established a relationship between the maximum likelihood, which is an estimation method used in many statistical analysis, and the EKLD. The model with the low EKLD value was usually consider a parsimonious model, because when we fit the data, it has lowest KLD. The proposed model having the lowest EKLD is referred to as the best EKLD model and is the model we wish to identify. Suppose a study resulted in outcome j, for a proposed model, M, its AIC value is defined as:

$$AIC(M) = 2k - 2lnL(\widehat{\theta}_\mathbf{j}) \tag{10}$$

this can also be written as:

AIC = 2 × (number of fitted parameters) − 2 × log(maximized likelihood for model)

Hence, K is the number of parameters in the statistical model, and L is the maximized value of the likelihood function for the estimated model. It can be shown that AIC estimates twice the models relative **EKLD**:

$$AIC(M) \approx 2(E_p\{I(\mathbf{g}, \mathbf{q})\} - c) \tag{11}$$

where $c = \sum_i g_i ln g_i$ is a constant common to all models that depends only on g. Hence, models with a low AIC value are more likely to be the best **EKLD** model. Although we know the method to find the model for the data with overdispersion, we have no idea to find the tolerance interval for these data. In This study, I try some method to find tolerance interval for data with overdispersion.

2 Testing for overdispersion data

The method of test is discussed in Ennis and Bi (1998), and Yinong and Chan (2008). Suppose an experiment involves n subjects being exposed to a treatment, and let p be the probability each individual in the replicate treatment exhibits a positive response. We will review the definition of the binomial distribution:

$$f(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad 0 < p < 1 \tag{12}$$

where x is the number of successes in a sequence of n independent experiments, and p is probability for success, and let the probability p follow a beta distribution with two parameters α and β, and the definition of the beta distribution is

$$f(p) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1} \tag{13}$$

where Γ is the gamma function $\Gamma(a) = (a-1)!$ and α and β are two positive parameters.

In this study, we assume that $f(p)$ can be well described by a beta density function because its flexibility and its ability to provide good approximations. The beta-binomial density function is

$$\begin{aligned} Pr(X = x) &= \int_{p=0}^{1} \binom{n}{x} p^x (1-p)^{n-x} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1} dp \\ &= \binom{n}{x} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_{p=0}^{1} p^{\alpha+x-1}(1-p)^{n-x+\beta+1} dp \\ &= \binom{n}{x} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha+x)\Gamma(n-x+\beta)}{\Gamma(\alpha+n+\beta)} \\ &= \frac{\Gamma(n+1)\Gamma(\alpha+\beta)\Gamma(x+\alpha)\Gamma(\beta+n-x)}{\Gamma(x+1)\Gamma(n-x+1)\Gamma(\alpha)\Gamma(\beta)\Gamma(\alpha+\beta+n)} \end{aligned}$$

where $x = 0, 1, 2...n$, α and β are two positive parameters, n is the total number of treatments, and x is the total number of subjects of defective units.

Hence, let $p_i = x_i/n_i, i = 1, 2, ...k$, where i indexes each studies, x_i is the number of events in the ith study and n_i is the sample size of the study. With the sample size n_i and binomial probability p_i that the proportion of defective units, the binomial distribution describes variation fully not possibly by itself, when p_i varies, the data therefore are fitted with a beta distribution with parameters (α, β). In short, we construct two model:

$$X_i|p_i \sim bin(n_i, p_i)$$

$$p_i \sim Beta(\alpha, \beta)(Each \;\; p_i \;\; is \;\; independent)$$

$$E(p_i) = \frac{\alpha}{\alpha + \beta} = \mu \tag{14}$$

$$\begin{aligned} var(p_i) &= \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \\ &= \frac{\alpha}{\alpha+\beta}\frac{\beta}{\alpha+\beta}\frac{1}{\alpha+\beta+1} \\ &= \mu(1-\mu)\gamma \end{aligned} \tag{15}$$

where $\gamma = 1/(\alpha + \beta + 1)$, hence, we can know the marginal expect value and variance of Y:

$$\begin{aligned} E(X_i) &= E(E(X_i|p_i)) \\ &= E(n_i p_i) \\ &= n_i \mu \end{aligned} \tag{16}$$

$$\begin{aligned} Var(X_i) &= E(var(X_i|p_i)) + Var(E(X_i|p_i)) \\ &= E(n_i p_i(1-p_i)) + Var(n_i p_i) \\ &= n_i(\mu - Var(p_i) - \mu^2) + n_i^2 \mu(1-\mu)\gamma \\ &= n_i(\mu - \mu(1-\mu)\gamma - \mu^2) + n_i^2 \mu(1-\mu)\gamma \\ &= n_i(\mu(1-\mu))(1-\gamma) + n_i^2 \mu(1-\mu)\gamma \\ &= n_i\mu(1-\mu)(1+(n-1)\gamma) = n_i\mu(1-\mu)\cdot\varphi \end{aligned} \tag{17}$$

