Confidence Intervals for Negative Binomial Random Variables of High Dispersion

David Shilane[*]   Steven N. Evans[†]

Alan E. Hubbard[‡]

[*]Stanford University, dshilane@stanford.edu

[†]University of California, Berkeley, evans@stat.berkeley.edu

[‡]University of California, Berkeley, hubbard@stat.berkeley.edu

Confidence Intervals for Negative Binomial Random Variables of High Dispersion[*]

David Shilane, Steven N. Evans, and Alan E. Hubbard

Abstract

We consider the problem of constructing confidence intervals for the mean of a Negative Binomial random variable based upon sampled data. When the sample size is large, it is a common practice to rely upon a Normal distribution approximation to construct these intervals. However, we demonstrate that the sample mean of highly dispersed Negative Binomials exhibits a slow convergence in distribution to the Normal as a function of the sample size. As a result, standard techniques (such as the Normal approximation and bootstrap) will construct confidence intervals for the mean that are typically too narrow and significantly undercover at small sample sizes or high dispersions. To address this problem, we propose techniques based upon Bernstein's inequality or the Gamma and Chi Square distributions as alternatives to the standard methods. We investigate the impact of imposing a heuristic assumption of boundedness on the data as a means of improving the Bernstein method. Furthermore, we propose a ratio statistic relating the Negative Binomial's parameters that can be used to ascertain the applicability of the Chi Square method and to provide guidelines on evaluating the length of all proposed methods. We compare the proposed methods to the standard techniques in a variety of simulation experiments and consider data arising in the serial analysis of gene expression and traffic flow in a communications network.

KEYWORDS: Bernstein's inequality, Chi Square distribution, confidence intervals, Gamma distribution, negative binomial distribution, serial analysis of gene expression (SAGE)

---

1   Introduction

Given a sample of n independent, identically distributed (i.i.d.) random variables with finite variance, the Central Limit Theorem (CLT) states that the distribution of the sample mean $\bar{X}$ is approximately Normal when the sample size n is large. As discussed in Rosenblum and van der Laan (2008), the Normal approximation and a bootstrap method are standard techniques used in the construction of confidence intervals for the mean μ even for moderately small sample sizes (e.g. $n = 30$). However, any application of Normal theory in these settings relies upon an assumption that n is large enough to render the differences between the distributions of $\bar{X}$ and the Normal inconsequential. At moderate sample sizes, this *CLT assumption* cannot be assured in the case of random variables with highly skewed distributions (Wilcox, 2005). In particular, we will demonstrate that a sample mean constructed from i.i.d. Negative Binomial random variables of high dispersion exhibits a probability mass function with an extremely heavy right tail. Moreover, the variability of estimates of the standard error of $\bar{X}$ provides an additional degree of uncertainty. In practice, researchers who rely on a relatively small number of independent samples (such as the investigation of (Lloyd-Smith et al., 2005) on the secondary transmission of infectious disease) should exercise caution to ensure that their conclusions are not greatly impacted by biased estimates of variability. Because $\bar{X}$ exhibits a skewed distribution, the Normal approximation may result in poor coverage and correspondingly poor inferences. Similarly, the bootstrap Bias Corrected and Accelerated (BCA) method (Efron and Tibshirani, 1994) also relies upon imbedded Normal theory that is impacted by skewness. This paper investigates the performance of these methods through simulation studies and proposes a variety of improvements based upon Bernstein's Inequality, a Gamma model, and the Chi Square (χ^2) distribution for the construction of confidence intervals for the mean of Negative Binomial random variables of high dispersion.

For any *significance level* $\alpha \in (0, 1)$, standard techniques for constructing $1 - \alpha$ confidence intervals often rely upon inverting hypothesis testing procedures under specific parametric assumptions (Casella and Berger, 1990; Clopper and Pearson, 1934; Crow and Gardner, 1959; Sterne, 1954). When these assumptions are satisfied, the resulting $1 - \alpha$ confidence intervals are *exact* in that the infimum coverage probability over all sample sizes is at least $1 - \alpha$ (Blyth and Still, 1963). Rosenblum and van der Laan (2008) investigate the use of exact methods in constructing confidence intervals when the assumptions underlying standard techniques are not valid. In particular, they employ tail probability bounds such as Bernstein's Inequality (Bernstein, 1934), Ben-

nett's Inequality (Bennett, 1962), and methods based on the work of Hoeffding (1963) and Berry-Esseen (Berry, 1941; Esseen, 1942). These bounds all require much weaker hypotheses that do not involve distributional assumptions on the data. As a result, Rosenblum and van der Laan (2008) are able to construct confidence intervals for a wide variety of parameters based upon the corresponding estimators' empirical influence curves. Such intervals will in general be more conservative than those based upon the Normal distribution but are not necessarily exact due to the influence curve approximation.

However, determining confidence intervals for the mean of i.i.d. Negative Binomial random variables of high dispersion is not so straightforward, particularly for small sample sizes. Even the relatively weak assumptions underlying methods such as the variant of Bernstein's Inequality employed by Rosenblum and van der Laan (2008) are not necessarily valid for Negative Binomials because the maximum deviation from the mean is not bounded. Since this assumption is violated in our setting, the resulting confidence intervals are not guaranteed to cover well. Therefore, we also investigate the Chi Square and Gamma distributions as practical alternatives to standard techniques and refinements to Bernstein confidence intervals that can lead to improved coverage of $1 - \alpha$ confidence intervals. We use simulation studies to compare the performance of these proposed techniques to those of standard methods for constructing confidence intervals for the mean of i.i.d. Negative Binomial random variables when the dispersion is high and the sample size n is small. We subsequently consider examples arising in the serial analysis of gene expression (SAGE) and network traffic flow data.

Section 2 reviews the Negative Binomial distribution. Section 3 describes Bernstein's Inequality's role in constructing $1 - \alpha$ confidence intervals for the mean, proves a limit theorem on the convergence of $\bar{X}$ to a Gamma distribution at large sample sizes and high dispersions, and also proposes the Chi Square distribution as an alternative approximation under suitable conditions. Section 4 summarizes a variety of simulation experiments that compare the coverage probabilities of the proposed methods, investigates the quality of approximations to the Negative Binomial's dispersion parameter, and examines potential refinements of the Bernstein method. Section 5 applies these techniques to data from the serial analysis of gene expression (SAGE) and network traffic flow studies. Finally, we will conclude the paper with a discussion in Section 6.

2   The Negative Binomial Distribution

A Negative Binomial distribution is conventionally used to compute the probability that a total of k failures will result before the rth success is observed when each trial is independent of all others and results in success with a fixed probability p. As described in Hilbe (2007), a Negative Binomial distribution may instead be parameterized in terms of a mean parameter $\mu = r\left(\frac{1}{p} - 1\right)$ and a dispersion parameter $\theta = r$. (We will adopt this alternative parameterization for the remainder of this paper.) Then, for any $\mu \in \mathbb{R}^+$ and $\theta \in \mathbb{R}^+$, the resulting probability mass function for the Negative Binomial random variable $X \sim NB(\mu, \theta)$ is

$$P(X = k) = \frac{\mu^k}{k!} \frac{\Gamma(\theta + k)}{\Gamma(\theta)[\mu + \theta]^k} \frac{1}{\left(1 + \frac{\mu}{\theta}\right)^\theta}, k \in \mathbb{Z}^+. \tag{1}$$

Equation (1) can be shown to converge to the probability mass function of a Poisson random variable with mean parameter μ as $\theta \to \infty$ (Hilbe, 2007). For this reason, the Negative Binomial may be considered as an over-dispersed Poisson random variable with the dispersion controlled by the value of θ. Negative Binomial models are useful as robust alternatives to the Poisson that allow the variance parameter to exceed the mean. For instance, smaller values of θ result in a higher dispersion by adding more weight to the right tail of the probability mass function, which necessarily results in a higher variance. When the value of θ is very small, the Negative Binomial distribution exhibits a high degree of skewness. As a result of the extreme dispersion of the Negative Binomial from the Poisson in this case, the sample mean $\bar{X}$ of n i.i.d. $NB(\mu, \theta)$ observations may not be reasonably close to the Normal in distribution for small values of n. Wilcox (2005) warns that standard confidence intervals based upon a Normal approximation may result in poor coverage in scenarios such as this.

