Statistical Papers
https://doi.org/10.1007/s00362-021-01255-y

REGULAR ARTICLE

Confidence intervals, prediction intervals and tolerance intervals for negative binomial distributions

Bao-Anh Dang[1] · K. Krishnamoorthy[1]

Received: 10 February 2021 / Revised: 13 May 2021 / Accepted: 16 July 2021

Abstract

The problems of constructing confidence intervals (CIs) for a proportion, prediction intervals (PIs) for a future sample size in a negative binomial sampling to observe a specified number of successes and tolerance intervals (TIs) for negative binomial distributions are considered. For interval estimating the success probability, we propose CIs based on the fiducial approach and the score method, evaluate them and compare them with available CIs with respect to coverage probability and precision. We propose PIs based on the fiducial approach and joint sampling approach, and compare them with the exact and other approximate PIs. We also propose TIs on the basis of our new CIs and evaluate them with respect to coverage probability and expected width. All three statistical intervals are illustrated using two examples with real data.

Keywords Beta-negative binomial · Coverage probability · Equal-tailed tolerance interval · Fiducial approach · Highest probability mass function · Inverse sampling · Precision

1 Introduction

Binomial is one of the most popular discrete probability distributions that has been received great attention in the literature. Statistical intervals for one- and two-sample problems involving binomial models are widely available. In comparison with the binomial, inferential results on negative binomial distributions are very limited. In binomial sampling, a fixed number of sampling units are drawn from an infinite population and the number of units with an attribute (success) of interest is counted whereas in negative binomial sampling (also known as inverse sampling) units are drawn from an infinite population until a prespecified number of successes is observed. Thus, in

✉ K. Krishnamoorthy
krishna@louisiana.edu

[1] Department of Mathematics, University of Louisiana at Lafayette, Lafayette, LA 70504, USA

Published online: 13 August 2021

binomial sampling the sample size is fixed and the number of successes is a random variable whereas in negative binomial sampling the number of success is fixed while the sample size or the number of failures is a random variable. Negative binomial sampling is commonly used in situations where one encounters events that occur with small probability. Haldane (1945) has noted that in epidemiological studies on rare disease, negative binomial sampling design may be used to ensure that a reasonable number of cases are observed. Tian et al. (2009) have noted the applications of negative binomial distributions in biological and medical studies. Kikuchi (1987) has used negative binomial distributions in case-control study involving a rare exposure of maternal congenital heart disease (see Example 1), and Madden et al. (1996) have noted the applications in botanical study of plant diseases. In mail-in survey sampling, non-responses are quite common, and the initial sample size needs to be determined in order to get a specified number of final responses. Recently, Young (2014) has shown application of negative binomial tolerance intervals (TIs) in such survey sampling; see Example 2.

The problem of finding statistical intervals for negative binomial distributions has received only little attention. George and Elston (1993) have considered the problem of finding confidence intervals (CIs) for proportions based on inverse sampling until the occurrence of the first event. Lui (1995) has noted that the confidence interval given in Clemans (1959), which was calculated on the basis of the first event, may be too wide for general utility. Tian et al. (2009) have proposed some approximate confidence intervals (CIs) for the success probability p of a negative binomial distribution. The comparison studies by these authors have indicated that the available exact CI is too conservative and so they have proposed some approximate confidence intervals which are less conservative. Even though the exact CIs for binomial, Poisson and negative binomial distributions are optimal among strictly nested intervals (Thulin and Zwanzig 2017), it is well-known that the exact CIs are often too conservative and unnecessarily wide. Alternative simple closed-form approximate CIs based on the score method and Bayesian methods are proposed for binomial and Poisson distributions. However, such CIs are not available for negative binomial distributions.

Another statistical interval that is commonly used in applications is the prediction interval (PI). The prediction problem that we will address concerns two independent negative binomial distributions with the same "success probability" p, but possibly different target numbers of successes. Given that n independent Bernoulli trials are needed to observe r successes, we like to predict the number of trials required in another negative binomial sampling to observe s successes with confidence $1-\alpha$. In particular, we like to find a prediction interval $[L(X; r, s, \alpha),\ U(X; r, s, \alpha)]$ so that

$$P_{X,Y}\left(L(X; r, s, \alpha) \le Y \le U(X; r, s, \alpha)\right) \ge 1-\alpha. \qquad (1)$$

In the above, the random variable X represents the number of failures before the rth success (so that $n = X + r$) and it has a negative binomial distribution with success probability p and the number of successes r, say, NBin(r, p) and Y has a NBin(s, p) distribution independently of X. Note that by adding s to the PI for Y, we find the PI for the number of trials needed to observe s successes in a future negative binomial sampling. Although many authors (see Knüsel 1994; Dunsmore 1976; Krishnamoorthy and Peng 2011) have addressed the problem of finding PIs for binomial and Poisson distributions, to the best of our knowledge, no PI is available for a negative binomial distribution.

The problem of constructing TIs for a discrete distribution has received some attention in the literature. TIs for a discrete distribution are used to assess the magnitude of discrete quality characteristics of a product, for example, the number of defective components in a system. Methods for finding TIs for the binomial and Poisson models are proposed in Hahn and Chandra (1981), Hahn and Meeker (1991), and Krishnamoorthy et al. (2011). These authors have provided exact and some approximate methods of obtaining TIs. Wang and Tsung (2009) provided an example where it is desired to find a TI for a binomial distribution to assess the number of defective chips in a wafer. Young (2014) has proposed some TIs for negative binomial distributions showing applications to survey sampling (see Example 2).

In this article, we address the following problems. (i) Construction of CIs for the success probability, (ii) finding PIs for $Y \sim \mathrm{NBin}(s, p)$ based on $X \sim \mathrm{NBin}(r, p)$, and (iii) construction of equal-tailed tolerance intervals. Since we propose the fiducial approach for the problems (i) and (ii), we first describe fiducial distributions for the success probability p in a negative binomial sampling. In Sect. 3, we propose fiducial and score CIs for p and compare them with an available large sample CI in terms of coverage probabilities and precisions. In Sect. 4, we propose a few CIs for the expected number of trials required to observe a fixed number of successes in a future negative binomial experiment. In Sect. 5, we describe an exact PI, and propose a fiducial PI, highest probability mass (HPM) prediction interval and a PI based on a joint sampling approach. All these PIs are evaluated with respect to coverage probabilities and expected widths. The problem of constructing equal-tailed TIs is addressed in Sect. 6. In Sect. 7, two examples with real data are used to illustrate the methods, and some concluding remarks are given in Sect. 8.

2 Fiducial distribution for *p*

The negative binomial probability mass function (PMF) is given by

$$P(X = x|r, p) = \binom{r + x - 1}{x} p^r (1 - p)^x, \quad x = 0, 1, 2, \ldots \tag{2}$$

where the random variable X represents the number of failures until the occurrence of the rth success in a sequence of independent Bernoulli trials each with success probability p. Let us denote the negative binomial distribution by $\mathrm{NBin}(r, p)$.

A fiducial distribution for a parameter can be obtained by inverting a hypothesis test as suggested by Fisher (1935) or by deducing from a random number generating method (Hannig 2009). As both methods produce similar fiducial distributions, we shall follow Hannig's approach. To identify the data generating mechanism in a negative binomial distribution, we note that x^* is a pseudo random number from the $\mathrm{NBin}(r, p)$ distribution if

$$P(X \leq x^* - 1|r, p) < U \leq P(X \leq x^*|r, p),$$

where U is a uniform(0,1) random variable (e.g., see Casella and Berger 2001, p. 249).

Let x be an observed value of $X \sim \text{NBin}(r, p)$. For a given x, the fiducial distribution of p is implicitly determined by

$$P(X \leq x - 1|r, p) < U \leq P(X \leq x|r, p), \tag{3}$$

where U has a uniform(0, 1) distribution. Let $B_{a,b}$ denote the beta random variable with shape parameters a and b. Using the result (see Patil 1960) that $P(X \leq x|r, p) = P(B_{r,x+1} \leq p)$, where $X \sim \text{NBin}(r, p)$ in (3), we see that the fiducial distribution of p is implicitly determined by

$$P(B_{r,x} \leq p) < U \leq P(B_{r,x+1} \leq p), \tag{4}$$

or equivalently,

$$B_{r,x+1;U} \leq p < B_{r,x;U}, \tag{5}$$

where $B_{a,b;q}$ denotes the qth quantile of a beta(a, b) distribution. Notice that if $u_1, \ldots, u_N$ are random numbers from uniform(0, 1) distribution, then $B_{a,b;u_1}, \ldots, B_{a,b;u_N}$ are random numbers from beta(a, b) distribution. Thus, a fiducial distribution of p lies between beta$(r, x + 1)$ and beta(r, x) distributions.