We can view the term γ as a multiplier of the binomial variance. In other words, it models the overdispersion. So we can test $\gamma \geq 0$ or $\varphi \geq 1$ to see whether the data is overdispersion or not.

For discrete distributions, Pearson's χ^2 statistic is often used for testing goodness-of-fit. The test hypothesis is:

$$H_0 : \varphi = 1$$

$$H_1 : \varphi > 1$$

and the test-statistic for Pearson's χ^2(Ismail, N. and Jemain, A.A., 2007) is:

$$\chi^2 = \frac{\sum_i (x_i - \mu_i)^2}{n_i \mu_i (1 - \mu_i)} \tag{18}$$

$$\hat{\varphi} = \frac{\chi^2}{DF} \tag{19}$$

Under H_0 the statistic χ^2 is chi-square distributed with $DF = n - k$ degrees of freedom, where n denotes total number of treatments and k the number of parameters. In other words, When the systematic part of the model is correct and the binomial assumption holds, χ^2 is approximately chi-square distributed with $DF = n - k$ degrees of freedom. Hence χ^2 has an expectation of DF and $\hat{\varphi} \approx 1$. However, this test, has been found to be less sensitive in detecting departure from the binomial model because boundary problems arise as we test whether a positive-valued parameter is greater than 0.

We may use a likelihood ratio test(Yinong and Chan 2008) that can also be used to test for overdispersion. The null hypothesis is that distribution is binomial versus the alternative hypothesis is that the distribution is beta-binomial. The log-likelihood for the binomial model is

$$lnL = ln\binom{n}{x} + xln(p) + (n - x)ln(1 - p) \tag{20}$$

The likelihood ratio test is

$$\chi^2 = 2(L_{betabinom} - L_{binom}) \tag{21}$$

where $L_{betabinom}$ is the log-likelihood value for the beta-binomial model and L_{binom} is log-likelihood value for the binomial model.

The third method, we can use Tarone's Z statistic(Ennis and Bi 1998) to test the goodness of fit test for binomial against beta binomial distribution. This has been shown to be more sensitive than the parameter test (e.g. test for γ being zero) and the log-likelihood ratio test:

$$Z = \frac{E - \sum_{i=1}^{k} n_i}{\sqrt{2\sum_{i=1}^{k} n_i(n_i - 1)}} \tag{22}$$

where

$$E = \sum_{i=1}^{k} \frac{(x_i - n_i\widehat{p})^2}{\widehat{p}(\widehat{1-p})} \tag{23}$$

$$\widehat{p} = \sum_{i=1}^{k} \frac{x_i}{n_k} \tag{24}$$

This statistic Z on (22) has an asymptotic standard normal distribution under the null hypothesis of a binomial distribution.

Rejection of test can be used to detect if the data of the model exit overdispersion. So, once we detect the presence of overdispersion, there are usually two approaches to deal effectively with overdispersion. The first approach involves modeling the causes of overdispersion implicitly using compound probability distributions(beta-binomial distribution). The second approach ignores the causes of overdispersion and uses Akaike′s information criterion (AIC) as a criterion.

3 Parameter estimation

For overdispersion data $x_1, ..., x_n$, we suggest to model them with a beta-binomial distribution. There are two true parameters, $\theta = (\alpha, \beta)$ that we need to estimate. We adopt two methods introduced as follows to estimate α and β.