3   Gamma, Chi Square, and Bernstein Confidence Intervals

3.1   The Gamma Model

We propose the Gamma distribution as an approximate distribution for the sample mean $\bar{X}$ of Negative Binomial random variables. The Gamma approximation may be established in a limit theorem using Laplace transforms. A

Negative Binomial random variable X_i with parameters μ and θ and PMF (1) has a Laplace transform given by:

$$F_{X_i}(\lambda) \equiv E\left[\exp\left(-\lambda X_i\right)\right] = \left(1 + (1 - e^{-\lambda})\frac{\mu}{\theta}\right)^{-\theta}. \tag{2}$$

Similarly, the sample mean $\bar{X}$ of n i.i.d. Negative Binomial random variables with parameters μ and θ has the Laplace transform

$$F_{\bar{X}}(\lambda) = \left(1 + (1 - e^{-\lambda/n})\frac{\mu}{\theta}\right)^{-\theta n}. \tag{3}$$

If θn converges to a positive constant γ as $n \to \infty$ and $\theta \to 0$, then the Laplace transform (3) converges to

$$F_{\bar{X}}(\lambda) = \left(1 + \frac{\lambda\mu}{\gamma}\right)^{-\gamma}. \tag{4}$$

Meanwhile, a Gamma random variable Y with a probability density function given by

$$f_{a,b}(x) = \frac{b^a x^{a-1} e^{-bx}}{\Gamma(a)} \tag{5}$$

has a Laplace transform of

$$F_Y(\lambda) = \left(1 + \frac{\lambda}{b}\right)^{-a}. \tag{6}$$

The Laplace transforms (4) and (6) are identical when $a = \gamma$ and $b = \frac{\gamma}{\mu}$. Therefore, the sample mean of Negative Binomial random variables converges to a Gamma distribution with shape parameter a and rate parameter b as $n \to \infty$ and $\theta \to 0$. Stated another way, the *Gamma assumption* requires that θn is sufficiently close to the limiting constant γ to ensure that the distribution of $\bar{X}$ is approximated well by the Gamma distribution. When this condition is satisfied, a Gamma confidence interval is expected to cover well.

Constructing a $1 - \alpha$ confidence interval using the Gamma distribution requires estimating the parameters $a = \gamma = \theta n$ and $b = \frac{\gamma}{\mu} = \frac{\theta n}{\mu}$ in terms of the estimates of μ and θ that are obtained from the data. Maximum likelihood estimates for these parameters may be obtained using numeric techniques, or the Method of Moments estimates $\hat{\mu} = \bar{X}$ and $\hat{\theta} = \bar{X}/((s^2/\bar{X}) - 1)$ may be employed. We will rely upon the latter choice as a default and provide a discussion of alternatives in Section 4.2. Once a and b are estimated, a $1 - \alpha$ confidence interval for μ is given by the $(\alpha/2)$nd and $(1 - \alpha/2)$th quantiles of the Gamma distribution.

3.2 The Chi Square Approximation

We also propose the Chi Square (χ^2) distribution as an approximate distribution for $\bar{X}$. Under the assumptions of Section 3.1, $\bar{X}$ approximately follows a Gamma distribution with parameters $a = \theta n$ and $b = \frac{\theta n}{\mu}$. The Chi Square family of distributions is a special case of the more general Gamma. The corresponding case occurs for the sample mean of Negative Binomial random variables when $\mu = 2n\theta$. If we consider the ratio quantity

$$\texttt{ratio} = \frac{\mu}{2n\theta}, \tag{7}$$

then, when ratio $= 1$, the Gamma parameters are $a = \theta n = \frac{\mu}{2}$ and $b = \frac{\theta n}{\mu} = \frac{1}{2}$, which collectively specify a Chi Square distribution with μ degrees of freedom.

In general, using a one–parameter Chi Square model to approximate a function of two–parameter Negative Binomial random variables lacks the two–parameter Gamma model's flexibility. This necessarily limits the use of Chi Square confidence intervals to situations in which the ratio quantity is reasonably close to 1. With this in mind, we emphasize that the Chi Square model's applicability should be carefully investigated before it is utilized in a particular context. However, when the Chi Square model is reasonable, it allows for the construction of a confidence interval based only upon the estimator $\bar{X}$ of μ. Other techniques such as the Gamma and Normal approximation also require an estimate s^2 of σ^2 that is considerably more variable than $\bar{X}$ at small sample sizes. We will propose some guidelines in Section 4.3 for the use of the Chi Square approximation based upon a study of the ratio quantity (7)'s relationship to the method's coverage probability. When applicable, Chi Square confidence intervals for μ may be constructed by estimating the degrees of freedom with the sample mean $\bar{X}$ and then computing the $(\alpha/2)nd$ and $(1 - \alpha/2)th$ percentiles of the corresponding Chi Square distribution.

The Chi Square distribution may also be used to construct confidence intervals for the success probability p under the traditional parametrization of the Negative Binomial distribution. On page 504, Casella and Berger (1990) demonstrate that the quantity $2pY$ converges in distribution to a Chi Square random variable with $2nr$ degrees of freedom, where Y is the sum of n i.i.d. Negative Binomial(r, p) random variables. When the number of successes r is known, a $1 - \alpha$ confidence interval for p may be written in terms of Y and the quantiles of the Chi Square distribution.

3.3 Bernstein's Inequality

Bernstein's Inequality (Bernstein, 1934) provides tail probability bounds on sums of independent random variables. Selecting an appropriate variant requires an examination of the assumptions underlying a particular study. More classical versions of Bernstein's Inequality were derived for uniformly bounded random variables, but Negative Binomial random variables are not bounded above. We will first address this problem through a version of Bernstein's Inequality that does not rely upon an assumption of boundedness. We will then provide an alternative methodology based upon this assumption. Although it does not directly apply to Negative Binomial random variables, this Bounded Bernstein method may be appropriate when a Negative Binomial model is considered as an approximate distribution for bounded data.

3.3.1 The Unbounded Bernstein Method

We will begin by deriving a confidence limit using a variant of Bernstein's Inequality that does not require an assumption of boundedness. As part of Lemma 8 on pages 366 and 367, Birge and Massart (1998) show that a knowledge of a random variable's moment generating function is sufficient to apply the classical version of Bernstein's Inequality given by Uspensky (1937). Suppose that Z_i, $1 \leq i \leq n$, are i.i.d. negative binomial with parameters μ and θ, and let $Y_i = Z_i - \mu$. Then,

$$\log E\left[\exp(yY_i)\right] = -\theta \log\left(\frac{(1 - e^y)\,\mu}{\theta} + 1\right) - y\mu. \tag{8}$$

From (Uspensky, 1937),

$$P\left[\sum_{i=1}^{n} Y_i \geq n\epsilon\right] \leq \exp\left[\inf_{y \geq 0}\left(-ny\epsilon + \sum_{i=1}^{n} \log E[e^{yY_i}]\right)\right]. \tag{9}$$

Substituting (8) into (9), we obtain the following probability bounds:

$$P\left(\sum_{i=1}^{n} Y_i > n\epsilon\right) \leq \left(\frac{\theta + \mu}{n\epsilon + \theta + \mu}\right)^{-\theta} \left(\frac{(\theta + \mu)(n\epsilon + \mu)}{\mu(n\epsilon + \theta + \mu)}\right)^{-n\epsilon - \mu}, \tag{10}$$

and

$$P\left(\sum_{i=1}^{n} Y_i < -n\epsilon\right) \leq \left(\frac{\theta + \mu}{n\epsilon + \theta + \mu}\right)^{-\theta} \left(\frac{\mu(n\epsilon + \theta + \mu)}{(\theta + \mu)(n\epsilon + \mu)}\right)^{n\epsilon + \mu}. \tag{11}$$

Equation (11) comes from using the classical inequality with Y_i replaced by $-Y_i$. With $\bar{X}$ serving as the empirical mean of Z_i, $1 \leq i \leq n$, the symmetry of Equations (10) and (11) imply that we can construct a $1 - \alpha$ confidence interval by setting the right side of either equation equal to $\alpha/2$. This amounts to constructing an interval of the form $\bar{X} \pm \epsilon$, where ϵ is the solution to the equation

$$\left(\frac{\theta + \mu}{n\epsilon + \theta + \mu} \right)^{-\theta} \left(\frac{(\theta + \mu)(n\epsilon + \mu)}{\mu(n\epsilon + \theta + \mu)} \right)^{-n\epsilon - \mu} = \alpha/2. \qquad (12)$$

However, solving Equation (12) is not easily amenable to analytic methods. For the purposes of implementation, we instead rely upon a simple numeric root–finding procedure that selects the best among candidate values of ϵ at evenly spaced intervals over a range (e.g. searching in increments of 0.1 from 0 to 100) and then searches within a small neighborhood of this candidate for an improved solution. Although this procedure is not guaranteed to provide a good approximation of the true value of ϵ that solves Equation (12), in practice it often performs reasonably well without requiring significant computation. However, further investigation of the solution to Equation (12) may lead to improved performance of the confidence interval.

