

# ON CONFIDENCE INTERVALS FOR THE NEGATIVE BINOMIAL DISTRIBUTION

Anwer Khurshid<sup>1</sup>, Department of Mathematics and Statistics, College of Science, Sultan Qaboos University, Muscat, Sultanate of Oman. Department of Statistics, University of Karachi, Pakistan.

Mohammed I. Ageel<sup>2</sup>, Department of Mathematics, Faculty of Science, King Khalid University, Abha, Saudi Arabia

Raheeq A. Lodhi, Department of Statistics, University of Karachi, Pakistan.

## ABSTRACT

The paper provides a comprehensive review of methodology, classical as well as bootstrapped, for setting confidence intervals for the parameter  $p$  of negative binomial distribution. The results are illustrated by a numerical example. The authors have developed a computer program in Visual BASIC which computes confidence limits using the procedures described in this paper.

**Key words:** negative binomial; confidence interval; normal approximation; Box-Cox transformation; chi-square distribution; bootstrap; bootstrap resampling.

## RESUMEN

Este trabajo proporciona una revisión comprensiva de metodología, clásica así como la asociada al método de bootstrap, para establecer intervalos de confianza para el parámetro  $p$  de distribución del binomio negativo. Los resultados se ilustran por un ejemplo numérico. Los autores han desarrollado un programa de computadora en Visual Basic que calcula límites de confianza usando los procedimientos descritos en este trabajo.

MSC: 62f25.

## 1. INTRODUCTION

The negative binomial distribution (NBD) is one of the most useful probability distributions and has been successfully employed to model a variety of natural phenomena. It has been used to construct useful models in many substantive fields: in accident statistics (Arbous and Kerrich, 1951, Greenwood and Yule, 1920, Weber, 1971); in biology (Anscombe, 1950; Bliss and Fisher, 1953; Anderson, 1965; Boswell and Patil, 1970; Elliot, 1977); in birth-death processes (Furry, 1937; Kendall, 1949); in ecology (Martin and Katti, 1965; Binn, 1986; White and Bennetle, 1996); in entomology (Wilson and Room, 1983); in epidemiology (Byers et al. 2003); in information sciences (Bookstein, 1997); in meteorology (Sakamoto, 1972); and in psychology (Sichel, 1951). In addition, many other physical and biological applications have been described by Biggeri (1998) and Eke *et al.* (2001). For an excellent exposition of the negative binomial distribution and extensive list of references, see Johnson *et al.* (1992).

## 2. THE NEGATIVE BINOMIAL DISTRIBUTION

Greenwood and Yule (1920) are credited for first deriving and applying the NBD in the literature even though some special forms of this distribution were already discussed by Pascal. The NBD is the mathematical counterpart of (positive) binomial distribution. Mathematically, it is given by the expansion  $(Q - P)^k$  where  $Q = 1 + P$ ,  $P > 0$ ,  $k$  is positive real number; and the  $(x + 1)$  term in the expression yields  $P[X = x]$ . This is analogous to the definition of the binomial distribution in terms of the binomial expansion  $(q + p)^n$ , where  $q = 1 - p$ ,  $0 < p < 1$ , and  $n$  is positive integer. Thus the algebraic properties of NBD are similar to that of the binomial distribution. The relationship between the mean and variance of the number of individuals per sampling unit is influenced by the underlying pattern of dispersal of the population. Three basic types of patterns, their variance-to-mean relationship and suitable probability models may be defined as follows: (i) random pattern;  $\sigma^2 = \mu$ ; Poisson; (ii) uniform pattern;  $\sigma^2 < \mu$ ; positive binomial and (iii) clumped pattern;  $\sigma^2 > \mu$ ; negative binomial.

---

**E-mail:** <sup>1</sup>anwer\_khurshid@yahoo.com  
<sup>2</sup>miaqeel@kku.edu.sa

A number of variants of the NBD are used in practice under different assumptions and situations. Some of the popular forms encountered in the literature are:

$$i) P(X = n) = P(n) = \binom{n+k-1}{k-1} p^k (1-p)^n, \quad n = 0, 1, 2, \dots, \quad k > 0, p > 0$$

$$ii) P(X = n) = P(n) = \binom{n-1}{k-1} p^k (1-p)^{n-k}, \quad n = k, k+1, \dots, \quad k > 0, p > 0$$

$$iii) P(X = x) = \binom{k+x-1}{x} p^k (1-p)^x, \quad x = 0, 1, 2, \dots, \quad k > 0, p > 0$$

$$iv) P(X = x) = \binom{k+x-1}{k-1} \left(\frac{P}{Q}\right)^x \left(1 - \frac{P}{Q}\right)^k, \quad x = 0, 1, 2, \dots, \quad P > 0, Q = 1 + P.$$

v) The probability of finding  $x$  individuals in a sampling unit, that is,  $P(x)$ , where  $x = 0, 1, 2, 3, \dots, r$  individuals, is given by

$$P(x) = \left[ \frac{\mu}{\mu + k} \right]^x \left\{ \frac{(k+x-1)!}{x!(k-1)!} \right\} \left[ 1 + \frac{\mu}{k} \right]^{-k}.$$

The form (iii) will be used in this paper. For a comparison of different forms and related characteristics see Shenton and Myers (1965).