For a given (x, r), random samples $\widehat{p}_{u_1}, \ldots, \widehat{p}_{u_N}$ from the fiducial distribution of p are determined by

$$B_{r,x+1;u_i} < \widehat{p}_{u_i} \leq B_{r,x;u_i}, \quad i = 1, \ldots, N.$$

Like in the binomial case (see Krishnamoorthy and Lee 2010), a random quantity that is "stochastically between" $B_{r,x+1}$ and $B_{r,x}$ can be used as a single fiducial variable for p. A simple choice is

$$B_{r,x+.5}, \tag{6}$$

which stochastically lies between $B_{r,x+1}$ and $B_{r,x}$. That is, for any given (r, x),

$$B_{r,x+1;U} \leq B_{r,x+.5;U} \leq B_{r,x;U} \qquad \text{for all } U \in (0, 1). \tag{7}$$

3 Confidence intervals

3.1 Fiducial confidence interval

For a given confidence coefficient $1 - \alpha$, the lower and upper $\alpha/2$ quantiles of $B_{r,x+.5}$ form a $1 - \alpha$ generalized fiducial CI for p. That is, the fiducial CI is given by

$$\left(B_{r,x+.5;\alpha/2},\ B_{r,x+.5;1-\alpha/2}\right). \tag{8}$$

3.2 Exact confidence interval

It follows from (5) that, for a given x,

$$(p_L, p_U) = \left(B_{r,x+1;\alpha/2},\ B_{r,x;1-\alpha/2}\right) \tag{9}$$

is also a $1-\alpha$ fiducial CI for p. Note that the above interval is an observed value of the random interval

$$\left(B_{r,X+1;\alpha/2},\ B_{r,X;1-\alpha/2}\right), \tag{10}$$

where $X \sim \text{NBin}(r, p)$. This random CI is obtained by using the "pivoting a CDF" approach and so it is exact in the frequentist sense; see Theorem 9.2.14 of Casella and Berger (2001) and the paper by Lui (1995). In particular, the endpoints of the exact CI $(p_L,\ p_U)$ are the solutions of

$$F_X(x|r, p_L) = \frac{\alpha}{2} \quad \text{and} \quad \bar{F}_X(x|r, p_U) = \frac{\alpha}{2}, \tag{11}$$

where x is an observed value of $X \sim \text{NBin}(r, p)$, $F_X(x|r, p) = P(X \le x|r, p)$ and $\bar{F}_X(x|r, p) = P(X \ge x|r, p)$. The solutions of the above equations are the endpoints of the CI (9), which can be verified using the distributional results that $P(X \le x|r, p) = P(B_{r,x+1} \le p)$ and $P(X \ge x|r, p) = P(B_{r,x} \ge p)$.

Remark 1 Tian et al. (2009) have found an approximation, say, $\widehat{F}_X(x|r, p)$, to the distribution function $F_X(x|r, p)$ using the saddle point approximation, and then determined p_L and p_U as solutions of $\widehat{F}_X(x|r, p_L) = \frac{\alpha}{2}$ and $\widehat{\bar{F}}_X(x|r, p_U) = \frac{\alpha}{2}$, respectively. This approximate CI is not in closed-form and can be obtained only numerically. However, the solutions of the exact method determined by equations in (11) are in closed-form given in (10), and simple to compute. So we will not consider this saddle point approximate CI for further studies.

3.3 Score confidence interval

Let $\eta = (1-p)/p$. For $X \sim \text{NBin}(r, p)$, $E(X) = r\eta$. The score CI for p can be obtained from the one for η. Noting that $\widehat{\eta} = X/r$ is an unbiased estimate of η and $\text{Var}(\widehat{\eta}) = \eta/(rp)$, we consider the quantity $\sqrt{r}(\widehat{\eta}-\eta)/\sqrt{\eta/p}$, which is asymptotically normally distributed (Wald 1943). Replacing p with the maximum likelihood estimate (MLE) $\widehat{p} = r/(r+X)$, we find a CI for η on the basis of the result that

$$Z_\eta = \frac{\sqrt{r}(\widehat{\eta}-\eta)}{\sqrt{\eta/\widehat{p}}} \sim N(0, 1), \quad \text{asymptotically.} \tag{12}$$

Let $z_{\alpha/2}$ denote the upper $100\alpha/2$ percentile of the standard normal distribution. Solving the equation $Z_\eta^2 = z_{\alpha/2}^2$ for η, we find a $1-\alpha$ CI for η as

$$(L, U) = \widehat{\eta} + \frac{z_{\alpha/2}^2}{2r\widehat{p}} \pm \frac{z_{\alpha/2}}{r}\sqrt{\frac{z_{\alpha/2}^2}{4\widehat{p}^2} + \frac{X}{\widehat{p}}}. \tag{13}$$

A CI for p, deduced from the above CI, is given by $(1/(1+U), 1/(1+L))$.

3.4 Large sample confidence interval

The large sample CI, proposed in Tian et al. (2009), is based on the asymptotic normality of the MLE of p. The MLE of p is $\widehat{p} = r/(r+x)$ with variance $\text{Var}(\widehat{p}) = p^2(1-p)/r$. These results lead to the large sample CI as

$$\widehat{p} \pm z_{\alpha/2}\sqrt{\frac{\widehat{p}^2(1-\widehat{p})}{r}}. \tag{14}$$

The left endpoints of all CIs are truncated at 0 if they are negative, and the right endpoints are truncated at 1 if they are greater than 1.

3.5 Coverage probabilities and expected widths of confidence intervals

To judge the coverage probabilities and precisions of the exact, fiducial, score and large sample CIs, we computed the coverage probabilities as follows. For a given x of $X \sim \text{NBin}(r, p)$, let $(L(x; r, \alpha), U(x; r, \alpha))$ be a $1-\alpha$ CI for p. Then the coverage probability of the CI can be computed using the negative binomial probabilities as

$$\sum_{x=0}^{\infty} \binom{r+x-1}{x} p^r (1-p)^x I[L(x; r, \alpha) \le p \le U(x; r, \alpha)], \tag{15}$$

where $I[x]$ is the indicator function. Expected width of the CI can be computed using the above expression with the indicator function replaced by the width $[U(x; r, \alpha) - L(x; r, \alpha)]$.

We computed the coverage probabilities and expected widths of the (i) exact CI, (ii) fiducial CI, (iii) score CI and (iv) large sample CI for $r = 5$, 10, 20, 40, 80 and 120, and plotted them in Fig. 1. Examination of the plots clearly indicates that the large sample CIs are too liberal having coverage probabilities below 0.80 in many cases; it is liberal even for large values of r. The exact CI is too conservative even for large values of r, or equivalently, for large expected number of trials. The over coverage of the exact CI is increasing with increasing p. The fiducial and score CIs are also somewhat liberal for very small values of r and large values of p; see the plot for $r = 5$. For $r \ge 10$, both the score and fiducial CIs perform very similar, except for a few cases. For instance, the fiducial CI appears to be more liberal than the score CI for large p, $r = 10$ and 20.

In order to compare the coverage probabilities and expected widths simultaneously, we plotted the expected widths on the right pane in Fig. 1. We first note from Fig. 1 that the expected widths are in agreement with the coverage probabilities of the CIs. The coverage probabilities of the large sample CI are much lower than the nominal level 0.95, as a result, they are narrower than other CIs. The exact CIs are too conservative and so they are wider than others in all cases. Between the fiducial and score CIs,

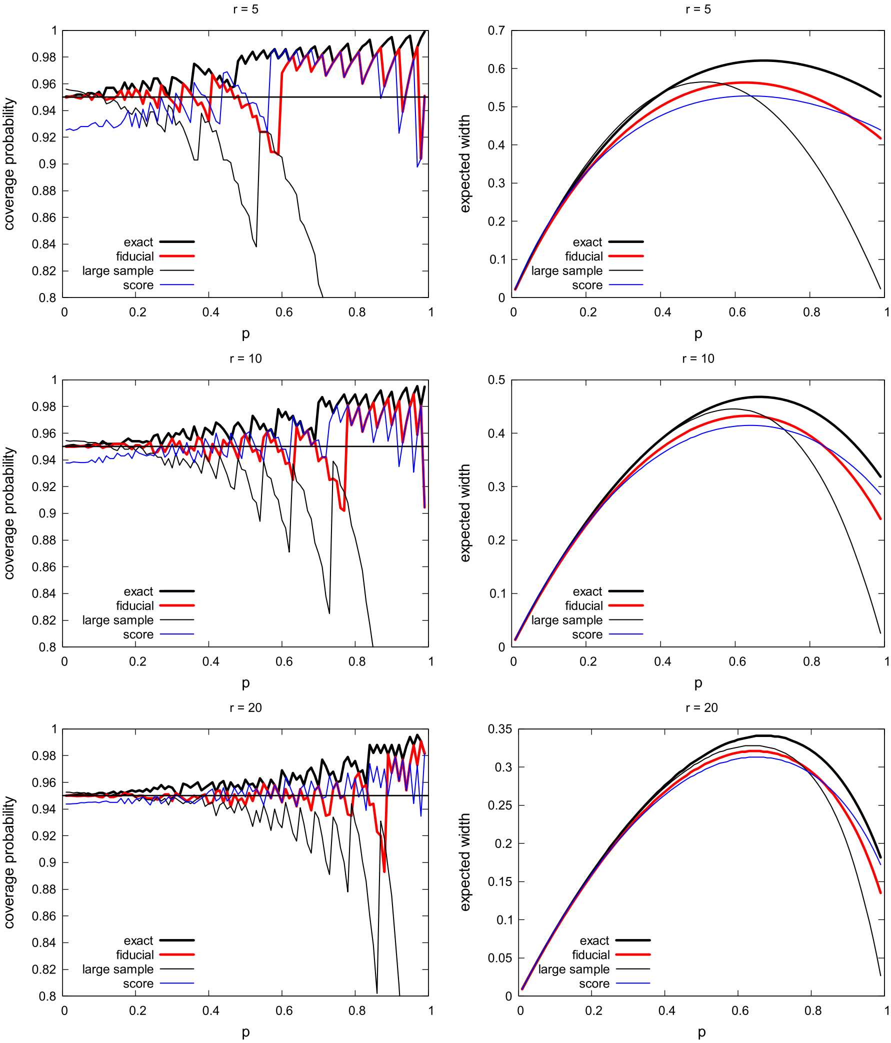

Fig. 1 Coverage probabilities and expected widths of 95% confidence intervals for p