3.3.2 The Bounded Bernstein Method

We will also construct a Bernstein confidence interval under a heuristic assumption of uniformly bounded data. For notational purposes, we will refer to the interval constructed for unbounded data as the *Unbounded Bernstein* method and that proposed for bounded data as the *Bounded Bernstein* procedure. As stated in van der Laan and Rubin (2005), suppose that $Z_1, \ldots, Z_n$ are independent random variables such that $Z_i \in [a, b] \in \mathbb{R}$ with probability one and $0 < \sum_{i=1}^{n} Var(Z_i)/n \leq \sigma^2$. Then, for all $\epsilon > 0$,

$$P \left(\frac{1}{n} \sum_{i=1}^{n} (Z_i - E[Z_i]) > \epsilon \right) \leq \exp \left[\frac{-1}{2} \left(\frac{n\epsilon^2}{\sigma^2 + \epsilon(b - a)/3} \right) \right], \qquad (13)$$

which in turn implies that

$$P \left(\frac{1}{n} | \sum_{i=1}^{n} (Z_i - E[Z_i])| > \epsilon \right) \leq 2 \exp \left[\frac{-1}{2} \left(\frac{n\epsilon^2}{\sigma^2 + \epsilon(b - a)/3} \right) \right]. \qquad (14)$$

We seek a $1 - \alpha$ confidence interval (CI) of the form $\bar{X} \pm \epsilon$ so that the distance ϵ extends sufficiently far to ensure with a probability of at least $1 - \alpha$ that the experiment will result in a confidence interval containing the true mean μ. The appropriate value of ϵ may be selected by setting the right hand side of Equation (14) equal to α. Then, by applying the Quadratic Formula, the value of ϵ is given by

$$
\epsilon = \frac{\frac{-2}{3}(b-a)\log(\alpha/2) \pm \sqrt{\frac{4}{9}(b-a)^2[\log(\alpha/2)]^2 - 8n\sigma^2\log(\alpha/2)}}{2n}. \tag{15}
$$

We will select the value of ϵ that adds the square root in Equation (15). The appeal of using the Bounded Bernstein method in the construction of $1 - \alpha$ confidence intervals is that it only requires three assumptions (Rosenblum and van der Laan, 2008): (i) all observations are independent, (ii) the maximum deviation from the mean is bounded by a known constant, and (iii) the variance is bounded by a known constant. By contrast, the CLT assumption underlying the Wald method's Normal approximation and the Gamma assumption underlying the Gamma and Chi Square methods are considerably stronger requirements. Therefore, Bounded Bernstein confidence intervals may be applied more widely than parametric methods. Similarly, confidence intervals may also be constructed from other tail bounds such as Bennett's Inequality (Bennett, 1962, 1963). Hoeffding's Inequality (Hoeffding, 1963) may in fact be applied under only assumptions (i) and (ii). The Berry-Esseen Inequality (Berry, 1941; Esseen, 1942, 1956; van Beek, 1972) also requires just the three above assumptions but only results in non-vacuous confidence intervals for $n \geq 1024$ (Rosenblum and van der Laan, 2008), which necessarily limits its application as an alternative to the Normal approximation.

The formulation for ϵ in Equation (15) depends upon the data's bounding range $[a, b]$ and the variance σ^2. In the case of i.i.d. observations of Negative Binomial random variables, the lower bound is $a = 0$ because these variables draw from a non-negative sample space. However, Negative Binomial variables are unbounded, which violates assumption (ii) underlying Bernstein's Inequality in the Bounded Bernstein method. Therefore, the Bounded Bernstein confidence interval is only appropriate if the Negative Binomial distribution is considered as an approximate distribution for bounded data. Furthermore, in small sample sizes, the data-based unbiased estimate s^2 of the variance σ^2 exhibits a high degree of variability and therefore may greatly underestimate the value of σ^2. Without accurate upper bounds for b and σ^2, Bounded Bernstein $1 - \alpha$ confidence intervals for μ are not necessarily exact. Rosenblum

and van der Laan (2008) provide some practical recommendations to address these concerns by relying upon known information about b and σ^2 collected in previous studies. Other possibilities include selecting b via a heuristic such as the 99.99th percentile of the Negative Binomial distribution with μ and θ estimated from the data. The only strict requirement for the Bounded Bernstein method is that we select a value of b at least as large as the maximum observed value. By default, we will rely upon the following data–based heuristic:

$$b = \frac{n+1}{n} \max\left(X_1, \ldots, X_n\right).$$

(16)

This heuristic was selected to provide an estimated upper bound in terms of the data and the sample size n. This choice of b will be considered in the simulation studies of Section 4, and Section 4.4 will examine b's impact on the coverage probability of corresponding Bounded Bernstein $1 - \alpha$ confidence intervals.

Although the Bounded Bernstein method is not justified for unbounded data, the Negative Binomial model is often nonetheless considered as an approximate distribution for bounded data. The case studies of Section 5 provide examples in the serial analysis of gene expression and an examination of traffic flow in a communications network in which the underlying data are bounded but are reasonably approximated by Negative Binomial models.

4 Simulation Studies: Comparing the Wald, Bootstrap, Chi Square, Gamma, and Bernstein Confidence Intervals for μ

4.1 Coverage Probabilities and Lengths of the Proposed Methods

We designed two simulation studies to compare the proposed methods for efficacy. The first simulation compared the Wald (Normal Approximation), bootstrap, Chi Square, Gamma, and both the Bounded and Unbounded Bernstein methods of constructing $1 - \alpha$ confidence intervals for μ. We selected the computational parameter sets $\mu = \{5, 10\}$, $\theta = \{0.1, 1, 10, 10000\}$, and $n = \{10, 20, \ldots, 100\}$, which are summarized in Table 1. Each combination of values for μ, θ, and n led to a unique and independent simulation experiment. We selected these values of θ to allow for both high dispersion (when θ is low) and low dispersion (when θ is high), and we considered both small and moder-

ate values of n to determine cut-off points at which standard methods like the Wald and bootstrap would overtake the proposed methods in terms of coverage. Each experiment consisted of 10,000 independent trials, and on each trial we generated n i.i.d. $NB(\mu, \theta)$ random variables in the **R** statistical programming language. With $\alpha = 0.05$, we then computed 95% confidence intervals for μ based upon the data collected in the trial. The Wald method constructed confidence intervals by adding and subtracting 1.96 estimated standard errors to the sample mean. Bootstrap confidence intervals were computed according to the BCA method (Efron and Tibshirani, 1994) based upon $B = 10,000$ resamplings from the data collected in each trial. We estimated the coverage probability of each method at each choice of parameters by computing the empirical proportion of trials within the experiment that resulted in a confidence interval containing the true value of μ.

We then undertook a second independent simulation to examine a greater variety of small θ values and sample sizes. As summarized in Table 1, we considered values of $\theta \in \{0.025, 0.05, 0.075, 0.1, 0.2, 0.3, 0.4, 0.5\}$ at sample sizes $n \in \{5, 10, \ldots, 100\}$ while maintaining the μ values and number of trials as in the first simulation. In this second simulation, the bootstrap method was not employed because of its heavy computational requirements and its observed similarity in coverage to the Wald method in the first simulation. Computing a single coverage probability with the bootstrap with $10,000$ resamplings in each of 10,000 size n data sets requires generating $n * 10^8$ random numbers. In total, Simulations 1 and 2 required nearly a week of continuous computation, of which all but a few hours were spent on the bootstrap.

Figures 1–6 provide summaries of each method's coverage and length across the simulation experiments. In judging the quality of a method's performance, we adopt the view that its coverage probability is of primary concern and that length is a secondary characteristic that can be used to choose among methods that produce similar results. Because a confidence interval is interpreted as a plausible range of values for μ in inferential settings, a shorter interval is typically preferred, but this is only the case so long as the method can be shown to cover reliably. Therefore, in comparing the simulation results obtained by the proposed and standard methods, we primarily seek methods that can produce coverages that are reasonably close to the desired level of $1 - \alpha$.

As expected, the bootstrap (Figure 1) and Wald (Figure 2) methods appear to cover well at large sample sizes. When $\theta \geq 1$, the coverage probability begins to exceed 0.9 even for sample sizes as small as 20. However, these standard techniques perform considerably worse at higher dispersions. For instance, when $\mu = 5$ and $\theta = 0.05$ or smaller, even a sample size of 100 is insufficient for the Wald method to exhibit a coverage probability of at least

0.85. Likewise, at the more moderate case of $\mu = 5$ and $\theta = 0.2$, a sample size of 30 also leads to a coverage of less than 0.85.

Meanwhile, the Bounded Bernstein (Figure 3) and Chi Square (Figure 4) confidence intervals appear to improve considerably on the standard techniques in terms of coverage at small sample sizes and high dispersions. When $\mu = 5$ and $\theta = 0.05$, the Bounded Bernstein confidence interval crosses a coverage threshold of 0.9 and the Chi Square that of 0.95, both as early as $n = 60$. These methods appear to exhibit greater coverages uniformly across the small values of θ considered in Simulation 2. Furthermore, when $\mu = 5$ and $\theta = 0.2$, the Bounded Bernstein and Chi Square methods both cover above 0.95 at $n = 30$.