It is clear from the form of the distribution that  $k$  needs not to be integer. In the case of integer  $k$  the NBD is sometimes known as the Pascal distribution although many authors do not distinguish between Pascal and negative binomial distribution. The name 'Pascal distribution' is more often applied to the distribution shifted  $k$  units from the origin, i.e. with support  $k, k+1, k+2, \dots$ ; this is also called the binomial waiting-time distribution.

NBD has two parameters, namely, the exponent  $k$ , dispersion parameter and  $p$  which is related to the mean of the NBD as

$$\text{Mean} = \mu = \frac{k(1-p)}{p}.$$

The variance of NBD is given by

$$\text{Variance} = \mu + (\mu^2/k) = \frac{k(1-p)}{p^2}$$

which clearly shows that the variance of the NBD is always greater than the mean. The parameter  $k$  is frequently referred to as an index of clustering; the larger  $k$ , less the clustering; and conversely, the smaller the  $k$ , the greater the clustering.  $k$  has been found to be quite stable for many populations, even though the mean count may be changed considerably.

A common inferential problem in dealing with the NBD is the determination of a confidence interval for the parameter  $p$ . Several methods for approximating confidence limits were earlier reviewed by Scheaffer (1976). The methods considered by Scheaffer (1976) were: central limit theorem approach and its modification; variance-stabilizing transformation approach; and an  $\chi^2$  approach and its modification. Lui (1995) discussed confidence limits on the expected number of trials in reliability studies. The purpose of this paper is to review some methods, classical as well as bootstrapped, of obtaining a confidence interval for the parameter  $p$ . These procedures are not generally covered in standard references and introductory textbooks and would be of considerable interest to students and instructors as well as research workers in substantive fields. The authors have developed a computer program in Visual BASIC that computes confidence limits using the procedures described in this paper.

### 3. AN EXACT METHOD FOR A CONFIDENCE INTERVAL

Let  $X$  be a negative binomial random variable with parameter  $p$ . To construct an exact confidence interval for  $p$ , we need to find an interval consisting of all values of  $p$  such that the realized value of the negative count  $x$  would result in acceptance of the null hypothesis  $H_0: p = p^*$  if one were using a two-tailed test. More precisely, if we want to form a  $100(1 - \alpha)\%$  confidence interval for  $p$ , we observe the value of  $X$ , say  $x$ , and then ask, "For the given values of  $x$ , which values of  $p^*$  we may use in the null hypothesis such that a two tailed test would result in the acceptance of  $H_0$ ?" These values of  $p$  would be in our confidence interval. Since each of the test has probability not exceeding  $\alpha/2$ , the values of the lower confidence limit  $p_L$  is selected as the value of  $p^*$  that would barely result in rejection of  $H_0$  for the given value of  $x$ , or a larger value. Thus,  $p_L$  is determined such that

$$P(X \geq n | p = p_L) = \sum_{r=n}^{\infty} \binom{r+k-1}{k-1} p_L^k (1-p_L)^r \leq \frac{\alpha}{2}. \quad (1)$$

Similarly, the upper confidence limit  $p_U$  is determined such that:

$$P(X \leq n | p = p_U) = \sum_{r=0}^n \binom{r+k-1}{k-1} p_U^k (1-p_U)^r \leq \frac{\alpha}{2} \quad (2)$$

The main problem in using this method is the difficulty in computing the cumulative probability expressions such as:

$$P(X \leq n | p) = \sum_{r=0}^n \binom{r+k-1}{k-1} p^k (1-p)^r \quad (3)$$

Specially prepared tables are available for evaluating the expression in (3) (for example, Williamson and Bretherton, 1963). However, with the advent of fast, inexpensive computing and with the widespread availability of powerful personal computers and statistical software it is no longer difficult to evaluate. Computer algebra systems provide a flexible tool to solve mathematical and statistical problems like this. Examples of popular computer algebra systems are Mathematica, Maple and MathCAD. We can use the following small Mathematica code to solve the expression (3):

```
<<Statistics 'DiscreteDistributions'
k= ; n= ; p= ;
CDF[NegativeBinomialDistribution [k,p],n]
```