3.1 Method 1 to estimate α and β

Let $\alpha = \frac{\eta}{\phi}$ and $\beta = \frac{1-\eta}{\phi}$, $0 < \eta < 1, 0 < \phi < 1$. The likelihood function of $x_1, ..., x_n$ is

$$L(\eta, \phi) = \prod_i \frac{\Gamma(n+1)\Gamma(\alpha+\beta)\Gamma(x_i+\alpha)\Gamma(n-x_i+\beta)}{\Gamma(x_i+1)\Gamma(n-x_i+1)\Gamma(\alpha)\Gamma(\beta)\Gamma(n+\alpha+\beta)}. \tag{25}$$

We need to derive $\hat{\eta}$ and $\hat{\phi}$ such that

$$max_{\eta,\phi} L(\eta, \phi) = L(\hat{\eta}, \hat{\phi}).$$

Then we can derive α and β estimators

$$\widehat{\alpha} = \frac{\widehat{\eta}}{\widehat{\phi}}, \widehat{\beta} = \frac{1-\widehat{\eta}}{\widehat{\phi}}. \tag{26}$$

Note that the condition

$$0 < \eta < 1, \quad 0 < \phi < 1.$$

leads to

$$0 < \alpha < \infty, \quad 0 < \beta < \infty.$$

Let

$$f(\eta, \phi) = ln(L) = ln(\prod_{i=1}^{m} f(x_i)) = \sum_{i=1}^{m} lnf((x_i)) \tag{27}$$

The following steps is used to find a value of (η, ϕ) to maximize L:

- Step 1: Let A=$\{(\eta,\phi):\eta=\frac{i}{1000},\phi=\frac{j}{1000},i=0,1,2,\cdots,999,j=0,1,2,\cdots,999\}$
- Step 2: Calculate the likelihood function L at the points in A. The likelihood function L is

$$L=\prod_i \frac{\Gamma(n+1)\Gamma(\frac{\eta}{\phi}+\frac{1-\eta}{\phi})\Gamma(x_i+\frac{\eta}{\phi})\Gamma(n-x_i+\frac{1-\eta}{\phi})}{\Gamma(x_i+1)\Gamma(n-x_i+1)\Gamma(\frac{\eta}{\phi})\Gamma(\frac{1-\eta}{\phi})\Gamma(n+\frac{1}{\phi})}$$

- Step 3: We select the point (η^*,ϕ^*) in set A of Step 1such that the likelihood function L is maximized at the point (η^*,ϕ^*). Then (η^*,ϕ^*) are desirable estimators of (η,ϕ)

3.2 Method 2 to estimate α and β

The second method is to derive the maximum likelihood estimators of α and β. Although the method 1 is also a method to derive estimators of α and β such that the likelihood function has the maximum value when α and β occurs at these estimators, it transform the α and β to η and ϕ. The second method directly maximize likelihood estimators by a numerical approach. The Maximum likelihood estimation(MLE) seeks the value of the parameter vector to maximizes the likelihood function. It is a totally analytic maximization procedure. It begins with a mathematical expression known as the Likelihood function, that is the probability of obtaining that particular set of data, given the chosen probability distribution model.

MLE is a common statistical method used to fitting the model by sample data, and corresponds to many famous estimation method in statistics. From a statistical viewpoint, the mle is more robust and yields estimators with good properties by considering. The MLE method is many-sided, and is suitable for the major part model and the different type data. Moreover, they provide the high efficiency method for the quota uncertainty through the confidence region.

Although the maximum likelihood estimate's methodology is simple, implements in mathematics are intense. However, complex mathematics formula is not a big problem by using present's computer science and technology.

4 Distribution-free tolerance interval

We briefly describe the distribution-free tolerance interval in this section. Let P_{LE} and P_{GE} is the proportion of the population fall within the range specified by the tolerance interval. If we want to find an upper bound, because we hoped that the bound is smaller, therefore it is essential to seek for the smallest integer ν such that

$$Pr(y \leq \nu) = B(\nu; n, p) \geq P_{LE} \tag{28}$$

is an upper probability bound on the number of defective units. And if we want to find the lower bound, because we hoped that the bound is greater, therefore it is essential to seeks for the the largest integer ω such that

$$Pr(y \geq \omega) = 1 - B(\omega; n, p) \geq P_{GE} \tag{29}$$

For example, the sample size n is 30, $P_{LE} = 0.90$, and $\tilde{p} = 0.03$ is the one-sided upper 95% confidence bound for the population proportion p of defective units the population proportion p of defective units. Because, $Pr(y \leq 1) = B(1; 30, 0.03) = 0.77$, and $Pr(y \leq 2) = B(2; 30, 0.03) = 0.94$, the upper tolerance bound is 2. In other words, we are 95% confident that at least 90%(more precisely, at least 0.94) of the products, the number of defective units will less or equal than 2.