However, the proposed Gamma (Figure 5) and Unbounded Bernstein (Figure 6) methods appear to perform erratically. The Unbounded Bernstein's coverages do not necessarily improve as the sample size increases. The Gamma method covers similarly to the Wald method in some regions (though it generally lags in coverage by a few percent), but its performance drops considerably at larger values of θ. While both techniques are grounded in theory, we can postulate a number of reasons why these methods may not cover well. The tail probability bound used in the variant of Bernstein's Inequality considered for the Unbounded version is not necessarily tight, and the numeric root–finding method used to approximate the value of ϵ that solves Equation (12) may introduce additional error. By contrast, the Gamma method only applies if n is sufficiently large and θ sufficiently small to ensure that θn is close to a limiting constant γ. This Gamma assumption appears to be violated at the larger values of θ considered in the simulations. Moreover, both the Unbounded Bernstein and Gamma methods rely upon accurate estimates of the dispersion parameter θ, which are typically unreliable in many of the situations encountered in the simulation studies. Section 4.2 will examine this issue in greater detail. With this said, it does appear that the Gamma method covers reasonably well when both θ is small and n is large, which correspond to the lower right corner of the coverage plots in Figure 5. Section 4.5 will investigate the Gamma's performance in large n and small θ settings in greater detail.

The candidate methods' coverages in the simulation experiments may be compared directly by examining Figures 1–6. In comparing any two estimated coverage probabilities at a specific value of n, μ, and θ, the difference in proportions has a margin of error of no more than 1.39% for a two-sided test based upon the simulation experiment's 10,000 trials. This worst-case error margin is obtained under the extreme assumption that the true coverage of each method is actually 50%. If the coverage of each method is actually 95%,

then this margin of error drops to 0.6%. Any observed difference that is larger than the margin of error may be considered significant at the 5% level.

Meanwhile, each method's coverage summaries are accompanied by corresponding plots depicting the average length of the confidence intervals in the simulation experiments. As expected, the Chi Square and Bounded Bernstein methods generally produce wider intervals than the Wald and bootstrap under high dispersion, and this increase in length corresponds to greater coverages. However, the Chi Square interval does not decrease in length as the sample size increases because its degrees of freedom is only specified by the sample mean. For fixed values of μ and θ, the ratio quantity of Equation (7) decreases for larger sample sizes. Taken together, these facts imply that ratios below one will result in a Chi Square confidence interval that over–covers the mean. We will further substantiate this claim in Section 4.3.

Figure 7 provides concrete recommendations on which confidence intervals perform best across all values of μ, θ, and n considered in the two simulation experiments. These recommendations are based upon which method exhibited a coverage closest to 95% in the simulation experiments. It is important to note that these recommendations allow for under-coverage; in the case of $\mu = 5$, $\theta = 0.5$, and $n = 100$, the Wald Method covers with an estimated probability of 93.53% and the Gamma at 92.34% whereas the Bounded Bernstein and Chi Square methods cover at 99.39% and 100%, respectively. Also, it is important to remember that none of the proposed methods perform particularly well when both n and θ are extremely small. We also provide corresponding recommendation plots for length that indicate approximate cut-offs at which the Wald confidence interval shrinks to a smaller length than that of the Chi Square method. Because of their erratic coverages, the recommendation plot for length does not incorporate results obtained from the Gamma or Unbounded Bernstein methods.

One caveat to the simulation results presented here is the special case of a data set containing all zeros. Figure 8 shows the percentage of all–zero data sets generated in the simulation experiments. When such data arose, we adopted the convention that all methods should produce a condence interval containing only the point zero, and therefore the interval would not cover the mean in this circumstance. This only affected the results at small values of both n and θ; for instance, in the experiment with $\mu = 10$, $\theta = 0.025$, and $n = 5$, 51.44% of all data sets produced contained only zeros. However, most scientic studies would not produce a condence interval for μ based upon a data set of all zeros. If one is willing to condition on observing non–zero data, then the coverage probabilities at these small values of n and θ may be adjusted by computing the proportion of intervals that covered μ among the non–zero

data set. For instance, the Bounded Bernsteins total coverage was 33.66% when $\mu = 10$, $\theta = 0.025$, and $n = 5$, and so its coverage among the non–zero data sets is actually $100\% * 3366/(10000 - 5144) = 69.316\%$.

Sim.	μ	θ	n	trials
1	$\{5, 10\}$	$\{0.1, 1, 10, 10000\}$	$\{10, 20, \ldots, 100\}$	10000
2	$\{5, 10\}$	$\{0.025, 0.05, 0.075, 0.1, 0.2, \ldots, 0.5\}$	$\{5, 10, \ldots, 100\}$	10000

Table 1: Parameter values for μ, θ, n, and the number of trials in the two simulation experiments of Section 4. The first simulation compared the Unbounded and Bounded Bernstein, Chi Square, Gamma, Wald, and bootstrap confidence interval methods at each combination of the first set of parameter values. The second simulation compared the Wald, Unbounded and Bounded Bernstein, Gamma, and Chi Square methods at each combination of the second set of parameter values.

4.2   The Accuracy of θ Estimates

Both the Gamma model and the Unbounded Bernstein method rely upon an estimate of θ to produce a $1 - \alpha$ confidence interval for μ. In addition to the Method of Moments estimator $\hat{\theta} = \bar{X}/((s^2/\bar{X}) - 1)$, Piegorsch (1990) and Clark and Perry (1989) have proposed iterative maximum likelihood estimation (MLE) procedures. Aragón et al. (1992) and Ferreri (1997) provide conditions for the existence and uniqueness of the MLE. Meanwhile, Pieters et al. (1977) compares an MLE procedure to the Method of Moments at small sample sizes. The general consensus of these previous studies is that estimating θ is a difficult problem; MLE methods appear to break down when the estimate s^2 of σ^2 is less than or equal to the estimate $\bar{X}$ of μ. Although the MLE estimator was preferred, implementations such as that in the **glm.nb** function of the **R** statistical programming language tend to produce computational errors that prevented its application in the simulation experiments of the previous section. Similarly, the Method of Moments estimator frequently results in a non–positive approximation of the strictly positive parameter θ in this situation. For the purposes of the simulations, we chose to handle this issue by truncating all non–positive estimates of θ to the value of 0.001 before applying the confidence interval procedures.

Figure 9 provides summary information about the Method of Moments estimates (without truncation) of θ over the range of experiments conducted in Simulations 1 and 2. For each combination of μ, θ, and n in the simulations, we provide the average estimation error across the 10,000 simulated data sets.
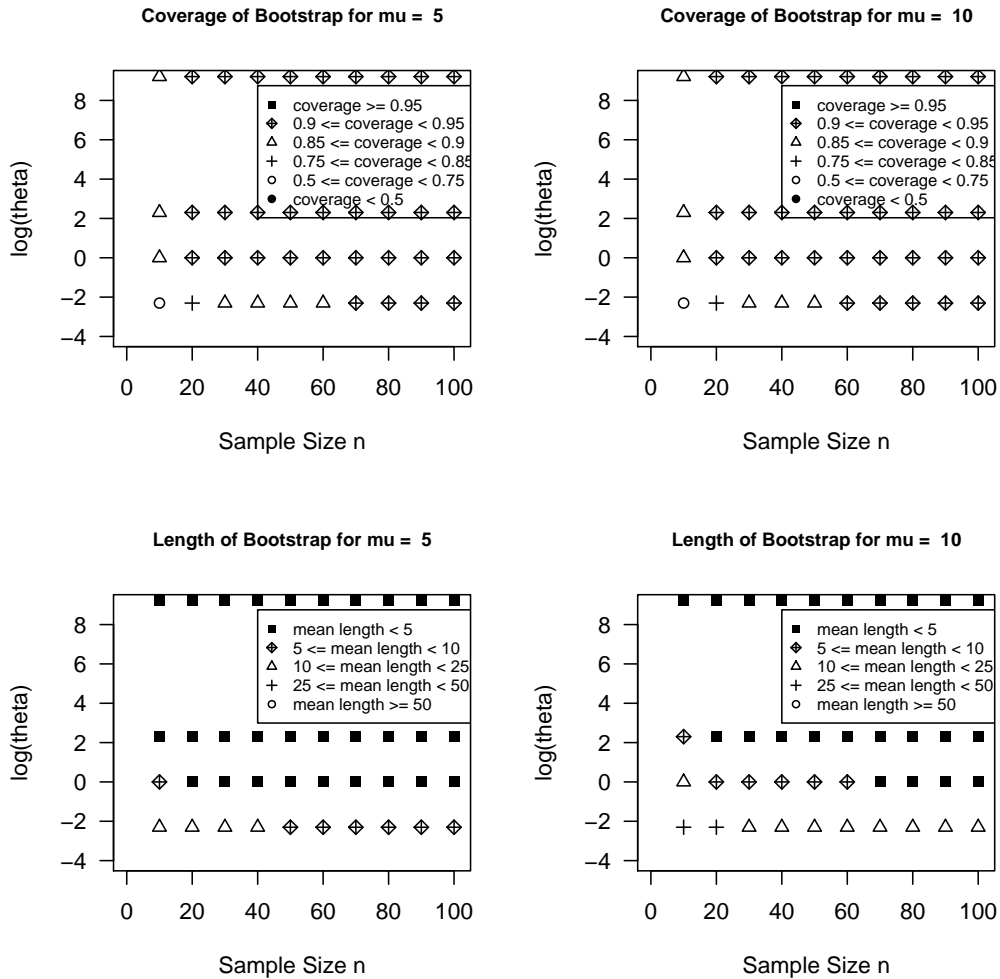
Coverage of Bootstrap for mu = 5