#### 3.1 Normal approximation using transformations

There are several transformations that approximately normalize and equalize the variance. These methods are applicable when the NBD is successfully fitted to the sample data. The choice of a suitable transformation depends upon the value of  $k$  estimated from the data. Some such transformations were considered by Elliott (1977) and are briefly described below: i) If  $2 \leq k \leq 5$  and  $\bar{x} \geq 15$ , each observed value of  $x_i$  in the sample is replaced by

$$y_i = \log\left(x_i + \frac{k}{2}\right), \quad i = 1, 2, 3, \dots, n \quad (4)$$

and the mean of transformed count  $\bar{y}$  is given by

$$\bar{y} = \frac{\sum_{i=1}^n \log\left(x_i + \frac{k}{2}\right)}{n} \quad (5)$$

where  $n$  is the number of sampling units or the number of counts and  $k$  is the negative binomial exponent. In as much as the distribution of transformed counts is approximately normally distributed with the variance  $0.1886 \text{ trigamma}(k)$  (Elliott, 1977, p. 57), an approximate  $100(1 - \alpha)\%$  confidence interval for  $p$  is given by

$$\bar{y} \pm t_{n-1,1-\alpha/2} \sqrt{\frac{0.1886 \text{ trigamma}(k)}{n}} \quad (6)$$

where  $t_{n-1,1-\alpha/2}$  is 100(1 -  $\alpha/2$ )-th percentile of the t-distribution with n-1 degrees of freedom and trigamma (k) is a mathematical derivative of the gamma function.

ii) For  $k \geq 2$  and  $\bar{x} \geq 4$ , each observed value of  $x_i$  in the sample is replaced by

$$y_i = \sinh^{-1} \sqrt{\frac{x_i + 0.375}{k - 0.75}}, \quad i = 1, 2, 3, \dots, n \quad (7)$$

where x is the observed count and  $\sinh^{-1}(x) = \log_e \left( x + \sqrt{x^2 + 1} \right)$ . Since the distribution of transformed count is approximately normally distributed with variance 0.25 trigamma(k) (Elliott, 1977, p. 57), an approximate 100(1 -  $\alpha$ )% confidence interval for p is given by

$$\bar{y} \pm t_{n-1,1-\alpha/2} \sqrt{\frac{0.25 \text{ trigamma}(k)}{n}} \quad (8)$$

These confidence intervals must be transformed back to the original scale of observations by inverting the original transformation.

### 3.2. Normal approximation using Box-Cox transformation

Another approach for an approximate confidence interval for p is to use a generalized transformation like Box and Cox (1964) to normalize the data and then use the conventional methods based on the t-distribution. Each observed value of  $x_i$  in the sample is replaced by

$$y_i = \frac{x_i^k - 1}{k}, \quad k \neq 0 \quad (9)$$

$$= \log_e x_i, \quad k = 0, \quad i = 1, 2, 3, \dots, n$$

and the mean of transformed counts  $\bar{y}$  is given by

$$\bar{y} = \begin{cases} \frac{\sum_{i=1}^n \left( \frac{x_i^k - 1}{k} \right)}{n}, & k \neq 0 \\ \frac{\sum_{i=1}^n \log x_i}{n}, & k = 0 \end{cases} \quad (10)$$

To use the Box-Cox transformation, one selects the value of k that minimizes the log-likelihood function (L) given by

$$L = \frac{(n-1)}{2} \log_e S_t^2 + (k-1) \frac{(n-1)}{n} \sum_{i=1}^n (\log_e x_i)$$

where  $S_t^2$  is variance of transformed scores using (9) and k is the provisional estimate of the power transformation parameter.

This must be solved iteratively to find the value of k that maximizes L. Since this is tedious; it is usually done by computer. In as much as the distribution of transformed counts is approximately normal, an approximate 100(1 -  $\alpha$ )% confidence interval for p is

$$\bar{y} \pm t_{n-1, 1-\alpha/2} \sqrt{\frac{S_t^2}{n}}. \quad (11)$$

### 3.3. Large sample normal approximation without transformation

For a sufficiently large value of  $n$  one can apply the central limit theorem and assume that the sample mean  $\bar{X}$  is approximately normally distributed with mean  $kp$  and variance  $\frac{kp(1-p)}{n}$ , and an approximate  $100(1 - \alpha)\%$  confidence interval for  $p$  is given by:

$$\hat{p} \pm z_{1-\alpha/2} \sqrt{\text{Var}(\hat{p})} \quad (12)$$

where  $z_{1-\alpha/2}$  is the  $100(1 - \alpha/2)$  - th percentile of the standard normal distribution.