4.1 Tolerance bound

We briefly describe the approach of deriving the distribution-free tolerance bound. Let $X_1, \ldots, X_n$ be n independent random variables. Suppose $(X_{(1)}, \ldots, X_{(n)})$

are the order statistics that comes from $(X_1, \ldots, X_n)$ is one of the data set from $(X_1, \ldots, X_n)$. That is, $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$. Hence, $X_{(1)}$ has the minimum observed value, and $X_{(n)}$ has the maximum observed value.

Let $F_r(x)$ denote the cumulative density function(cdf) of the rth order statistic $X_{(r)}$, and $r = 1, 2, \ldots, n$. Then the cdf of the largest order statistic $X_{(n)}$ is given by

$$\begin{aligned} F_n(x) &= Pr\{X_{(n)} \leq x\} \\ &= P(max\ X_i \leq x) \\ &= Pr(X_1 \leq x, \cdots, X_n \leq x) \\ &= \prod_{i=1}^{n} P(X_i \leq x) \\ &= (F_X(x))^n \end{aligned} \tag{30}$$

Likewise we have the smallest order statistic $X_{(1)}$ is given by

$$\begin{aligned} F_1(x) &= Pr\{X_{(1)} \leq x\} \\ &= P(min\ X_i \leq x) \\ &= 1 - p(min\ X_i \geq x) \\ &= 1 - Pr(X_1 \geq x, \cdots, X_n \geq x) \\ &= 1 - \prod_{i=1}^{n} P(X_i \geq x) \\ &= 1 - (1 - F_X(x))^n \end{aligned} \tag{31}$$

There are important special cases of the general result for $F_r(x)$:

$$\begin{aligned} F_r(x) &= Pr\{X_{(r)} \leq x\} \\ &= Pr\{at\ least\ r\ of\ the\ X_i \leq x\} \\ &= \sum_{j=r}^{n} \binom{n}{j} (F_X(x))^j (1 - F_X(x))^{n-j} \end{aligned} \tag{32}$$

We now assume that X_i is continuous with probability density function(*pdf*) $p(x) = F'(x)$. If $f_r(x)$ denote the *pdf* of $X_{(r)}$ we have

$$f_r(x) = \frac{n!}{(r-1)!(n-r)!}(F(x_{(r)}))^{r-1}(1-F(x_{(r)}))^{n-r}f(x_{(r)}) \tag{33}$$

hence, the general result for $F_r(x)$:

$$\begin{aligned} F_r(x) &= \sum_{j=r}^{n} \binom{n}{j} (F_X(x))^j (1-F_X(x))^{n-j} \\ &= \int_0^{F_X(x)} \frac{n!}{(r-1)!(n-r)!}(F(x_{(r)}))^{r-1}(1-F(x_{(r)}))^{n-r}f(x_{(r)})\,dx_{(r)} \\ &= I_{F_X(x)}(r, n-r+1) \end{aligned} \tag{34}$$

where

$$\begin{aligned} I_p(a,b) &= \int_0^p t^{a-1}(1-t)^{b-1}dt / \int_0^1 t^{a-1}(1-t)^{b-1}dt \\ &= \int_0^p t^{a-1}(1-t)^{b-1}dt B(a,b) \end{aligned} \tag{35}$$

and

$$B(a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}$$

The joint density function of $X_{(r)}$ and $X_{(s)}(1 \le r < s \le n)$ is conveniently denoted by $f_{rs}(x,y)$. It follows that for $x \le y$

$$f_{rs}(x,y) = \frac{n!}{(r-1)!(s-r-1)!(n-s)!}(F(x))^{r-1}(F(y)-F(x))^{s-r-1}(1-F(y))^{n-s}p(x)p(y) \tag{36}$$

Let ξ_p be the *pth* quantile. We shall now show that if X is continuous the random interval $(X_{(r)}, X_{(s)})$ covers ξ_p with a probability which depends on r,s,n and p, thus allowing the construction of distribution-free confidence intervals for ξ_p. Since $X_{(s)} < \xi_p$, we have, whether X is continuous or not,

$$P_r\{X_{(r)} \le \xi_p \le X_{(s)}\} = P_r\{X_{(r)} \le \xi_p\} - P_r\{X_{(s)} \le \xi_p\} \tag{37}$$

It follows from $F_r(x) = I_{p(x)}(r, n-r+1)$ that $(X_{(r)}, X_{(s)})$ covers ξ_p with probability $\pi(r, s, n, p)$ given by

$$\begin{aligned}\pi(r,s,n,p) &= F_r(x) - F_s(x) \\ &= I_p(r, n-r+1) - I_p(s, n-s+1) \\ &= \sum_{i=r}^{s-1} \binom{n}{i} P^i (1-p)^{n-i} \end{aligned} \tag{38}$$