Legend:
- ■ coverage >= 0.95
- ⬖ 0.9 <= coverage < 0.95
- △ 0.85 <= coverage < 0.9
- + 0.75 <= coverage < 0.85
- ○ 0.5 <= coverage < 0.75
- ● coverage < 0.5

Coverage of Bootstrap for mu = 10

Legend:
- ■ coverage >= 0.95
- ⬖ 0.9 <= coverage < 0.95
- △ 0.85 <= coverage < 0.9
- + 0.75 <= coverage < 0.85
- ○ 0.5 <= coverage < 0.75
- ● coverage < 0.5

Length of Bootstrap for mu = 5

Legend:
- ■ mean length < 5
- ⬖ 5 <= mean length < 10
- △ 10 <= mean length < 25
- + 25 <= mean length < 50
- ○ mean length >= 50

Length of Bootstrap for mu = 10

Legend:
- ■ mean length < 5
- ⬖ 5 <= mean length < 10
- △ 10 <= mean length < 25
- + 25 <= mean length < 50
- ○ mean length >= 50

Figure 1: Simulation results for the Bootstrap method. The plots include summaries of coverage and length for each value of μ. The Bootstrap method was only considered for the first simulation experiment due to its computational requirements.
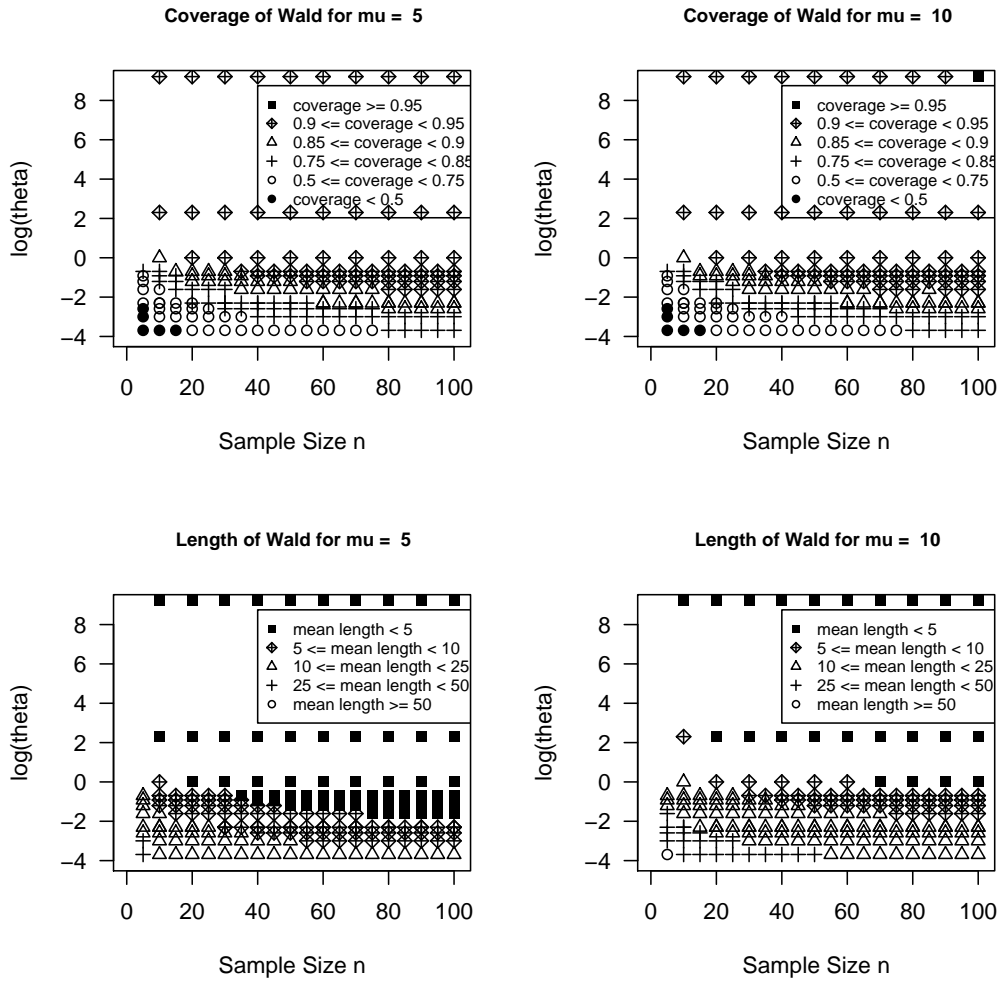
Figure 2: Simulation results for the Wald method. The plots include summaries of coverage and length for each value of μ. Results from the two simulation experiments are concatenated onto a single plot.
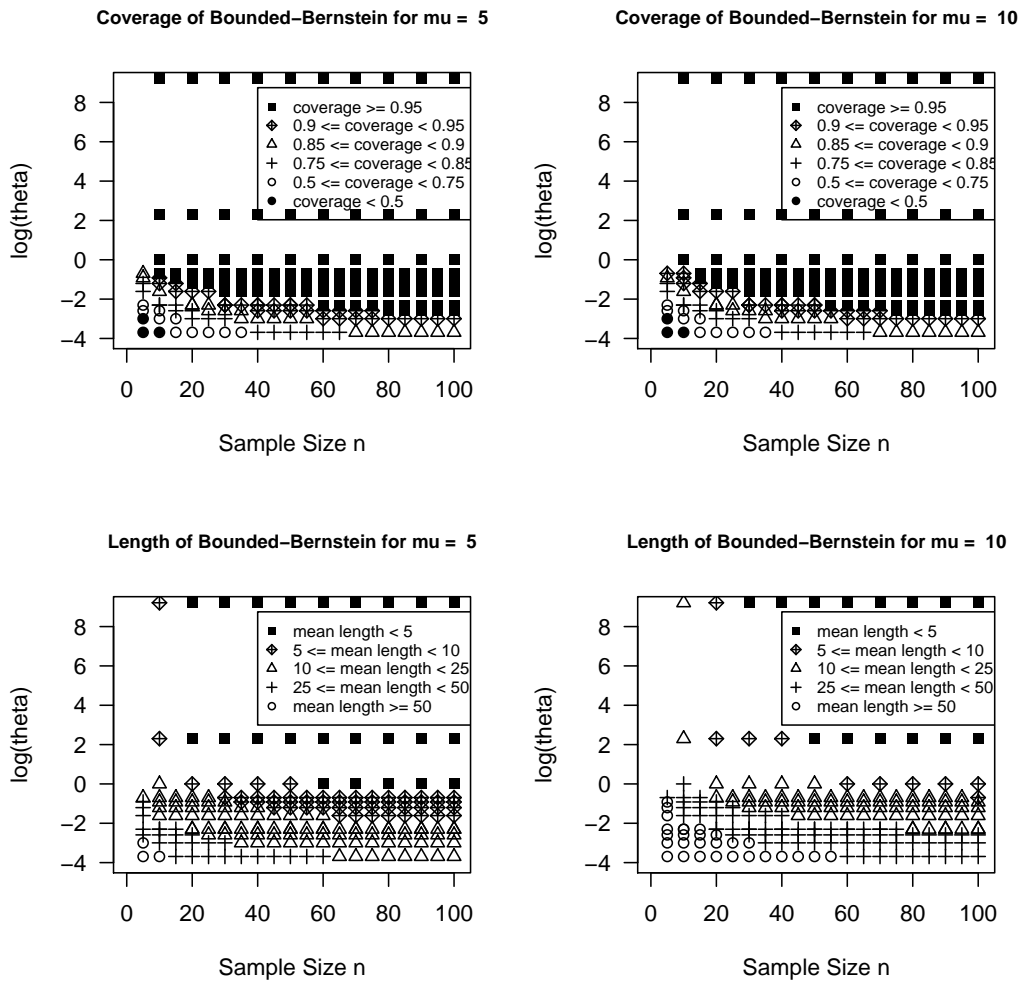
Figure 3: Simulation results for the Bounded Bernstein method. The plots include summaries of coverage and length for each value of μ. Results from the two simulation experiments are concatenated onto a single plot.
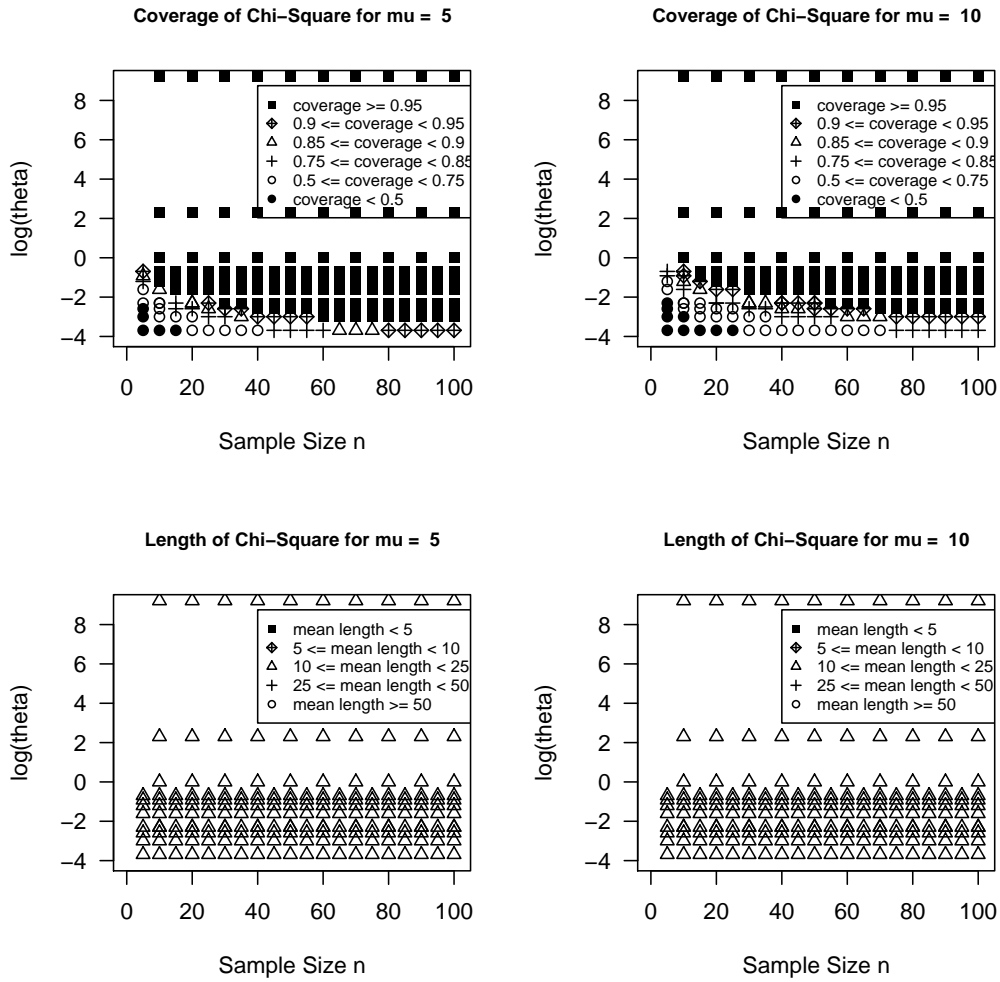
Figure 4: Simulation results for the Chi Square method. The plots include summaries of coverage and length for each value of μ. Results from the two simulation experiments are concatenated onto a single plot.
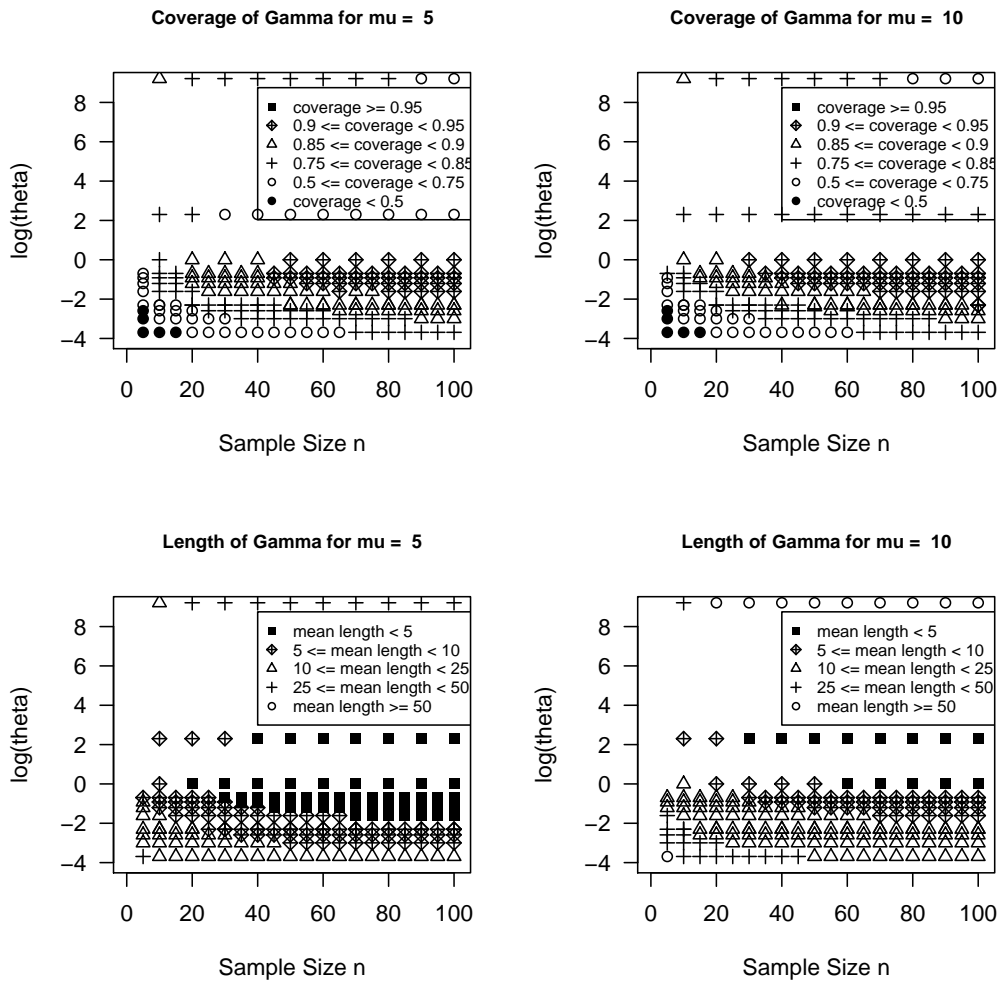