### 3.4. Normal approximation using variance stabilizing transformation

The formulas  $E(X_i) = k \frac{(1-p)}{p}$  and  $\text{Var}(X_i) = k \frac{(1-p)}{p^2}$  suggest that the transformation

$$y_i = \sinh^{-1} \sqrt{\frac{x_i}{k}}, \quad (13)$$

where  $y_i$  is approximately normally distributed with some mean,  $\mu$ , and variance  $\frac{1}{4k}$ . It therefore follows that

$S_n = \sum_{i=1}^n x_i$  is negative binomial with parameter  $nk$  and  $p$ . Thus, the transformed random variable  $Y = \sinh^{-1} \sqrt{\frac{S_n}{nk}}$  is approximately normally distributed with mean  $\mu$  and variance  $\frac{1}{4nk}$  for large values of mean. Considering the normality of  $Y$ , it can be observed that

$$\begin{aligned} E[\sinh Y]^2 &= E[e^{2Y} + e^{-2Y} - 2] \\ &= \frac{1}{4} \left[ e^{2\sigma_y^2} 2 \cosh 2\mu - 2 \right] \\ &= \frac{1}{2} \left[ e^{2\sigma_y^2} \cosh 2\mu - 1 \right] \end{aligned} \quad (14)$$

Thus, a confidence interval for  $p$  can be obtained by first finding a confidence interval for  $\mu$  and then transforming it according to equation (14). The confidence interval for  $\mu$  is given by:

$$(\mu_L, \mu_U) = y \pm z_{1-\alpha/2} 2\sigma_Y \quad (15)$$

and the transformed interval for  $p$  is

$$\left( \frac{1}{2} \left[ e^{2\sigma_y^2} \cosh 2\mu_L - 1 \right], \frac{1}{2} \left[ e^{2\sigma_y^2} \cosh 2\mu_U - 1 \right] \right) \quad (16)$$

for nonnegative  $\mu_L$  and known  $\sigma_y^2$ .

### 3.5. Use of chi-square distribution

Using the moment generating function for  $X_i$ ,  $[(1+p) - pe^t]^{-k}$ , it follows that the moment generating function of  $\frac{2(X_i + k)}{p+1}$  is

$$\left[ (1+p) - pe^{\frac{2t}{p+1}} \right]^{-k} e^{\frac{2kt}{p+1}} \quad (17)$$

and, therefore, that  $\frac{2n(\bar{X} + k)}{p+1}$  has approximately a  $\chi^2_{2nk}$  distribution for large  $k$ . A confidence interval for  $p$  is

$$\left[ \frac{2n(\bar{X} + k)}{\chi^2_{2nk, 1-\alpha/2}} - 1, \frac{2n(\bar{X} + k)}{\chi^2_{2nk, \alpha/2}} - 1 \right] \quad (18)$$

### 3.6. An improved large sample approximation

For large sample approximation in section 3.3, we have

$$P \left[ \left( \frac{\bar{X} - kp}{\sqrt{\frac{kp(1+p)}{n}}} \right)^2 \leq z^2_{1-\alpha} \right] \approx 1 - \alpha \quad (19)$$

and an approximate confidence interval for  $p$  in this case is determined by

$$\frac{\frac{2\bar{X}}{k} + \frac{z^2_{1-\alpha/2}}{nk} \pm \sqrt{\left( \frac{2\bar{X}}{k} + \frac{z^2_{1-\alpha/2}}{nk} \right)^2 - 4 \left( 1 - \frac{z^2_{1-\alpha/2}}{nk} \right) \left( \frac{\bar{X}}{k} \right)^2}}{2 \left( 1 - \frac{z^2_{1-\alpha/2}}{nk} \right)} \quad (20)$$

provided that  $\left( 1 - \frac{z^2_{1-\alpha/2}}{nk} \right) > 0$ .

The classical interval estimation methods use a fully parametric model to express the uncertainty of the corresponding point estimator. That is, all elements of the assumed statistical model are required to derive the confidence interval. However, it is often not possible to make a distributional assumption, or a distributional assumption can be made but derivation of a confidence interval is mathematically intractable. The bootstrap (Efron, 1982; Efron and Tibshirani, 1993; Mooney and Duval, 1993) provides a solution in both the instances.

In recent years the bootstrap theory is gaining support as an alternative to classical parametric inference methods. The principal goal of bootstrap theory is to produce good confidence intervals automatically. Here the term good means that the bootstrap intervals should closely match exact confidence intervals in those special situations where statistical theory yields an exact answer, and should give dependably accurate coverage probabilities in all situations. It should be noted that every thing in the bootstrap procedure depends on the original sample values. A different set of sample values yields a different set of estimates.

Bootstrap methods, justified by mathematical arguments, are means of assessing the precision of estimates, and require programming ability, rather than statistical expertise, for their implementation. Taffe and Garnham (1996), for example, have given more details about practical applications, including MINITAB code.

## 4. BOOTSTRAP PROCEDURES FOR CONFIDENCE INTERVALS

There are different bootstrap procedures for finding the confidence intervals; each procedure is different from other due to its exactness and its complexity level. When they are appropriate, bootstrap confidence intervals are second order accurate (Singh, 1981; Bickel and Freedman, 1981) which suggests that the bootstrap could provide good approximate confidence intervals, better than the usual 'standard intervals'. For details on bootstrap procedures for confidence intervals see DiCiccio and Efron (1996) and Martinez and Louzada-Neto (2001). In this section we present some bootstrap procedures for confidence intervals.