To obtain a distribution-free TI, the requirement of a TI$(X_{(r)}, X_{(s)})$ is that it contains at least a proportion γ of the population with probability $1-a$. Thus, if $h(x)$ is continuous we seek $X_{(r)}, X_{(s)}$ such that

$$P_r\{\int_{X_{(r)}}^{X_{(s)}} h(x)dx \geq \gamma\} = 1 - a \tag{39}$$

and we can write as

$$P_r\{\int_{X_{(r)}}^{X_{(s)}} h(x)dx \geq \gamma\} = P_r\{P(X_{(s)}) - P(X_{(r)}) \geq \gamma\} \tag{40}$$

Let $W_{rs} = P(X_{(s)}) - P(X_{(r)})$. The density function is

$$f(w_{rs}) = \frac{1}{B(s-r, n-s+r+1)} w_{rs}^{s-r-1}(1-w_{rs})^{n-s+r} \quad 0 \leq w_{rs} \leq 1 \tag{41}$$

which is a beta distribution Beta(s-r,n+1-s+r). Therefore,

$$Pr\{W_{rs} \geq \gamma\} = 1 - I_\gamma(s-r, n-s+r+1) \tag{42}$$

Hence, we want to find r such that $Pr\{W_{rs} \geq \gamma\} = 1 - I_\gamma(s-r, n-s+r+1) = 0.95$ with the proportion $\gamma = 0.9$, $n = 1000$, and $s = 1000$. By numerical calculation, we have $r = 84$. It also means that the TI $(X_{(84)}, X_{(1000)})$ contains at least a proportion 0.9 of the population with probability 0.95

4.2 Procedure

The following is the simulation procedure for obtain the coverage probabilities of $(0.1, 0.95)$ distribution-free tolerance bound.

Table 1: The coverage probability for tolerance bound of x_{80}, x_{84} and x_{90} quantile

quantile \ a	1	2	3	4	5	6	7	8	9
80th	0.9680	0.9728	0.9472	0.9641	0.9699	0.9606	0.9644	0.9583	0.9711
90th	0.9230	0.9209	0.9225	0.9189	0.9197	0.9119	0.9259	0.9302	0.9333
84th	0.9539	0.9602	0.9400	0.9468	0.9508	0.9538	0.9426	0.9414	0.9687

- Step 1: Suppose α and β are known. Generate $p_i, i = 1, 1000$ from the beta distribution $beta(\alpha, \beta)$.
- Step 2: Generate $x_i \sim binomial(n, p_i), i = 1, ..., 1000.$ of x. Let TRUETI=$x_{(100)}$ and then we derive estimators $\hat{\alpha}$, $\hat{\beta}$ for α and β, by method 1 or method 2.
- Step 3: Generate $p_i \sim beta(\hat{\alpha}, \hat{\beta})$, $i = 1, ..., 1000$. Then generate $x_i \sim binomial(n, p_i)$, $x_i, i = 1, ..., 1000$. Let TI=$x_{(90)}$.
- Step 4: Repeat Step 1 to Step 3 to calculate the proportion that TIs derived in Step 3 are less than TRUETI. The proportion is the coverage probability of the tolerance bound.

4.3 Simulation result

The tolerance bound $x_{(84)}$ derived before, we also calculate the coverage probability for different tolerance bound based on the order statistics $x_{(80)}$ and $x_{(90)}$ for different α and β values. The results are presented in Table 1 for $\beta = 1, ..., 9$ and $\alpha = 1, ..., 9$.

Table 1 shows that the coverage probability based on the tolerance bound x_{80} is always higher than 0.95, and the coverage probability of tolerance bound x_{90} is always lower than 0.95. Compare with these two tolerance bounds, the coverage probability of $84th$ is closer to 0.95.

5 Real data example

5.1 Introduction of data

In this section, we use a real data example to illustrate the method. The real data is from a semiconductor manufacturing process. The location of chips on a wafer that is measured on 30 wafers. On each wafer, 50 chips are measured and a defective is defined whenever a misregistration, in terms of horizontal and/or vertical distances from the center, is recorded. The chips data are listed in Table 2.