the former appears to be slightly wider than the latter for p around 0.5; see the plots for $r = 5$, 10 and 20. In general, the fiducial and score CIs perform very similar in terms of coverage probabilities and expected widths, except that the score CI has an edge over the fiducial CI in a few cases. Overall, we see that the score CI followed by the fiducial CI are satisfactory in controlling the coverage probabilities close to the nominal level and maintaining the precision.

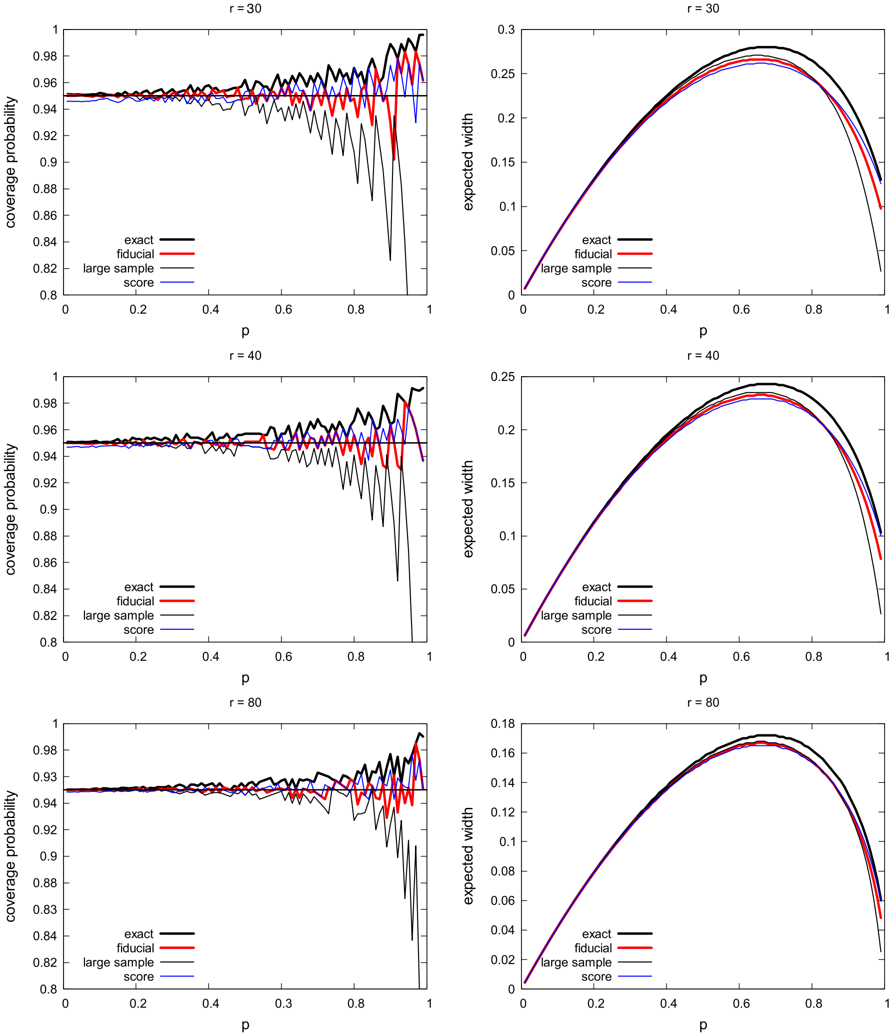

Fig. 1 continued

4 Confidence intervals for the expected number of trials

Let $X \sim \text{NBin}(r, p)$. On the basis of (X, r), we would like to estimate the expected number of trials required to observe s successes in a future negative binomial experiment with the same success probability. That is, we would like to find a CI for $E(s+Y)$, where $Y \sim \text{NBin}(s, p)$. Since $E(Y) = s(1-p)/p = s\eta$ and s is known, it is enough to find CI for η. As η is a decreasing function of p, any CI for p can be used to find a CI for the expected number of trials to observe s successes. Let (p_L, p_U) is a $1-\alpha$ CI for p. Then $((1-p_U)/p_U,\ (1-p_L)/p_L)$ is a $1-\alpha$ CI for η. By using the exact

Table 1 Coverage probabilities and expected widths (in parentheses) of 95% CIs for the mean number of trials needed to observe s successes

p	$r = 10, s = 5$		$r = 10, s = 10$		$r = 10, s = 20$	
	Exact	Score	Exact	Score	Exact	Score
.05	.950(146.8)	.938(126.7)	.950(293.7)	.938(253.5)	.950(587.4)	.938(507.1)
.15	.952(46.7)	.945(40.1)	.952(93.54)	.945(80.3)	.952(187.0)	.945(160.7)
.20	.954(34.2)	.947(29.3)	.954(68.42)	.947(58.6)	.954(136.8)	.947(117.2)
.25	.958(26.6)	.943(22.7)	.958(53.31)	.943(45.5)	.958(106.6)	.943(91.1)
.50	.959(11.2)	.959(9.52)	.959(22.58)	.959(19.0)	.959(45.16)	.959(38.0)
.60	.972(8.59)	.948(7.21)	.972(17.19)	.948(14.4)	.972(34.39)	.948(28.8)
.70	.983(6.57)	.955(5.50)	.983(13.15)	.955(11.0)	.983(26.30)	.955(22.0)
.80	.989(4.94)	.973(4.13)	.989(9.88)	.973(8.26)	.989(19.77)	.973(16.5)
.90	.991(3.53)	.966(2.97)	.991(7.06)	.966(5.94)	.991(14.12)	.966(11.8)
.95	.980(2.87)	.980(2.44)	.980(5.74)	.980(4.88)	.980(11.40)	.980(9.76)
p	$r = 30, s = 5$		$r = 30, s = 15$		$r = 30, s = 45$	
.05	.950(74.4)	.946(70.9)	.950(223)	.946(212)	.950(669)	.946(638)
.15	.952(23.5)	.945(22.4)	.952(70.7)	.945(67.2)	.952(212)	.945(201)
.20	.953(17.2)	.946(16.3)	.953(51.6)	.946(48.9)	.953(154)	.946(146)
.25	.953(13.3)	.946(12.6)	.953(40.1)	.946(37.9)	.953(120)	.946(113)
.50	.962(5.57)	.953(5.21)	.962(16.7)	.953(15.6)	.962(50.1)	.953(46.8)
.60	.956(4.20)	.956(3.90)	.956(12.6)	.956(11.7)	.956(37.8)	.956(35.1)
.70	.967(3.17)	.952(2.92)	.967(9.5)	.952(8.7)	.967(28.5)	.952(26.3)
.80	.971(2.32)	.971(2.13)	.971(6.9)	.971(6.3)	.971(20.8)	.971(19.1)
.90	.989(1.53)	.973(1.41)	.989(4.6)	.973(4.2)	.989(13.8)	.973(12.7)
.95	.993(1.12)	.974(1.04)	.993(3.3)	.974(3.1)	.993(10.1)	.974(9.43)

CI for p, we can find the exact CI for η. Note that the score CI for η is already defined in (13).

The coverage probabilities of these CIs for η should be similar to those for the CIs for p. In the following Table 1, we reported the coverage probabilities and expected widths of confidence intervals for the mean number of trials required to observe s successes for some values of r, s and confidence level 0.95. We observe from Table 1 that the coverage probabilities of the score CIs are very close to or greater than the nominal level for all the cases. Furthermore, the score CIs are shorter than the corresponding exact CIs for all values of r and s reported in the Table 1.