### 4.1. The standard bootstrap confidence limits

With the standard bootstrap confidence interval is estimated by the standard deviation of estimates of a parameter  $\theta$  that are found by bootstrap resampling of the values in the original sample of data. The interval is then

$$\text{Estimate} \pm z_{\alpha/2} (\text{Bootstrap standard deviation})$$

### 4.2. The first percentile method (Efron, 1979)

Bootstrap resampling of the original data is used to generate the bootstrap parameter of the interest. The  $100(1 - \alpha)\%$  confidence interval for the true value of the parameter is then given by the two values that encompass the central  $100(1 - \alpha)\%$  of this distribution. For example, a 95% confidence interval is given by the value that exceeds 2.5% and 97.5% of the generated distribution.

### 4.3. The second percentile method (Hall, 1992)

Bootstrap resampling is used to generate a distribution of estimates  $\hat{\theta}_B$  for a parameter  $\theta$  of interest. The bootstrap distribution of difference between the bootstrap estimates and the estimate of  $\theta$  in original sample  $\varepsilon_B = \hat{\theta}_B - \hat{\theta}$  is then assumed to approximate the distribution of errors for  $\hat{\theta}$  itself. On this basis the bootstrap distribution of  $\varepsilon_B$  is used to find limits  $\varepsilon_L$  and  $\varepsilon_H$  for the sampling error such that  $100(1 - \alpha)\%$  of errors are between these limits.  $100(1 - \alpha)\%$  confidence limits for  $\theta$  are then  $\hat{\theta} - \varepsilon_H < \theta < \hat{\theta} - \varepsilon_L$ . For example, to obtain 95% confidence interval  $\varepsilon_L$  and  $\varepsilon_H$  should be chosen as the two values that define the central 95% part of the distribution of the bootstrap sampling errors  $\varepsilon_B$ .

### 4.4. Bias-corrected percentile $100(1 - \alpha)\%$ confidence limits

The following steps are needed to obtain bias-corrected percentile confidence intervals:

Generate values  $\hat{\theta}_B$  from the bootstrap distribution for estimates of the parameter  $\theta$  of interest. Find the proportion of times  $p$  that  $\hat{\theta}_B$  exceeds  $\hat{\theta}$ , the estimate of  $\theta$  from the original sample. Hence, calculate  $z_0$ , the value from the standard normal distribution that is exceeded with probability  $p$ . (This is  $z_0 = 0$  if  $p = 0.5$ )

Calculate  $\phi(2z_0 - z_{\alpha/2})$  and  $\phi(2z_0 + z_{\alpha/2})$  which are the proportions of the standard normal distribution that is less than  $2z_0 - z_{\alpha/2}$  and  $2z_0 + z_{\alpha/2}$  respectively where  $z_{\alpha/2}$  is the value of that is exceeded with probability  $\alpha/2$  for the standard normal distribution.

The lower and upper confidence limits for  $\hat{\theta}$  are the values that just exceed a proportion  $\phi(2z_0 - z_{\alpha/2})$  and  $\phi(2z_0 + z_{\alpha/2})$  of all values in the bootstrap distribution of estimates  $\hat{\theta}_B$  respectively.

### 4.5. Bootstrap-t $100(1 - \alpha)\%$ Confidence limits

As the name suggests this method uses the t distribution. The bootstrap-t confidence limits are found by performing the following steps:

Approximate  $t_{\alpha/2}$  and  $t_{1-\alpha/2}$  using the bootstrap t-distribution, i.e. by finding the values that satisfy the two equations

$$\Pr[(\hat{\theta}_B - \hat{\theta})/\widehat{SE}(\hat{\theta}_B) > t_{\alpha/2}] = \alpha/2 \text{ and } \Pr[(\hat{\theta}_B - \hat{\theta})/\widehat{SE}(\hat{\theta}_B) > t_{1-\alpha/2}] = 1 - \alpha/2$$

for the generated bootstrap estimates. For example, for a 95% confidence interval the two values of t will encompass the central 95% of the bootstrap-t distribution.

The confidence interval is given by

$$\hat{\theta} - t_{\alpha/2} \widehat{SE}(\hat{\theta}) < \theta < \hat{\theta} - t_{1-\alpha/2} \widehat{SE}(\hat{\theta}).$$

## 5. AN ILLUSTRATIVE EXAMPLE

A shovel sampler was used to take a large sample of 80 sampling units from the bottom of a stony stream, and fresh water shrimps were counted in each of sampling unit (0.05 sq. m). The chi-square goodness of fit test was applied on the observed counts, which justify that the negative binomial distribution provides good approximate for the data. The data in frequency distribution form is. (Elliot, 1977, p. 53)

X	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	Total
Obs.	3	7	9	12	10	6	7	6	5	4	3	2	2	1	1	1	1	80

From the table, we obtain the mean as 5.3125 and the variance 13.534. The method of moments estimator of p and P are 0.3925 and 1.5157 respectively. The maximum likelihood estimates of p and P are 0.3871 and 1.5832 respectively.