Table 2: The semiconductor data

Sample number	fraction defectives	Sample number	fraction defectives
1	0.24	16	0.16
2	0.3	17	0.2
3	0.16	18	0.1
4	0.2	19	0.26
5	0.08	20	0.22
6	0.14	21	0.4
7	0.32	22	0.36
8	0.18	23	0.48
9	0.28	24	0.3
10	0.2	25	0.18
11	0.1	26	0.24
12	0.12	27	0.14
13	0.34	28	0.26
14	0.24	29	0.18
15	0.44	30	0.12

The data can be obtained from the NIST/SEMATECH e-Handbook of Statistical Methods: http:// www.itl.nist.gov/div898/handbook/pmc/section3/pmc332.htm. We first assume that the defective numbers follow a binomial distribution, Bin(50,θ), then test if the overdispersion parameter ϕ is equal to 1. It reveals that the data are overdisperion. Therefore, we fit the data with a beta-binomial model.

To obtain tolerance bounds for the data, we adopted method 1 and method 2(MLE method) to estimate the parameters (α, β) in a beta distribution. If we let real data be the population, and sampling 15 values from population without replacement. The estimated α and β values are $\hat{\alpha} = 4.3333$, $\hat{\beta} = 6.7778$ and $\hat{\alpha} = 3.6078$, $\hat{\beta} = 6.8662$(MLE method) respectively. The true tolerance bound calculating by real data is 5. The coverage probability by method 1 is about 0.9828 by average, and by method 2 is about 0.9332 by average.

If we treat real data as the sample values which sampling from 100 population without replacement. The estimated α and β values are $\hat{\alpha} = 2.8571$, $\hat{\beta} = 4.2857$ and $\hat{\alpha} = 3.0737$, $\hat{\beta} = 4.8203$(MLE method) respectively. The mean lower tolerance bounds corresponding to these two sets of $(\hat{\alpha}, \hat{\beta})$ are 5 and 4, respectively.

5.2 AIC value comparison

We know $AIC = 2k - 2lnL(\widehat{\theta}_{\mathbf{j}})$, and at this real data example:

$$L = \prod_{i=1}^{10} \int_{p=0}^{1} \frac{\Gamma(\widehat{\alpha}+\widehat{\beta})}{\Gamma(\widehat{\alpha})\Gamma(\widehat{\beta})} p^{\widehat{\alpha}-1} 1 - p^{\widehat{\beta}-1} \binom{n}{x_i} p_i^{x} (1-p)^{n-x_i}\, dp \tag{43}$$

where

$$k = 2$$

and

$$\begin{aligned} &\int_{p=0}^{1} \frac{\Gamma(\widehat{\alpha}+\widehat{\beta})}{\Gamma(\widehat{\alpha})\Gamma(\widehat{\beta})} p^{\widehat{\alpha}-1} (1-p)^{\widehat{\beta}-1} \binom{n}{x_i} p_i^{x} (1-p)^{n-x_i}\, dp \\ &= \binom{n}{x_i} \frac{\Gamma(\widehat{\beta}+n-x_i)}{\Gamma(\widehat{\beta})\Gamma(\widehat{\alpha}+\widehat{\beta})\Gamma(\widehat{\alpha})\Gamma(x_i+\widehat{\alpha})\Gamma(\widehat{\beta}+n+\widehat{\alpha})} \end{aligned}$$

The AIC value for Method 1 is about 183, and it is 61 for Method 2. Thus, we conclude that Method 2 leads to better estimators for the unknown parameters α and β because it has a lower AIC value.

6 Conclusion

In industrial or other applications, the overdispersion phenomenon usually occurs for binomial data or poisson data. Since in this case, the conventional tolerance limit does not lead to a satisfactory result, in this study, we develop procedures for constructing a one-sided tolerance bounds for discrete overdispersion data.

In this study, we fit the data with a beta-binomial distribution and use two methods to estimate the unknown parameters for the beta distribution. Procedures for deriving distribution-free tolerance intervals are established. We also conduct a simulation to calculate the coverage probability of the derived tolerance bound. The results show that the proposed method can lead to a satisfactory result.