Figure 5: Simulation results for the Gamma method. The plots include summaries of coverage and length for each value of μ. Results from the two simulation experiments are concatenated onto a single plot.
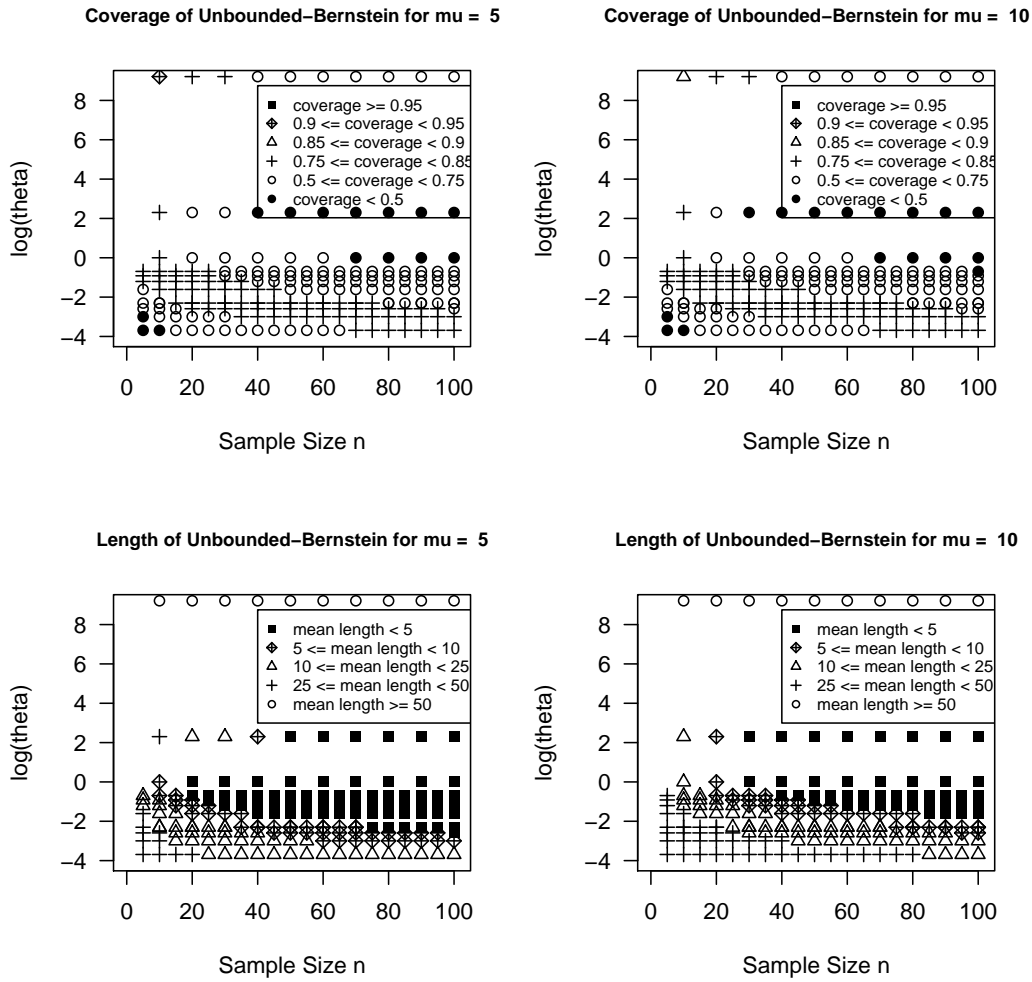
Figure 6: Simulation results for the Unbounded Bernstein method. The plots include summaries of coverage and length for each value of μ. Results from the two simulation experiments are concatenated onto a single plot.
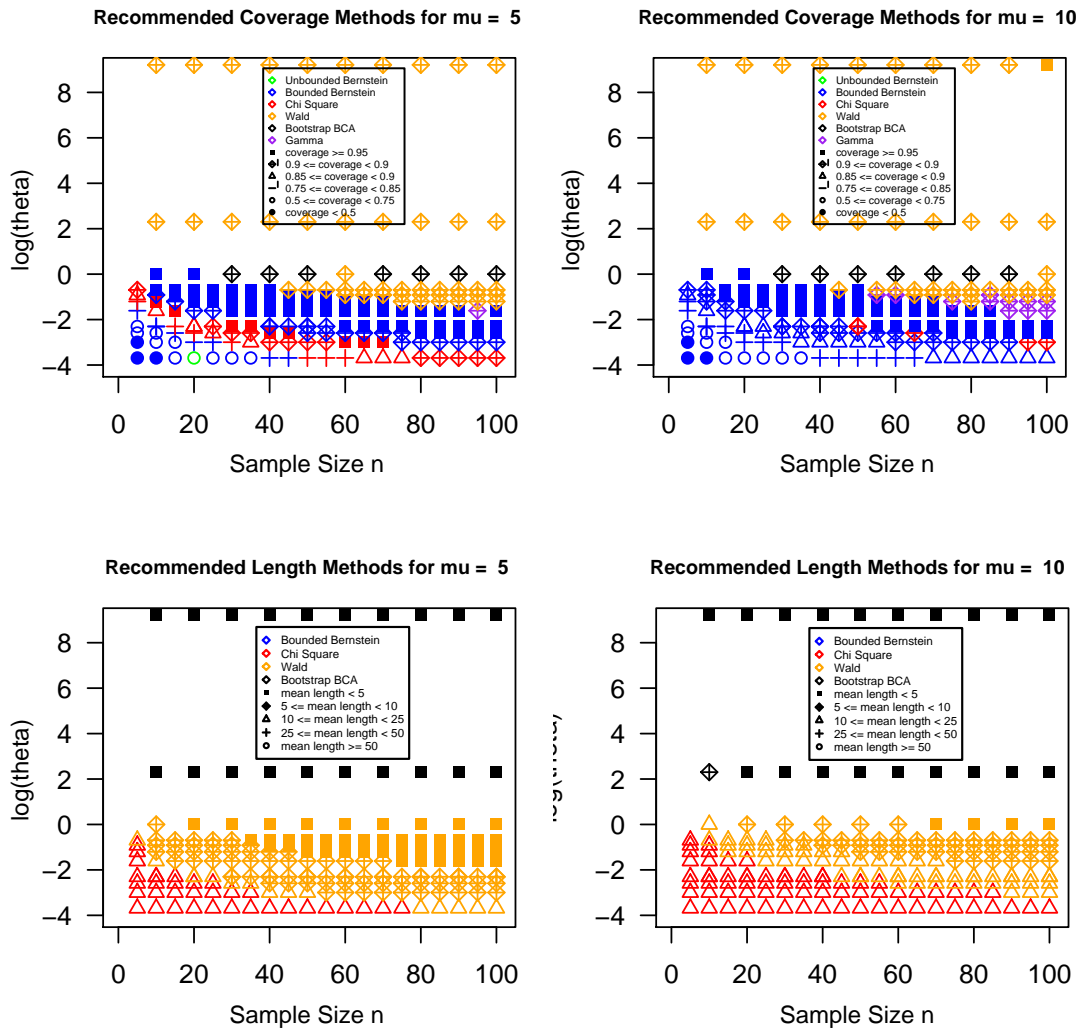
Figure 7: Recommendation plots based upon coverage and length in the simulation experiments. Selections were made according to the method whose coverage was closest to 95%. Length recommendations were only chosen among the Wald, Bootstrap, Chi Square, and Bounded Bernstein methods.
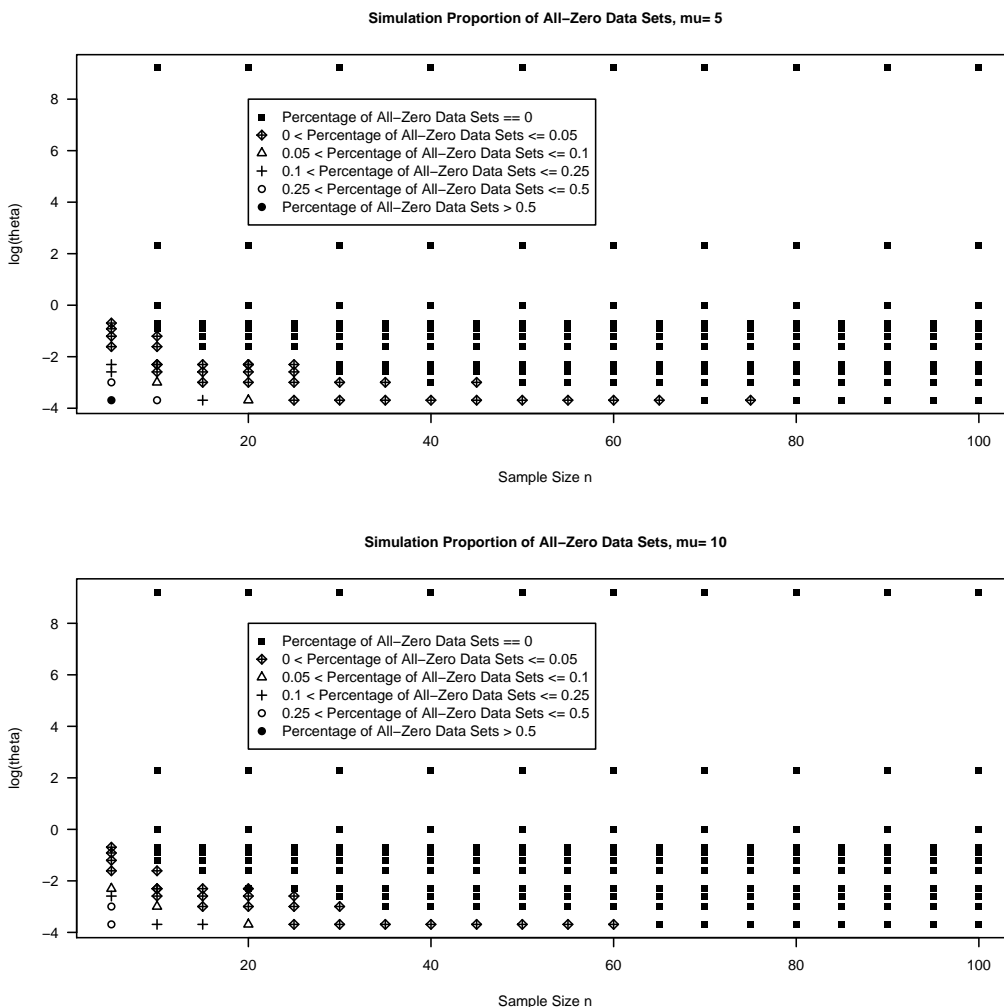
Simulation Proportion of All–Zero Data Sets, mu= 5