### 5.1. Classical methods for confidence intervals

*Large sample normal approximation without transformation*

Using equation (12), 95% confidence interval of p is given as

$$\hat{p} \pm (1.96)\sqrt{\text{Var}(\hat{p})} = (0.3527, 0.4233)$$

*Normal approximation using variance stabilizing transformation*

Using equation (15), confidence interval for  $\mu$  is:

$$1.05275 \pm (1.96)(0.0305) = (1.35912, 1.84504)$$

and using equation (16), the transformed interval for p is (0.3515, 0.4239).

*Use of chi square distribution*

The tabulated values of chi square were found as  $\chi_{640,0.975}^2 = 571.7912$  and  $\chi_{640,0.025}^2 = 711.9964$  at approximate k value to 4. Then using equation (18) the confidence interval of p was (0.3838, 0.4777).

*An improved large sample approximation*

An approximate confidence interval for p using equation (19) gives (0.3500, 0.4225).

### 5.2. Bootstrap confidence intervals

For bootstrap confidence intervals 10000 bootstrap samples were obtained from the original sample of size n = 80. With replacement sampling was done to draw each sample of size n = 80. The mean and variance of the original sample were 5.3125 and 13.534 respectively. The estimate of k was 3.3555. While

from the original sample the moment estimator of  $p$  and the maximum likelihood estimate of  $p$  were 0.3975 and 0.3871 respectively.

Confidence intervals for the parameter  $p$  were obtained from these 10000 bootstrap samples by the following procedures.

*Standard bootstrap C.I.:*

The bootstrap standard deviation from all 10000 bootstrap confidence intervals was 0.0627, hence the confidence limits can be found as

$$0.3871 \pm (1.96)(0.0627) = (0.2642, 0.5099)$$

The mean of the 10000 bootstrap estimates of  $p$  was 0.4002. The bias is therefore estimated to be  $0.4002 - 0.3871 = 0.0131$ . This suggests that the original sample estimate of 0.3871 is also low by this amount as well, so the bias-corrected estimate of  $p$  of the original population sampled is  $0.3871 - 0.0131 = 0.3740$ . Hence the confidence interval based on this estimate of  $p$  is

$$0.3740 \pm (1.96)(0.0627) = (0.2511, 0.4968)$$

*1<sup>st</sup> percentile method:*

First the bootstrap distribution of the parameter  $p$  estimated from the 10000 samples was constructed by arranging them into an array of ascending order. Now the 95% confidence limits were found by selecting two values that define the 95% proportion of the generated distribution. The value that exceeds 2.5% and 97.5% of the generated distribution were taken as the lower and the upper confidence limits respectively. So the confidence limits were (0.3137, 0.5804).

*2<sup>nd</sup> percentile method:*

The bootstrap distribution of the parameter  $p$  estimated from the 10000 samples was constructed by arranging them into an array of ascending order. The bootstrap distribution of the difference between the bootstrap estimate and the estimate of  $p$  in the original sample  $\varepsilon_B = \hat{p}_B - \hat{p}$  is then constructed. Now the 95% confidence limits of the sampling error were found by selecting two values  $\varepsilon_L$  and  $\varepsilon_H$  that define the 95% proportion of the generated distribution. The value that exceeds 2.5% and 97.5% of the generated distribution were taken as the lower  $\varepsilon_L$  and the upper confidence  $\varepsilon_H$  limits respectively. So the confidence limits for the  $p$  were found as

$$0.3871 - 0.1545 < p < 0.3871 - (-0.0908) = (0.2326, 0.4779)$$

*Bias corrected percentile method:*

In this method from the generated distribution of  $p$  the proportion of times that the bootstrap estimate of  $p$  exceeds the estimate of  $p$  from the original sample was 0.5454. Hence  $z_0 = -0.1764$  and the proportion of the standard normal distribution that is less than  $2z_0 - z_{\alpha/2}$  is 0.0617. Now the lower limit of  $p$  is 0.0617 quantile value of the generated distribution which just exceeds a proportion 0.0104 of all values in the bootstrap distribution of estimate of  $p$ . Similarly the proportion of the standard normal distribution that is less than  $2z_0 + z_{\alpha/2}$  is 0.9913. Now the upper limit of  $p$  is 0.9913 quantile value of the generated distribution, which just exceeds a proportion 0.9913 of all values in the bootstrap distribution of estimate of  $p$ .