Remark 2 A CI for the mean of a NBin(r, p) distribution can also be obtained by parameterizing $\theta = r$ and $\mu = r(1-p)/p$. In this formulation, the PMF can be written as

$$P(X = k) = \frac{\Gamma(\theta + k)}{\Gamma(k+1)\Gamma(\theta)} \left(\frac{\mu}{\mu + \theta}\right)^k \left(\frac{\theta}{\mu + \theta}\right)^\theta, \quad k = 0, 1, 2, \ldots.$$

Table 2 95% confidence intervals based on simulated data from NB(r, p)

(n, r, p)	$(\sum_{i=1}^{n} x_i, s^2)$	Gamma CI	Exact	Fiducial	Score
(5,5,.3)	(56, 27.7)	(7.93, 15.0)	(7.11, 18.7)	(7.18, 18.6)	(7.02, 17.8)
(10,5,.3)	(111,67.9)	(6.93, 16.2)	(8.02, 15.8)	(8.06, 15.8)	(7.96, 15.5)
(20,5,.3)	(270, 45.1)	(11.1, 16.1)	(10.8, 17.2)	(10.8, 17.1)	(10.7, 17.0)
(50,5,.3)	(619, 44.2)	(10.9, 14.0)	(10.7, 14.4)	(10.7, 14.4)	(10.7, 14.3)

where θ is real positive, and both θ and μ are unknown. Shilane et al. (2010) have addressed the problem of finding CIs for the mean μ based on a sample $X_1, \ldots, X_n$ from a distribution with the above PMF and using the above parameterized model (Hilbe 2011). They have provided some CIs based on the result that the sample mean $\bar{X}$ has a gamma distribution asymptotically. Since $\sum_{i=1}^{n} X_i$ has the NB(nr, p) distribution, the asymptotic approaches in Shilane et al. can be used to find a CI for the mean $r(1-p)/p$. However, to apply our methods, the value of r should be known, and so they are not applicable to find a CI for μ in this parameterized model. For large n, as indicated by the CIs based on simulated data given in Table 2, the gamma asymptotic CI and our CIs are in agreement, but they are appreciably different for small to moderate sample sizes.

5 Prediction intervals for the number of trials

Some of the PIs that we consider below are based on the hypergeometric distribution and for easy reference, we describe the probability mass function (PMF) of the hypergeometric distribution with n = sample size, a = number of items with an attribute of interest and b = the number of items without the attribute, as

$$h(x; a, b, n) = P(X = x | a, b, n) = \frac{\binom{a}{x}\binom{b}{n-x}}{\binom{a+b}{n}}, \quad L_x \le x \le U_x, \tag{16}$$

where $L_x = \max\{0, n-b\}$ and $U_x = \min\{n, a\}$. Let us denote the PMF of the hypergeometric distribution by $h(x; a, b, n)$ and the cumulative distribution function (CDF) by $H(x; a, b, n)$.

5.1 Exact prediction interval

Let $X \sim$ NBin(r, p) independently of $Y \sim$ NBin(s, p). For a given (X, r), we like to predict Y or, equivalently, the number of trials $Y + s$ required to have s successes in a future negative binomial sampling with the same success probability p. The conditional PMF can be expressed as

$$P(X = x|X + Y = f) = \frac{\binom{x+r-1}{x}\binom{s+f-x-1}{f-x}}{\binom{f+r+s-1}{f}} = \frac{s}{f+s-x} h(x; x+r-1, s+f-x, f), \quad (17)$$

for $\max\{0, x-s\} \le x \le \min\{f, x+r-1\}$. In the above equation, $h(x; a, b, n)$ is the hypergeometric PMF in (16). To find an exact PI for the number of trials required to have s successes, we shall use the approach given in Thatcher (1964) for the binomial case. Following Thatcher's approach, we find the lower prediction limit for Y as the smallest integer L for which

$$P(X \ge x|X + L = f) = s \sum_{i=x}^{x+L} \frac{h(i; i+r-1, s+x+L-i, x+L)}{f+s-i} > \alpha.$$

The upper prediction limit for Y is the largest integer U for which

$$P(X \le x|X + U = f) = s \sum_{i=0}^{x} \frac{h(i; i+r-1, s+x+U-i, x+U)}{f+s-i} > \alpha.$$

The interval $[L, U]$ is the $(1-2\alpha)$ exact PI for Y.

5.2 Fiducial prediction interval

To find a fiducial PI, we shall use the general approach of Wang et al. (2012). For a given (r, x, s), the fiducial PI is based on the predictive distribution which is described by

$$\widetilde{Y}|W \sim \text{NBin}(s, W) \quad \text{and} \quad W \sim \text{beta}(r, x+.5),$$

where beta(a, b) denotes the beta distribution with shape parameters a and b. The probability mass function (PMF) of $\widetilde{Y}$ can be obtained as

$$\begin{aligned} P(\widetilde{Y} = y) &= E_W P(\widetilde{Y} = y|W) \\ &= \binom{s+y-1}{y} \frac{1}{\text{beta}(r, x+.5)} \int_0^1 w^{s+r-1}(1-w)^{y+x+.5-1} dw \\ &= \binom{s+y-1}{y} \frac{\text{beta}(r+s, y+x+.5)}{\text{beta}(r, x+.5)}. \end{aligned} \quad (18)$$

The above PMF is called the beta-negative binomial. See Sect. 6.2.3 of Johnson et al. (2005).

5.2.1 Equal-tailed prediction interval

For a given (r, x, s), the lower $100\alpha/2$ and the upper $100\alpha/2$ percentiles of $\widetilde{Y}$ form a $1-\alpha$ fiducial PI for Y. This PI $[\widetilde{L}, \widetilde{U}]$ can be computed as follows. The left endpoint is smallest integer $\widetilde{L}$ so that $\sum_{y=0}^{\widetilde{L}} P(\widetilde{Y} = y) > \alpha/2$ and the right endpoint is the largest integer $\widetilde{U}$ so that $\sum_{y=\widetilde{U}}^{\infty} P(\widetilde{Y} = y) > \alpha/2$.

5.2.2 Highest posterior mass prediction interval

The highest posterior mass (HPM) fiducial PIs are constructed by collecting integers with large probability masses according to the predicting distribution. The HPM-PIs are expected to be shorter than equal-tailed PIs. To compute the HPM-PI, let y_m denote the mode of the predicting distribution. The HPM-PI can be obtained by first adding y_m to the predicting set S and then adding the integers in decreasing order of their probability mass until $P(\widetilde{Y} \in S) \geq 1-\alpha$.

To find the mode of the distribution of $\widetilde{Y}$ defined in (18), it can be easily verified that

$$\frac{P(\widetilde{Y} = y+1)}{P(\widetilde{Y} = y)} = \frac{(s+y)(y+x+.5)}{(y+1)(r+s+x+y+.5)} > 1$$

if

$$y \leq \left\lceil \frac{s(x-1) - (r+x-.5(s-1))}{r+1} \right\rceil = y_m,$$

where $\lceil x \rceil$ is the ceiling function. Thus, the PMF of $P(\widetilde{Y} = y)$ is an increasing function for $y \leq y_m$ and decreasing for $y > y_m$. Therefore, y_m is the mode. For R code to compute the HPM-PI, see the appendix.

5.3 Joint sampling approach

We now propose a closed-form approximate PI based on the "joint sampling approach" which is similar to the one used to find confidence interval in a calibration problem (e.g., see Brown 1982, Sect. 1.2). This approach was also used to find an approximate PI for binomial distributions (Krishnamoorthy and Peng 2011). To describe this approach, we first note that $X \sim \text{NBin}(r, p)$ independently of $Y \sim \text{NBin}(s, p)$ and the sum $X+Y$ has also $\text{NBin}(r+s, p)$ distribution. Let $\eta = (1-p)/p$. Then $E\left(\frac{X+Y}{r+s}\right) = \eta$. Let $\widehat{\eta}_{xy} = \frac{X+Y}{r+s}$. Consider the quantity

$$\frac{rY - sX}{\sqrt{\text{Var}(rY - sX)}} = \frac{rY - sX}{\sqrt{rs(r+s)\eta/p}}.$$

Since $E(rY - rX) = 0$, by the Wald result, the above quantity follows the standard normal distribution asymptotically. Replacing η with $\widehat{\eta}_{xy}$ and p with the MLE $\widehat{p} = r/(r+X)$ in the above expression, we see that

$$C_X \frac{(sX - rY)^2}{X + Y} \sim Z^2, \quad \text{asymptotically,}$$

where $Z \sim N(0, 1)$ and $C_X = 1/[s(r + X)]$. Let $z_{\alpha/2}$ denote the upper $100\alpha/2$ percentile of the standard normal distribution. Then solving the equation

$$C_X \frac{(sX - rY)^2}{X + Y} = z_{\alpha/2}^2$$

for Y, we find the roots as

$$(L, U) = s\left(\frac{X}{r} + \frac{z_{\alpha/2}^2}{2r^2}(r + X)\right) \mp \frac{z_{\alpha/2}}{r}\sqrt{s(r + X)}\sqrt{s\left(\frac{X}{r} + \frac{z_{\alpha/2}^2}{4r^2}(r + X)\right) + X}.$$

The $1 - \alpha$ PI for the number of trials required to have s successes in a future negative binomial experiment with success probability p is given by

$$[\lceil L + s \rceil, \lfloor U + s \rfloor], \tag{19}$$

where $\lceil x \rceil$ and $\lfloor x \rfloor$ are the ceiling and floor functions, respectively.