A The main matlab code

A.1 real data example by method 1

```
err=0;
m=30;
n=30; mm=50; s=0;
eta=0;
#To read data from computer
x=load('D:/data.txt');
x=x*50;
#To find eta and phi such that maximizes likelihood function
for i = 1:mm
 max=-10000;
  while(eta<1)
   phi=0;
   eta=eta+0.01;
   while(phi<1)
     phi=phi+0.01;
     j=1;
     if(eta<1)
       sumf=0;
       for j = 1:m
         f=log(gamma(n+1)/(gamma(x(j)+1)*gamma(n-x(j)+1))
           *gamma(x(j)+eta/phi)*gamma(n-x(j)+(1-eta)/phi));
         sumf=sumf+f;
       end
```

```
      ff=m*log(gamma(1/phi)/gamma(eta/phi)/gamma((1-eta)/phi)/gamma(1/phi+n));
      total=sumf+ff;
      #To choose the phi and eta such that maximizes likelihood function
      if(max<total)
        max=total;
        truephi=phi;
        trueeta=eta;
      end
     end
    end
 end
ahead=trueeta/truephi; % alpha=eta/phi
bhead=(1-trueeta)/truephi; % beta=(1-eta)/phi
a=ahead;
b=bhead;
x=sort(x); %sort the x of the population
TRUETI=x(3); % let TRUETI=the 10th quantile of x
k=0;
jj=1;
  for jj = 1:mm
    p=betarnd(a,b,1000,1);
    xx=binornd(n,p);
    xx=sort(xx);
    TI(jj)=xx(84);
    if(TI(jj)<=TRUETI)
       k=k+1;
```

```
      end
   end
 point = k/mm;
 s=s+point;
end
#To calculate the mean of the proportion
s/mm
```

A.2 real data example by method 2

```
nn=30;
m=50;
s=0;
#To start iterations
for i = 1:m
#To read data from the computer}
population=load('D:/data.txt');
population=population*50;
x=sort(population);
#To be TRUETI=the 10th quantile of x
TRUETI=x(3); let TRUETI=the 10th quantile of x
#To rand sampling from population by 15 times without replacement
sample=randsample(population,15,'false')
data = sample/nn;
j=1;
#To be all data unequal 0 or 1
for j = 1:15
  if data(j) == 1
   data(j) = data(j)-0.01;
  end
  if data(j) == 0
   data(j) = data(j)+0.01;
  end
  j=j+1;
end
```

```
#To find the MLE of alpha and beta
phat = betafit(data);
k=0;
j=1;
#To compare TI with TRUETI m times
for j = 1:m
  #To generate p from beta distribution with parameter alpha and beta of MLE
  p=betarnd(phat(1,1),phat(1,2),1000,1);
  #To generate x from beta distribution with parameter nn and p
  x=binornd(nn,p);
  x=sort(x);
  #To be TI=the 8.4th quantile of x
  TI(j)=x(84);
  #To calculates how many TI is smaller than TRUETI in m times
  if(TI(j) <= TRUETI)
   k=k+1;
  end
 end
 #To calculate the proportion that TI is smaller than TRUETI in m times
 point = k/m;
 To add all proportion that TI is smaller than TRUETI in m times
 s=s+point;
end
#To calculate the coverage probability
s/m
```

Table 3: The coverage probability with $80th$

b	1	2	3	4	5	6	7	8	9	average
a=1	0.9116	0.9408	0.9388	0.9960	0.9440	0.9872	0.9964	0.9976	1.0000	0.968
a=2	0.8872	0.9336	0.9872	0.9656	0.9960	0.9964	1.0000	0.9960	0.9928	0.9728
a=3	0.8448	0.9136	0.9316	0.9480	0.9704	0.9408	0.9960	0.9980	0.9812	0.9472
a=4	0.8816	0.9668	0.9548	0.9528	0.9880	0.9880	0.9808	0.9812	0.9832	0.9641
a=5	0.9136	0.9368	0.9616	0.9868	0.9656	0.9928	0.9812	0.9984	0.9924	0.9699
a=6	0.8404	0.9440	0.9680	0.9784	0.9656	0.9884	0.9984	0.9792	0.9828	0.9606
a=7	0.8904	0.9112	0.9744	0.9732	0.9812	0.9828	0.9844	0.9964	0.9852	0.9644
a=8	0.8808	0.9416	0.9528	0.9720	0.9808	0.9496	0.9888	0.9652	0.9932	0.9583
a=9	0.9060	0.9520	0.9800	0.9900	0.9752	0.9968	0.9984	0.9688	0.9728	0.9711