*Bootstrap-t confidence limit:*

First we calculate  $t$  statistics for each bootstrap sample by the formula  $T_B = (\hat{p}_B - \hat{p}) / \hat{SE}(\hat{p}_B)$  where  $\hat{p} = 0.3871$ . Then the bootstrap  $t$ -distribution of the generated values of  $t$  was constructed. For the 95% confidence interval the two values of  $t$  that encompass the central 95% of the bootstrap- $t$  distribution are  $t_{\alpha/2} = 1.4717$ ,  $t_{1-\alpha/2} = -2.3709$ . Hence the confidence interval for  $p$  is given by

$$\begin{aligned} &0.3871 - (1.4717)(0.0627) < p < 0.3871 - (-2.3709)(0.0627) \\ &= (0.2949, 0.5357) \end{aligned}$$

According to one referee some newer procedures such as the ‘wild bootstrap’ can also be used for constructing confidence intervals. Wild or external bootstrap, first proposed by Wu (1986), is one of the most promising bootstrap resampling scheme (Helmers and Wegkam, 1995). Using two stage wild bootstrapping we found: mean and variance for  $p$  are 0.397047 and 0.153179; lower and upper confidence limits for  $p$  were 0.367024 and 0.42707 respectively and hence the length 0.060046.

### 5. RESULTS AND COMMENTS

The purpose of this paper has been to describe some methods of constructing confidence intervals for the parameter  $p$  of NBD. Various methods, classical as well as bootstrap, have been described with examples illustrating the application of each procedure. In order to assist the reader in assessing the options of the methods for construction of confidence intervals, the results of the example considered in the article are summarized below:

**Confidence interval results for the parameter  $p$  of NB distribution based on various methods**

Method of construction		95% C.I.		Length
		Lower limit	Upper limit	
<b>CLASSICAL</b>	Large sample normal approximation without transformation	0.3527	0.4233	0.0706
	Normal approximation using variance stabilizing transformation	0.3515	0.4239	0.0724
	Use of chi-square distribution	0.3838	0.4777	0.0939
	Improved large sample approximation	0.3500	0.4225	0.0725
<b>BOOTSTRAP</b>	Standard bootstrap C.I.	0.2642	0.5099	0.2457
	Standard bootstrap C.I on adjusted estimate	0.2511	0.4968	0.2457
	1 <sup>st</sup> percentile method	0.3137	0.5804	0.2667
	2 <sup>nd</sup> percentile method	0.2326	0.4779	0.2453
	Bias corrected percentile method	0.2963	0.5416	0.2453
	Bootstrap-t confidence limit	0.2949	0.5357	0.2408
	Wild bootstrap C. I.	0.3670	0.4270	0.0600

The question “which is the best method?” has no simple answer. The numerical results show that wild bootstrap produced the shortest confidence intervals among the classical as well as bootstrap methods. Among the classical procedures, narrowest confidence interval is given by large sample normal approximation without transformation. While bootstrap confidence intervals differ by a quantity  $o(n^{-1})$ , their coverage probabilities, except the wild bootstrap, are the same to  $o(n^{-1})$  and agree to that order, with coverage probability.

### ACKNOWLEDGMENTS

This study has benefited enormously from the critical comments especially about the wild bootstrap of one anonymous referee. Authors are also thankful to Professor Ghulam Hussain for pointing and sorting some inconsistencies in the computation.

### REFERENCES

ANDERSON, F.S. (1965): “The negative binomial distribution and the sampling of insect population”, **Proceedings of XII International Conference on Entomology**, 395-402.

ANSCOMBE, F.J. (1950): “Sampling theory of negative binomial and logarithmic series distributions”, **Biometrika**, 37, 358-382.

ARBOUS, A.G. and J.E. KERRICH (1951): “Accident statistics and the concept of accident proneness”, **Biometrics**, 7, 340-342.

BICKEL, P. and D. FREEDMAN (1981): “Some asymptotic theory for the bootstrap”, **Annals of Statistics**, 9, 1196-1227.