5.4 Coverage probabilities and expected widths of prediction intervals

For a given (r, s, p, α), the exact coverage probability of a PI $[L(x, r, s, \alpha), U(x, r, s, \alpha)]$ can be evaluated using the expression

$$\sum_{x=0}^{\infty}\sum_{y=0}^{\infty}\binom{r+x-1}{x}\binom{s+y-1}{y}p^{r+s}(1-p)^{x+y}I[L(x, r, s, \alpha) \le y \le U(x, r, s, \alpha)], \tag{20}$$

where $I[x]$ is the indicator function. The coverage probabilities of a good PI should be close to the nominal level. The above expression with the indicator function replaced by $U(x, r, s, \alpha) - L(x, r, s, \alpha)$ can be used to compute the expected width.

We evaluated the coverage probabilities of the exact, equal-tailed, HPM prediction intervals and the PI based on the joint sampling approach (JS-PI) for some values of (r, s) at the confidence level 0.95. These coverage probabilities were plotted in Fig. 2. We first observe from these plots that the exact PIs are too conservative for all values of (r, s). The HPM PI is also conservative but less conservative than the exact one. The equal-tailed PI is liberal for some values in the parameter space; see the plots for $(r, s) = (10, 10)$ and $(10, 20)$. The JS-PI is also slightly liberal for the values of p near zero, but it maintains the coverage probability very close to the nominal level for most of the cases.

As the magnitudes of the coverage probability and the expected width of a PI are quite different, the plots of such values are less informative. Instead, we tabulated the coverage probabilities and corresponding expected widths of all PIs for some values of p, r and s in Table 3. We first observe that the exact PI is too conservative and so they are wider than others for all the cases reported in Table 3. Between the equal-tailed PI

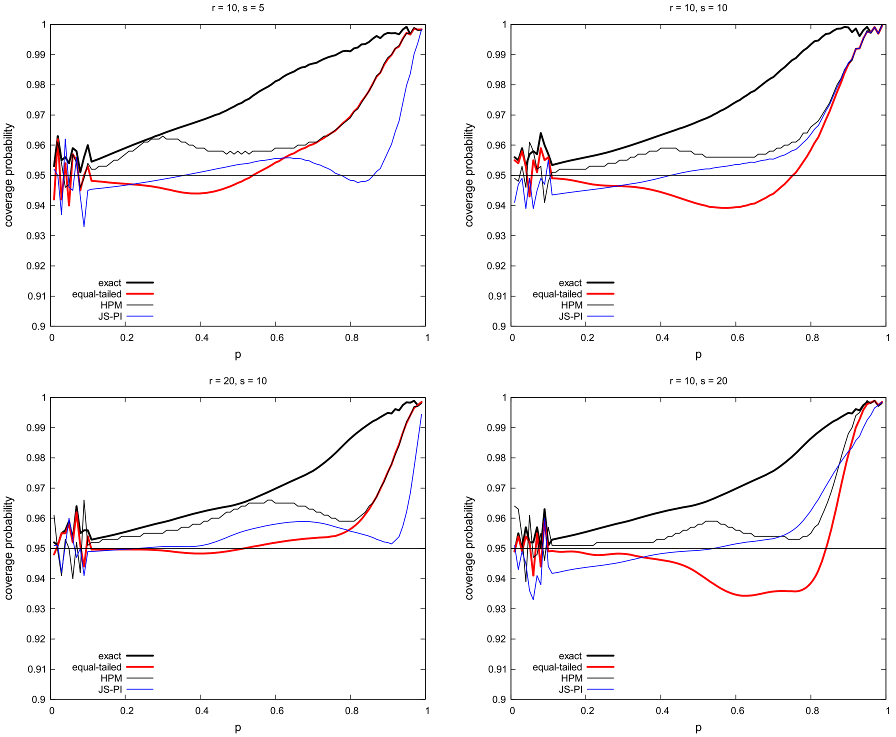

Fig. 2 Coverage probabilities of 95% prediction intervals for NB(s, p) distributions

and the HPM prediction interval, the latter has better coverage probabilities than the former for all the cases. So the HPM-PI is preferable to the equal-tailed PI. The PI by the joint sampling approach (JS-PI) is shorter than other PIs, and it also has coverage probabilities close to the nominal level for all the cases. We also note that between the HPM PI and the JS-PI, the latter is simple to compute.

6 Tolerance intervals

Let $X \sim \text{NBin}(r, p)$. On the basis of (X, r), we like to find an equal-tailed tolerance interval (TI) for a NBin(s, p) distribution. Let $\kappa_q(p; s)$ denote the qth quantile of a NBin(s, p) distribution. A γ content and $1-\alpha$ coverage equal-tailed TI or simply a $(\gamma, 1-\alpha)$ equal-tailed TI $[L(X, r, s), U(X, r, s)]$ for a NBin(s, p) distribution is constructed so that it includes the interval $\left[\kappa_{\frac{1-\gamma}{2}}(p; s), \kappa_{\frac{1+\gamma}{2}}(p; s)\right]$ with confidence $1-\alpha$; see Chap. 1 of Krishnamoorthy and Mathew (2009). That is,

$$P\left\{L(X, r, s) \le \kappa_{\frac{1-\gamma}{2}} \text{ and } \kappa_{\frac{1+\gamma}{2}} \le U(X, r, s)\right\} = 1-\alpha. \tag{21}$$

Table 3 Coverage probabilities and expected widths of 95% prediction intervals

p	$r = 10, s = 5$				$r = 10, s = 10$			
	Exact	Eq. tailed	HPM	JS-PI	Exact	Eq. tailed	HPM	JS-PI
.05	.954(243)	.940(236.)	.947(215)	.946(209)	.952(403)	.943(398)	.961(361)	.949(349)
.10	.960(118.)	.953(114.)	.954(106)	.945(102)	.957(198)	.956(193)	.948(177)	.955(170)
.15	.956(76.9)	.947(73.0)	.953(68.4)	.945(66.1)	.954(127)	.948(123)	.952(115)	.944(109)
.20	.959(56.1)	.947(52.5)	.957(49.9)	.946(47.8)	.955(93.3)	.947(89.0)	.952(84.1)	.945(79.6)
.25	.961(43.6)	.946(42.0)	.961(38.8)	.947(36.8)	.957(72.6)	.946(68.1)	.953(65.2)	.945(61.5)
.30	.963(35.3)	.945(31.9)	.963(31.4)	.948(29.4)	.959(58.6)	.946(54.9)	.955(52.5)	.946(49.4)
.35	.966(29.2)	.944(26.0)	.961(26.0)	.949(24.1)	.961(48.6)	.945(44.9)	.957(43.4)	.947(40.7)
.40	.968(24.7)	.944(21.5)	.959(21.8)	.951(21.0)	.963(41.1)	.944(37.3)	.959(36.5)	.949(34.1)
.45	.970(21.1)	.945(17.9)	.958(18.3)	.952(16.8)	.965(35.1)	.942(31.5)	.958(31.1)	.950(29.0)
.50	.973(18.2)	.947(15.1)	.958(15.4)	.953(14.2)	.967(33.4)	.940(26.7)	.957(26.6)	.951(24.8)
.55	.977(15.7)	.950(12.7)	.958(12.9)	.954(12.0)	.970(26.4)	.939(22.7)	.956(22.8)	.952(21.3)
.60	.981(13.6)	.954(11.7)	.959(10.9)	.955(10.3)	.974(23.0)	.939(19.3)	.956(19.5)	.953(18.4)
.70	.987(10.1)	.960(7.64)	.961(7.65)	.954(7.41)	.982(17.4)	.943(13.7)	.958(13.9)	.955(13.6)
.80	.991(7.37)	.969(5.29)	.969(5.29)	.948(5.13)	.993(12.9)	.958(9.4)	.966(9.48)	.963(9.6)
.90	.997(5.08)	.988(3.41)	.989(3.41)	.957(3.05)	.998(9.20)	.986(5.9)	.987(5.97)	.987(6.1)
.99	.998(3.19)	.998(2.10)	.998(2.10)	.998(1.14)	.999(6.29)	.999(3.2)	1.00(3.29)	.999(3.2)