Table 4: The coverage probability with $84th$

b	1	2	3	4	5	6	7	8	9	average
a=1	0.8360	0.9556	0.9368	0.9996	0.9488	0.9188	0.9912	0.9988	0.9996	0.9539
a=2	0.8492	0.9200	0.9464	0.9628	0.9928	0.9956	0.9892	0.9856	1.0000	0.9602
a=3	0.8092	0.9252	0.9160	0.9516	0.9324	0.9728	0.9664	1.0000	0.9864	0.9400
a=4	0.8552	0.9060	0.9552	0.9244	0.9752	0.9556	0.9864	0.9832	0.9800	0.9468
a=5	0.8768	0.9480	0.9188	0.9372	0.9748	0.9464	0.9920	0.9664	0.9972	0.9508
a=6	0.8240	0.9288	0.9712	0.9360	0.9940	0.9720	0.9736	0.9976	0.9868	0.9538
a=7	0.8184	0.9588	0.9432	0.9336	0.9588	0.9600	0.9596	0.9540	0.9968	0.9426
a=8	0.8116	0.9332	0.9532	0.9252	0.9720	0.9800	0.9616	0.9560	0.9796	0.9414
a=9	0.9188	0.9580	0.9652	0.9520	0.9796	0.9952	0.9616	0.9896	0.9984	0.9687

Table 5: The coverage probability with 90th

b	1	2	3	4	5	6	7	8	9	average
a=1	0.8576	0.8848	0.8276	0.9996	0.9200	0.8772	0.9564	0.9844	0.9996	0.9230
a=2	0.7216	0.8248	0.9164	0.9472	0.9760	0.9424	0.9840	0.9760	1.0000	0.9209
a=3	0.8692	0.7952	0.8964	0.9096	0.9728	0.9512	0.9800	0.9532	0.9748	0.9225
a=4	0.7528	0.8932	0.8788	0.9720	0.9316	0.9800	0.9360	0.9484	0.9772	0.9189
a=5	0.8628	0.8496	0.8608	0.8792	0.9568	0.9628	0.9432	0.9772	0.9848	0.9197
a=6	0.7132	0.8736	0.8540	0.9428	0.9704	0.9720	0.9592	0.9784	0.9432	0.9119
a=7	0.7836	0.9208	0.9460	0.9080	0.9284	0.9748	0.9368	0.9436	0.9908	0.9259
a=8	0.8120	0.9264	0.8732	0.9372	0.9668	0.9396	0.9636	0.9564	0.9968	0.9302
a=9	0.8020	0.9428	0.8988	0.9364	0.9304	0.9544	0.9920	0.9672	0.9756	0.9333

References

[1] Cai, T. and Wang, H. (2009). "Tolerance Intervals for Discrete Distributions in Exponential Families," Statistica Sinica, 19, 905-923.

[2] Cover, T.M and Thomas, J.A. (1991). Elements of Information Theory, New York: Wiley.

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

[4] Ennis, D.M. and Bi, J. (1998). "The beta-binomial model: accounting for intertrial variation in replicated difference and preference tests," Journal of Sensory Studies, 13, 389-412.

[5] Hahn, G. J., and Meeker, W. Q. (1991), Statistical Intervals: A Guide for Practitioners, New York: Wiley.

[6] Liang, K.-Y. and McCullagh, P. (1993) Case studies in binary dispersion. Biometrics, 49, 623V630.

[7] Ismail, N. and Jemain, A.A. (2007). "Handling overdispersion with Negative Binomial and Generalized Poisson regression models," Casualty Actuarial Society Forum Winter, 103- 158.

[8] 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.

[9] 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.

[10] Richards, S.A. (2008). "Dealing with overdispersed count data in applied ecology," Journal of Applied Ecology, 45, 218-227.

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

[12] Viele, K. (2007). "Nonparametric Estimation of Kullback-Leibler Information Illustrated by Evaluating Goodness of Fit," Bayesian Analysis, 2, 2, 239-280.

[13] Wang, H. and Tsung, F.(2009). "Tolerance Intervals With Improved Coverage Probabilities for Binomial and Poisson Variables," Technometrics, 51, 25-33.

[14] Williams, D.A. (1982). "Extra binomial variation in logistic linear models," Applied Statistics, 31, 144-148.

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

[16] Xiang, L., Lee, A.H., Yau, K.K.W. and McLachlan, G.J. (2007). "A score test for overdispersion in zero-inflated poisson mixed regression model," Statistic in Medicine, 26, 22-1608.

[17] Yinong, Y.X.and Chan, K.A. (2008). "Pooling overdispersed binomial data to estimate event rate," BMC Medical Research Methodology, 8, 58.