- BIGGERI, A. (1998): "Negative binomial distribution", In: **Encyclopedia of Biostatistics**, 4, 2962-2967. (Eds. P. Armitage and T. Colton). John Wiley: New York.
- BINNS, M. R. (1986): "Behavioral dynamics and the negative binomial distribution", **Oikos**, 47, 315-318.
- BLISS, C. I. and R.A. FISHER (1953): "Fitting the negative binomial distribution to biological data and note on the efficient fitting of negative binomial", **Biometrics**, 9, 176-200.
- BOOKSTEIN, A. (1997): "Infometric distributions, Part III: Ambiguity and randomness", **Journal of the American Society for Information Sciences**, 48, 2-10.
- BOSWELL, M. and G.P. PATIL (1970): "Change mechanisms generating the negative binomial distributions", In: **Random Counts in Scientific Work, Vol. 1: Random Counts in Models and Structures**, 3-22. (Ed. G. P. Patil). Pennsylvania State University Press: University Park, Pennsylvania.
- BOX, G. E. P. and D.R. COX (1964): "An analysis of transformations", **Journal of Royal Statistical Society**, Series B, 26, 211-252.
- BYERS, A.L.; H. ALLORE; T.M. GILL and P.N. PEDUZZI (2003): "Application of negative binomial modeling for discrete outcomes: a case study in aging research", **Journal of Clinical Epidemiology**, 56, 559-564.
- DICICCIO, T. and B. EFRON (1996): "Bootstrap confidence intervals" (with discussion), **Statistical Science**, 11, 189-228.
- EFRON, B. (1979): "Bootstrap methods: Another look at the jackknife", **Annals of Statistics**, 7, 1-26.
- \_\_\_\_\_ (1982): **The Jackknife, the Bootstrap and other Resampling Plans**, Society for Industrial and Applied Mathematics: Philadelphia.
- EFRON, B. and R. TIBSHIRANI (1993): **An Introduction to the Bootstrap**, Chapman & Hall: New York.
- EKE, A.C.; B.O. EKPENYUNG and G.I. OGBAN (2001): "Exploring the potentials of the negative binomial distribution", **Global Journal of Pure and Applied Sciences**, 7, 749-754.
- ELLIOT, J.M. (1977): **Some Models for the Statistical Analysis of Samples of Benthic Invertebrates**, Fresh Water Biological Association: Ambleside, England.
- FURRY, W.H. (1937): "On fluctuation phenomena in the passage of energy electrons through lead", **Physical Review**, 52, 569-581.
- GREENWOOD, M. and G.U. YULE (1920): "An inquiry into the nature of frequency distributions representative of multiple happening with particular reference to the occurrence of multiple attacks of disease or repeated accidents", **Journal of the Royal Statistical Society**, Series A, 83, 255-279.
- HALL, P. (1992): **The Bootstrap and Edgeworth Expansion**, Springer-Verlag: New York.
- HELMERS, R. and H.M. WEGKAMP (1995): **Wild bootstrapping in finite population with auxiliary information**, CW Report, BS-R95xx, Amsterdam.
- JOHNSON, N.L.; S. KOTZ and A. KEMP (1992): **Discrete Distributions**, Second edition. John Wiley: New York.
- KENDALL, D.G. (1949): "Stochastic processes and population growth", **Journal of the Royal Statistical Society**, Series B, 11, 230-282.
- LUI, K.J. (1995): "Confidence limits for the population prevalence rate based on the negative binomial distribution", **Statistics in Medicine**, 14:1471-1477.

- MARTÍN, D.C. and S.K. KATTI (1965): "Fitting of some contagious distribution to some available data by the maximum likelihood method", **Biometrics**, 21, 34-48.
- MARTÍNEZ, E.Z. and F. LOUZADA-NETO (2001): "Bootstrap confidence interval estimation", **Revista-de-Matematica-e-Estatistica**, 19, 217-251. (In Portuguese)
- MOONEY, C.Z. and R.D. DUVAL (1993): **Bootstrapping: A Nonparametric Approach to Statistical Inference**, Sage: California.
- SAKAMOTO, C.M. (1972): "Application of the Poisson and Negative Binomial Models to Thunderstorm and Hail Days Probabilities in Nevada", **Monthly Weather Review**, 101, 350-355.
- SCHEAFFER, R.L. (1976): "A note on approximate confidence limits for negative binomial Model", **Communications in Statistics: Theory and Methods**, 5, 149-158.
- SHENTON, L.R. and R.H. MYERS (1965): "Comments on Estimators for the Negative Binomial Distribution", **Proceedings of XII International Conference on Entomology**, 445-458.
- SICHEL, H.S. (1951): "The estimation of the parameters of a negative binomial distribution with special reference to psychological data", **Psychometrika**, 16, 107-127.
- SINGH, K. (1981): "On asymptotic accuracy of Efron's bootstrap", **Annals of Statistics**, 9, 1187-1195.
- TAFFE, J. and N. GARNHAM (1996): "Resampling: the bootstrap and MINITAB", **Teaching Statistics**, 18, 24-25.
- WEBER, D.C. (1971): "Accident rate potential: An application of multiple regression analysis of a Poisson process", **Journal of the American Statistical Association**, 66, 285-288.
- WHITE, G.C. and R.E. BENNETLE (1996): "Analysis of frequency count data using the negative binomial distribution", **Ecology**, 77, 2549-2557.
- WILLIAMSON, E. and M.H. BRETHERTON (1963): **Tables of the Negative Binomial Probability Distribution**, John Wiley: New York.
- WILSON, L. T. and P.M. ROOM (1983): "Clumping patterns of fruit and arthropods in cotton, with applications for binomial sampling", **Environmental Entomology**, 12, 50-54.
- WU, C.F.J. (1986): "Jackknife, bootstrap and other resampling methods in regression analysis" (with discussion), **Annals of Statistics**, 14, 1261-1350.