Table 3 continued

p	$r = 10, s = 5$				$r = 10, s = 10$			
	Exact	Eq. tailed	HPM	JS-PI	Exact	Eq. tailed	HPM	JS-PI
	$r = 10, s = 20$				$r = 20, s = 10$			
.05	.952(711)	.952(704)	.961(649)	.936(614)	.959(329)	.958(318)	.947(299)	.960(299)
.10	.951(346)	.949(334)	.957(311)	.944(298)	.956(156)	.954(154)	.945(146)	.949(145)
.15	.953(224)	.948(218)	.951(203)	.942(192)	.953(107)	.949(98.3)	.949(95.0)	.949(93.1)
.20	.955(164)	.948(158)	.951(148)	.943(139)	.955(73.5)	.949(71.1)	.950(69.2)	.949(67.5)
.25	.957(127)	.947(122)	.952(114)	.944(108)	.957(57.0)	.949(54.7)	.950(53.7)	.950(52.1)
.30	.958(103)	.948(98.0)	.952(92.3)	.945(87.3)	.958(46.0)	.949(43.7)	.951(43.3)	.950(41.7)
.35	.960(85.8)	.947(80.5)	.952(76.3)	.946(72.1)	.960(38.1)	.948(35.9)	.951(35.8)	.950(34.3)
.40	.962(72.6)	.946(67.3)	.953(64.1)	.948(66.6)	.962(32.1)	.948(29.9)	.951(30.1)	.950(28.7)
.45	.963(62.2)	.944(57.0)	.955(54.5)	.948(51.6)	.963(27.4)	.948(25.2)	.953(25.6)	.952(24.2)
.50	.964(53.8)	.941(48.6)	.958(46.8)	.949(44.3)	.964(23.6)	.949(21.4)	.955(22.0)	.955(21.7)
.55	.966(46.8)	.937(41.6)	.959(40.3)	.950(38.2)	.966(19.4)	.950(18.3)	.957(18.9)	.956(17.8)
.60	.969(40.8)	.934(35.6)	.956(34.8)	.951(33.1)	.969(17.6)	.951(15.6)	.958(16.3)	.958(15.4)
.70	.975(31.0)	.935(25.6)	.954(25.6)	.953(24.7)	.975(13.2)	.953(11.3)	.959(11.7)	.958(11.3)
.80	.986(23.0)	.938(17.5)	.955(18.0)	.962(18.0)	.986(9.4)	.956(7.6)	.956(7.75)	.955(7.6)
.90	.994(16.1)	.981(14.8)	.988(11.1)	.982(12.2)	.994(6.1)	.975(4.4)	.952(4.41)	.951(4.2)
.99	.998(11.5)	.998(9.5)	.998(5.58)	.998(7.4)	.998(3.3)	.998(2.1)	.994(2.19)	.994(1.3)

Note that 100γ percent of the $\text{NBin}(p, s)$ distribution falls in the interval $\left[\kappa_{\frac{1-\gamma}{2}}(p; s), \kappa_{\frac{1+\gamma}{2}}(p; s)\right]$. So a $1-\alpha$ CI for this interval will include at least 100γ percent of the distribution with confidence $1-\alpha$.

6.1 Tolerance intervals based on different confidence intervals

A CI for $\left[\kappa_{\frac{1-\gamma}{2}}(p; s),\ \kappa_{\frac{1+\gamma}{2}}(p; s)\right]$ can be found using the confidence limits for p based on $X \sim \text{NBin}(r, p)$. Since the negative binomial distribution is stochastically decreasing in p, the quantile $\kappa_q(p; s)$ is a decreasing function of p. Using this fact, we see that

$$\kappa_{\frac{1-\gamma}{2}}(p_U; s) \le \kappa_{\frac{1-\gamma}{2}}(p; s) \quad \text{and} \quad \kappa_{\frac{1+\gamma}{2}}(p; s) \le \kappa_{\frac{1+\gamma}{2}}(p_L; s) \text{ with probability } 1-\alpha,$$

where (p_L, p_U) is a $1-\alpha$ CI for p based on $X \sim \text{NBin}(r, p)$. In other words,

$$\left[\kappa_{\frac{1-\gamma}{2}}(p_U; s), \kappa_{\frac{1+\gamma}{2}}(p_L; s)\right] \tag{22}$$

is a $(\gamma, 1-\alpha)$ equal-tailed TI for the $\text{NBin}(s, p)$ distribution.

Remark 3 The approach of finding a TI given in Young (2011) is essentially the same as the above approach. Mathew and Young (2013) have also proposed a TI on the basis of fiducial approach which is equivalent to the above TI based on the fiducial CI for p. We follow the above approach as it is simple and is easy to implement in R as shown below.

The properties of the TIs defined above are similar to those of the CIs. For example, if (p_L, p_U) is an exact CI for p based on $X \sim \text{NBin}(r, p)$, then the TI defined above is an exact TI in the sense that the coverage probabilities are always greater than or equal to $1-\alpha$ for all p. We also note that these TIs are easy to compute using the quantile function available in software packages. For example, after finding the CI (p_L, p_U), the R function `qnbinom(q, s, p)` can be used to find the quantiles. Specifically, the $(\gamma, 1-\alpha)$ TI can be computed as

$$[\texttt{qnbinom}((1-\gamma)/2, s, p_U),\ \ \texttt{qnbinom}((1+\gamma)/2, s, p_L)].$$

Remark 4 Cai and Wang (2009) have proposed first-order and second-order probability matching tolerance intervals for discrete distributions in exponential families which include binomial, Poisson and negative binomial distributions. The two-sided TIs proposed in the Cai and Wang's paper is determined so that

$$P_X\{F(U(X)) - F(L(X)) \ge \gamma\} = 1-\alpha,$$

where $F(x)$ denotes the CDF of a $\text{NBin}(r, p)$ distribution. It is important to note that Cai and Wang (2009) have defined the random variable X as the number of successes until the rth failure, which is different from our commonly used definition.

This two-sided TI is expected to be shorter than an equal-tailed TI, because the latter is constructed to include the lower and upper $100(1+\gamma)/2$ percentiles. That is, an equal-tailed $(\gamma,\, 1-\alpha)$ TI not only includes at least $100\gamma\%$ of the population, but also includes the lower and upper $100(1+\gamma)/2$ percentiles whereas a two-sided TI is constructed just to include at least $100\gamma\%$ of the population. Our coverage studies indicated that the Cai and Wang TIs are satisfactory for large r and $.05 \le p \le .95$. They are not satisfactory for value of p at boundaries even when r is large. For example, for $r = 50$, the coverage probabilities of (.90, .95) two-sided TIs for a NBin(50, p) distribution at $p = .01, .02, .03, .04$ and $.05$ are **.303, .550, .707, .848** and .902 respectively; at $p = .97, .98$ and $.99$, they are .909, .866 and .781, respectively. So these two-sided TIs should be used with caution, and further research is needed to improve these TIs.

6.2 Coverage probabilities and expected widths of tolerance intervals

To judge the coverage probabilities and expected widths of the TIs, we computed the exact coverage probability of a TI $\left[\kappa_{\frac{1-\gamma}{2}}(p_U; s), \kappa_{\frac{1+\gamma}{2}}(p_L; s)\right]$ using the following expression:

$$\sum_{x=0}^{\infty} \binom{r+x-1}{x} p^r (1-p)^x I\left[\kappa_{\frac{1-\gamma}{2}}(p_U; s) \le \kappa_{\frac{1-\gamma}{2}}(p; s) \text{ and } \kappa_{\frac{1+\gamma}{2}}(p; s) \le \kappa_{\frac{1+\gamma}{2}}(p_L; s)\right], \tag{23}$$

where $I[x]$ is the indicator function.

We evaluated coverage probabilities of the following equal-tailed TIs: (i) the interval in (22) with the exact CI (10) for p (referred to as the exact TI), (ii) the interval in (22) with the score CI deduced from (13) for p (referred to as the score TI), and (iii) the interval in (22) with the large sample CI (14) for p (referred to as the large sample TI). The coverage probabilities of (.90, .95) TIs were computed for $(r, s) = (20, 10), (20, 30), (30, 10), (30, 20), (50, 10)$ and (50,40). These coverage probabilities were plotted in Fig. 3. We observe from these plots that the large sample TIs are, in general, too liberal having coverage probabilities much smaller than the nominal level .95. The score TI and the exact TI are conservative except that in some cases the score TIs are less conservative than the exact TIs. For example, see the cases $(r, s) = (20, 10), (20, 30), (30, 20)$ and (50,40).

We further compared the TIs with respect to expected widths. We tabulated some summary statistics of the expected widths of the exact and score TIs in Table 4. The summary statistics are based on expected widths for $p = .001(.001).999$. We considered only these two TIs because the third one, the large sample TI, is too liberal. The summary statistics of the expected widths of these two TIs clearly indicate the score TI is narrower than the exact TI for all the cases. This comparison result was anticipated because the exact TIs are more conservative than the score TIs.

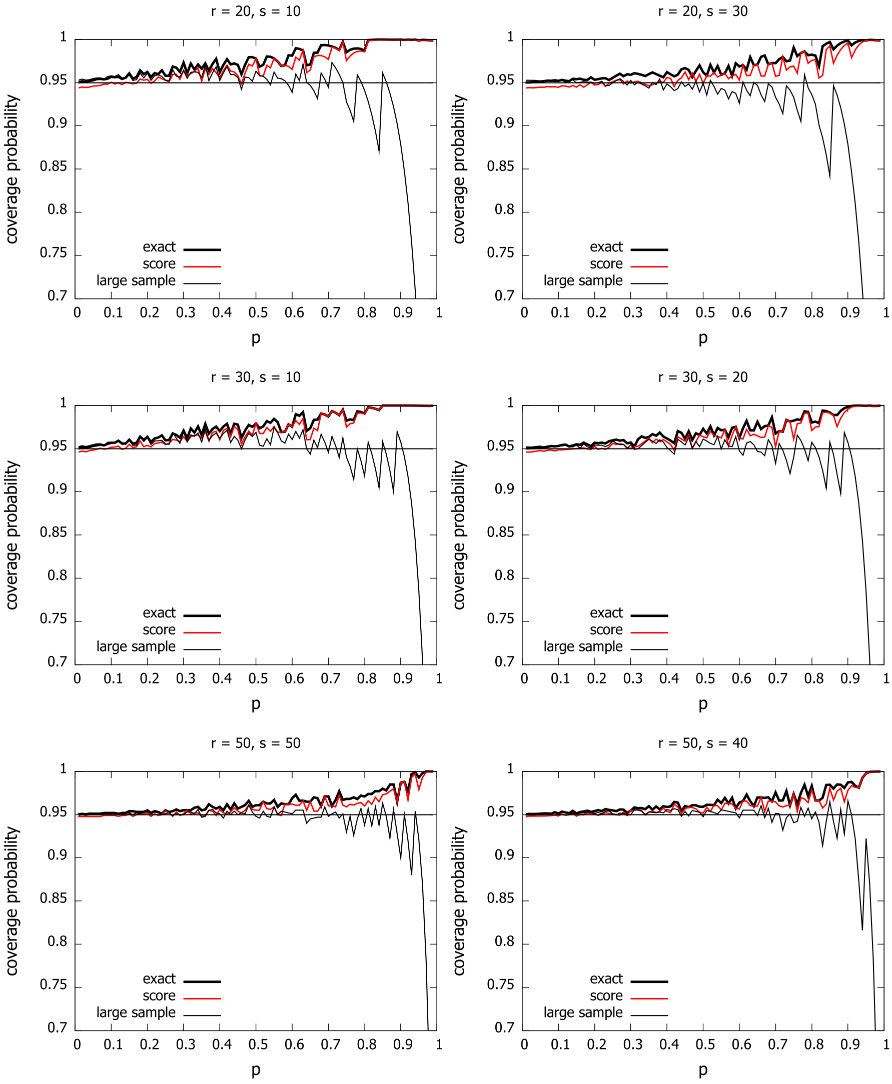

Fig. 3 Coverage probabilities of (.90, .95) tolerance intervals for NB(s, p) distributions

7 Examples

Example 1 A study was conducted to find the association between the maternal congenital heart disease and low birthweights of infants (Kikuchi 1987). A negative binomial sampling plan was used to recruit pregnant mothers until $r = 5$ maternal congenital heart disease mothers were observed. The 5th congenital heart disease mother was observed at the 146th selection. Thus, the number of normal mothers is $x = 141$ and the number of congenital heart disease mothers is $r = 5$. Tian et al. (2009) have used these results to illustrate different interval estimation methods for the proportion p

Table 4 Summary statistics of expected widths of (.90, .95) tolerance intervals for NBin$(s,\ p)$ based on $X \sim \text{NBin}(r,\ p)$

	(r, s)											
	(20, 10)		(20, 30)		(30, 10)		(30, 20)		(50, 10)		(50, 40)	
	Exact	Score	Exact	Score	Exact	Score	Exact	Score	Exact	Score	Exact	Score
mean	152.5	134.7	347.4	303.7	133.7	121.7	219.5	198.6	116.9	109.2	312.7	288.7
min	5.0	5.0	11.0	10.0	4.0	4.0	6.0	6.0	2.0	2.0	6.1	6.1
5%	7.0	6.8	15.9	14.8	5.6	5.6	9.1	8.9	4.3	4.3	11.9	11.7
25%	15.3	14.5	35.6	33.2	13.2	12.7	22.1	21.0	11.4	11.2	31.2	30.2
med	31.9	29.9	72.9	68.1	27.8	26.7	45.8	43.8	24.2	23.6	65.0	63.2
75%	76.9	72.3	175.4	164.2	67.4	64.6	110.6	106.0	58.9	57.5	157.5	153.5
95%	422.9	397.8	962.4	902.9	371.1	356.8	609.0	584.7	325.1	317.7	868.4	847.4
max	22040	12120	50140	25390	19350	11850	31750	18500	16960	11550	45290	27910

Table 5 95% confidence intervals for p and 95% prediction intervals

Methods	Confidence intervals	Methods	Prediction intervals
Exact	(.011, .069)	Exact	[178, 1438]
Large sample	(.005, .064)	Equal-tailed	[179, 1427]
Fiducial	(.011, .069)	HPM	[121, 1186]
Score	(.015, .078)	Score	[140, 1073]

of the mothers with the congenital heart disease. Young (2014) has used the data to construct a two-sided tolerance interval for future samples of pregnant mothers.

We computed the 95% CIs for the proportion p of the mothers with the congenital heart disease using different methods described earlier, and presented them in Table 5. Note that the uniformly minimum variance unbiased estimate for p is $(r-1)/(r+x-1) = 4/145 = .0276$, which indicates that the population proportion is likely to be small. For such cases, all interval estimation methods are satisfactory and similar in terms of coverage probabilities and expected widths; see Fig. 1. So, for this example, all the methods produced CIs that are not much different.

Suppose it is desired to estimate the expected number of pregnant mothers to be examined in order to find 15 congenial mothers. The 95% CI based on the exact CI for p is (202, 1323). That is, on average, 202 to 1323 mothers are to be examined in order to find 15 congenial mothers. The 95% CI based on the score method is [178, 1004], which is much shorter than the exact CI.

To illustrate the construction of the PIs, we computed 95% PIs for the number of pregnant mothers to be examined in order to find $s = 15$ congenital heart disease pregnant mothers. PIs based on various methods are given in Table 5. The PIs for the number of pregnant mothers to be examined to observe 15 mothers with congenital heart disease are quite different. Among all the PIs the score PI [140, 1073] is the shortest, and it means that 140 to 1073 pregnant mothers to be examined to find 15 congenital heart disease pregnant mothers.

Suppose it is desired to find an interval with 90% confidence where 90% of all future samples of healthy pregnant mothers would fall if a similar study were to be performed; that is, to capture a target number of $s = 5$ mothers with congenial heart disease. Using $(r, x) = (5, 141)$, we computed the 90% exact CI (9) as (.0136, .0620). On the basis of this exact CI, the (.90, .90) TI was computed using the R function as

$$[\texttt{qnbinom}(.05, 5, .0620), \texttt{qnbinom}(.95, 5, .0136)] = [28, 666].$$

Similarly, using the score CI (13), the (.90, .90) TI was computed as [25, 539]. The (.90, .90) TIs for the total sample sizes for negative binomial sampling scheme to observe $s = 5$ mothers with congenial heart disease can be obtained by adding five to these TIs.

Table 6 The American Community Survey data from 2006 to 2010

Year	Housing units		Group quarters	
	Initial	Final	Initial	Final
2010	2,935,687	1,940,293	197,970	145,552
2009	2,933,345	1,940,458	199,768	147,374
2008	2,930,800	1,954,659	187,783	146,619
2007	2,922,519	1,960,496	187,952	143,200
2006	2,921,218	1,991,487	190,572	145,995
Mean	2,928,714	1,957,479	192,809	145,748
Total	14,643,569	9,787,393	964,045	728,740

Example 2 The American Community Survey (ACS) collects data[1] to help local officials, community leaders and businesses to understand the changes taking place in their communities and regions. The data are collected annually from all counties and county equivalents for various purposes. Data were first collected by self-enumeration via mailback. The completed ACS captures roughly 65% of housing units (HUs) and 75% of group quarters (GQs) originally listed in the sample. For more details on data collection, see Young (2014) who used the data to demonstrate the construction of tolerance intervals for a negative binomial distribution. The following Table 6, taken from Young (2014), shows the initial sample sizes and final interview sizes for HUs and GQs during the years 2006–2010.

It is quite common in mail-in surveys that the final target sample size is considerably smaller than the initial samples. Young (2014) has used the average initial sample size as the total number of negative binomial trials required to obtain the average final interview size. The averages are given in the last row of Table 6. For example, Young has used $r = 1,957,479$ and $x = 2,928,714 - 1,957,479 = 971,235$ to construct a TI for HUs. As negative binomial distributions have additive property, we can use the total initial sample size 14,643,569 as $r + x$ and the total final interview size 9,787,393 as r.

7.1 Prediction intervals

For a future survey of target size of 2,000,000, we would like to predict the initial sample size for HUs with 95% confidence. In our present notations, $s = 2,000,000$, $r + x = 14,643,569$ and $r = 9,787,393$, and we like to find a 90% PI for $s + Y$, where $Y \sim \text{NBin}(s, p)$. The PIs based on all four methods are reported in Table 7. For GQs, we like to predict the initial sample size to obtain 200,000 final interviews. In this case, $r + x = 964,045$, $r = 728,740$ and $s = 200,000$. Using these data, we computed 90% PIs for the initial sample size and reported them in Table 7. All four methods produced similar PIs for both cases. To interpret the PI, we note that the exact PI for GQs is $[264,036,\ 265,122]$. This means that an initial sample size of 264,036

[1] http://www.census.gov/acs/www/methodology/sample_size_and_data_quality/.

Table 7 90% prediction intervals based on the ACS data

Methods	Housing units	Group quarters
Exact	[2,990,134, 2,994,534]	(264,036 265,122)
Equal-tailed	[2,990,134, 2,994,532]	(264,037, 265,122)
HPM	[2,990,133, 2,994,532]	(264,036, 265,122)
Score	[2,990,134, 2,994,532]	(264,037, 265,121)

to 265,122 is needed to obtain the final target size of 200,000 responses. Other PIs can be interpreted similarly.

7.2 Tolerance intervals

Based on the average of the initial sample and final interview sizes, Young (2014) has estimated (.99, .95) TI where 99% of all initial sample sizes will fall if target final sample size of 2,000,000 is required for HUs. Using the large sample CI, the (.99, .95) TI for the HUs was computed as **[2,986,789, 2,997,894]**. We computed the same TI using the total initial sample size and the total final interview size. On the basis of totals, the TIs based on the exact CI, score CI and the large sample CI are the same and is **[2,988,119, 2,996,556]**. This means that 99% of initial sample sizes from this interval would produce a target response size of 2,000,000 with confidence 95%. Young (2014) has also estimated (.99, .95) TI where 99% of all initial sample sizes fall if target final sample size of 200,000 is required for GQs. Using the large sample CI based on average sizes, Young has computed the (.99, .95) TI as **[263,163, 266,011]**. On the basis of the total sizes, we find the TIs based on the exact, score and large sample CIs are the same and is **[263,530, 265,636]**. That is, 99% of all initial sample sizes between 263,636 and 265,636 would produce 200,000 final responses with confidence 0.95.

Finally, we note that the TIs based on the total sizes are shorter than the corresponding TIs based on the average sizes for both HUs and GQs. Furthermore, all methods produced the same TI because the sample sizes are very large.

8 Concluding remarks

The classical exact methods for discrete distributions are known to be too conservative producing confidence intervals and prediction intervals that are unnecessarily wide or tests that are less powerful. For the binomial and Poisson distributions, Agresti and Coull (1988), Brown et al. (2001) and many other authors have recommended alternative approximate approaches for constructing confidence intervals and prediction intervals with satisfactory coverage probabilities and good precision. In this article, we have provided similar approximate methods for constructing CIs, PIs and TIs for negative binomial distributions. Furthermore, we showed that the approximate CIs and PIs have good coverage properties with expected widths narrower than those of the exact CIs and PIs. In terms of simplicity and accuracy, the PI based on the joint sampling approach, the score CI and the TI based on the score CI are preferable to others. These statistical intervals are not only easy to compute, but also safe to use in practical applications.

Acknowledgements The authors are grateful to three reviewers for providing valuable comments, suggestions and references relevant to this article.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

Appendix

R code to compute HPM-PI

```
NB.HPM.PI = function(r, x, s, cl){
al = (1-cl)/2
probs = function(y,r,x,s){
lpr = lgamma(s+y)-lgamma(y+1)-lgamma(s)+
    + lgamma(r+s) + lgamma(y+x+.5)-lgamma(r+s+y+x+.5)+
    + lgamma(r+x+.5)-lgamma(r)-lgamma(x+.5)
return(exp(lpr))
}
md = ceiling((s*(x-1)-(r+x-.5*(s-1)))/(r+1))
cdf = 0
if(md <= 3){
yi = 0
repeat{
cdf = cdf + probs(yi,r,x,s)
if(cdf >= cl){break}
yi = yi + 1}
return(c(s, yi+s))
}
cdf = probs(md,r,x,s)
yf = md+1; yb = md-1
probsf = probs(yf,r,x,s)
probsb = probs(yb,r,x,s)
repeat{
if(probsf >= probsb){
cdf = cdf + probsf
if(cdf >= cl){break}
yf = yf+1
probsf = probs(yf,r,x,s)
}
else{
cdf = cdf + probsb
if(yb <= 0 | cdf >= cl){break}
yb = yb-1
probsb = probs(yb,r,x,s)}
}
if(cdf >= cl){
return(c(yb+s, yf+s))}
repeat{
probf = probs(yf,r,x,s)
cdf = cdf + probf
if(cdf >= cl){break}
yf = yf+1
}
return(c(yb+s,yf+s))
}
> NB.HPM.PI(5,141,15,.95)
[1]  121 1186
```

References

Agresti A, Coull BA (1988) Approximate is better than “exact" for interval estimation of binomial proportion. Am Stat 52:119–125

Brown PJ (1982) Multivariate calibration. J R Stat Soc B 44:287–321

Brown LD, Cai T, Das Gupta A (2001) Interval estimation for a binomial proportion (with discussion). Stat Sci 16:101–133

Cai TT, Wang H (2009) Tolerance intervals for discrete distributions in exponential families. Stat Sin 19:905–923

Casella G, Berger RL (2001) Statistical inference, 2nd edn. Duxbury, Belmont

Clemans KG (1959) Confidence limits in the case of the geometric distribution. Biometrika 46:260–264

Dunsmore IR (1976) A note on Faulkenberry’s method of obtaining prediction intervals. J Am Stat Assoc 71:193–194

Fisher RA (1935) The fiducial argument in statistical inference. Ann Eugen 6(4):391–398

George VT, Elston RC (1993) Confidence limits based on the first occurrence of an event. Stat Med 11:685–690

Hahn GJ, Chandra R (1981) Tolerance intervals for Poisson and binomial random variables. J Qual Technol 13:100–110

Hahn GJ, Meeker WQ (1991) Statistical intervals. Wiley, Hoboken

Haldane JBS (1945) On a method of estimating frequencies. Biometrika 33:222–225

Hannig J (2009) On generalized fiducial inference. Stat Sin 19:491–544

Hilbe JM (2011) Negative binomial regression, 2nd edn. University Press, Cambridge

Johnson NL, Kemp AW, Kotz S (2005) Univariate discrete distributions, 3rd edn. Wiley, Hoboken

Kikuchi DA (1987) Inverse sampling in case control studies involving a rare exposure. Biom J 29:243–246

Knüsel L (1994) The prediction problem as the dual form of the two-sample problem with applications to the Poisson and the binomial distribution. Am Stat 48:214–219

Krishnamoorthy K, Lee M (2010) Inference for functions of parameters in discrete distributions based on fiducial approach: binomial and Poisson cases. J Stat Plan Inference 140:1182–1192

Krishnamoorthy K, Mathew T (2009) Statistical tolerance regions: theory, applications, and computation. Wiley, Hoboken

Krishnamoorthy K, Peng J (2011) Improved closed-form prediction intervals for binomial and Poisson distributions. J Stat Plan Inference 141:1709–1718

Krishnamoorthy K, Xia Y, Xie F (2011) A simple approximate procedure for constructing binomial and Poisson tolerance intervals. Commun Stat 40:2443–2458

Lui KJ (1995) Confidence limits for the population prevalence rate based on the negative binomial distribution. Stat Med 14:1471–1477

Madden LV, Hughes G, Munkvold GP (1996) Plant disease incidence: inverse sampling, sequential sampling, and confidence intervals when observed mean incidence is zero. Crop Prot 15:621–632

Mathew T, Young D (2013) Fiducial-based tolerance intervals for some discrete distributions. Comput Stat Data Anal 61:38–49

Patil GP (1960) On the evaluation of the negative binomial distribution with examples. Technometrics 2:501–505

Shilane D, Evans SN, Hubbard AE (2010) Confidence intervals for negative binomial random variables of high dispersion. Int J Biostat Article10 6:1–10

Thatcher AR (1964) Relationships between Bayesian and confidence limits for prediction. J Roy Stat Soc B 26:176–192

Thulin M, Zwanzig S (2017) Exact confidence intervals and hypothesis tests for parameters of discrete distributions. Bernoulli 23:479–502

Tian M, Tang ML, Ng HKT, Chan PS (2009) A comparative study of confidence intervals for negative binomial proportions. J Stat Comput Simul 79:241–249

Wald A (1943) Tests of statistical hypotheses concerning several parameters when the number of observations is large. Trans Am Math Soc 54:426–482

Wang H, Tsung F (2009) Tolerance intervals with improved coverage probabilities for binomial and Poisson variables. Technometrics 51:25–33

Wang CM, Hannig J, Iyer HK (2012) Fiducial prediction intervals. J Stat Plan Inference 142:1980–1990

Young D (2014) A procedure for approximate negative binomial tolerance intervals. J Stat Comput Simul 84:438–450