Simulation Proportion of All–Zero Data Sets, mu= 10

Figure 8: The proportion of all–zero data sets in the simulation experiments.

For small values of θ, the median estimate is reasonably reliable even at small sample sizes, but the average and standard deviation can be greatly affected by extreme data sets. In general, we observed that a sample size of at least 30 or 40 is required to ensure that the Method of Moments estimator is not highly vulnerable to extreme data sets and may need to be as large as 85 to ensure reliability in the smallest values of θ studied in the simulations.

While MLE or other improved estimators of θ may lead to stronger performance of the Gamma and Unbounded Bernstein confidence intervals, it appears that their simulation results were not greatly impacted by the selec-

tion of the Method of Moments estimator. At moderate sample sizes, the Method of Moments typically produced estimates that were reasonably close to the true value of θ, and its more erratic performance at smaller sample sizes corresponds to cases in which the MLE is also expected to have problems.

4.3   The Applicability of the Chi Square Method

Section 3.2 introduced the Chi Square approximation to the Gamma model of Section 3.1. This special case occurs when the ratio quantity of Equation (7) is equal to one. The principle advantage of the Chi Square method is that it allows for the construction of $1 - \alpha$ confidence intervals without relying upon the extremely variable estimator s^2 of the variance σ^2, which is especially useful at small sample sizes. However, this implies that the length of the interval is only dependent upon the sample mean, so the Chi Square interval's length will be approximately the same regardless of the value of θ. Furthermore, for fixed values of μ and θ, the ratio quantity decreases as a function of sample size. Therefore, we expect the Chi Square method to undercover for ratios above 1 and over–cover for ratios below one. We are interested in determining how robust the Chi Square approximation is to deviations of this ratio.

We conducted a third simulation to gain insight on the Chi Square's coverage at a variety of ratio quantities. Each experiment consisted of selecting μ uniformly on $(1, 50)$, θ uniformly on $(0, 1)$, and the sample size n uniformly on the integers in $\{5, 6, \ldots, 150\}$. For each combination of n, μ, and θ, we randomly generated $10,000$ data sets of n i.i.d. $NB(\mu, \theta)$ random variables, applied the Chi Square method to each data set, and estimated the method's coverage probability by the empirical proportion of Chi Square 95% intervals that contained the selected value of μ. We conducted a total of $100,000$ such experiments to collect data at a wide range of ratio quantities.

Figure 10 displays boxplots of the ratio distribution for the simulation data partitioned into coverage groups. For magnification purposes, the plot restricts attention to the cases that resulted in a coverage of at least 50%. The 4% of the simulations not pictured generated extremely large ratio quantities: approximately 1% of all simulations resulted in a ratio larger than 30, and the maximum observed value was $75,170$. Among the simulations with ratios less than 8, the correlation between the ratio quantity and the Chi Square method's coverage probability was -0.98. As expected, Figure 10 suggests that ratio values less than 1 typically over–cover the mean while ratios below 1 tend to undercover. It also appears that the Chi Square method will cover at a rate of at least 80% when the ratio quantity is below 2.
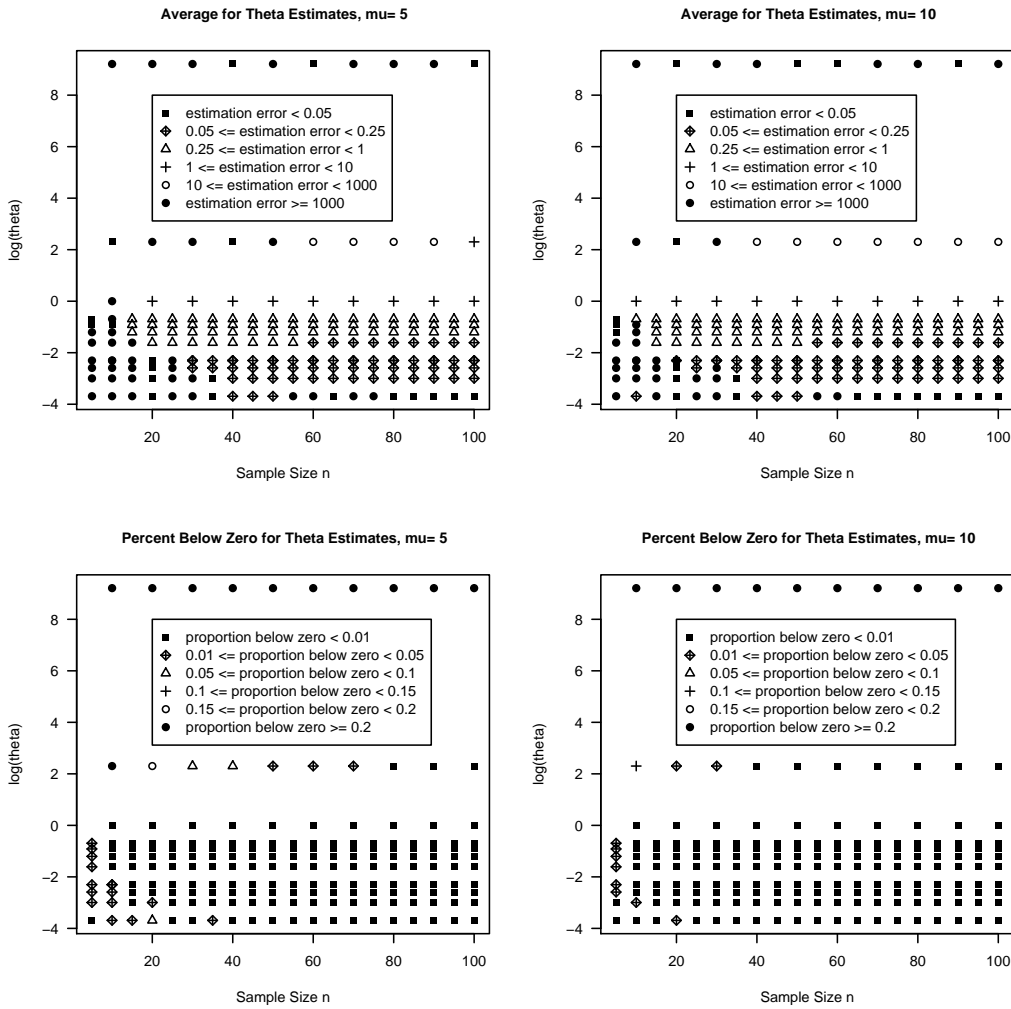
Figure 9: Quality measurements for the Method of Moments estimates of θ obtained in the simulation experiments. We present the average absolute estimation error along with the percentage of θ estimates that fell below zero in each experiment.

With these observations in mind, it appears that an empirical estimate of the ratio quantity of Equation (7) can provide some insight into the applicability of the Chi Square method. Ratios less than 1 will typically result in a confidence interval that over–covers because it is too wide, whereas ratios greater than 1 indicate intervals that are too narrow and will undercover μ. Because of its strong relationship with coverage, the ratio quantity can be used

23

as a guide in selecting among candidate confidence intervals even when the Chi Square procedure is not applicable. For instance, if the ratio quantity is 3, the Chi Square method might only produce an interval with a coverage of 75%, but the length of this interval can be used as a reference in the comparison of other candidate procedures so that a wider interval is ultimately selected. Furthermore, when the ratio quantity is less than 1, the Chi Square interval's length can be viewed as a maximum range so that any wider interval may be immediately excluded.

4.4 The Upper Bound b in Bounded Bernstein Confidence Intervals

Although it is intended to be an exact method, the simulation results of Section 4 show multiple examples in which the Bounded Bernstein confidence intervals as implemented result in estimated coverage probabilities well below 95%. These results were based upon the unbiased estimate s^2 of the variance σ^2 and an upper bound b given by Equation (16). Because s^2 is highly variable at small sample sizes, it may result in values much smaller than σ^2. Likewise, it is unclear how to optimally select the value of b because the Negative Binomial random variables in question are unbounded and highly skewed. If we are primarily concerned with producing exact confidence intervals through the Bounded Bernstein method, there is no harm in greatly overestimating σ^2 and b. However, in practice we would prefer to construct intervals that are as narrow as possible while maintaining the minimum desired coverage. In this section we will examine the impact of selecting various choices of b given an estimate s^2 of σ^2 in the context of the two simulation studies of Section 4.1.

A fourth simulation was conducted to repeat the simulations of Section 4.1 for only the Bounded Bernstein method. In this case, a variety of b values ranging up to 10^6 were substituted in place of the heuristic of Equation (16) used previously. Figure 11 depicts the best available value of b and the resulting coverage for each combination of μ, θ, and n for each simulation experiment. In general, larger values of the upper bound b increase the length and coverage of the Bounded Bernstein confidence interval, and sufficiently large values of b can be found to produce an exact method. Especially at high dispersions and low sample sizes, it appears that the heuristic (16) choice of b is considerably smaller than would be required to cover appropriately. Based upon our observation of the simulation results, it appears to be reasonable to roughly double the value of b given by the heuristic (16). For instance, when $\mu = 5$, $\theta = 0.05$, and $n = 60$, the Bounded Bernstein method resulted in a coverage of 90.85%, which roughly corresponds to a b value of 40. Meanwhile, a b value of 80 results in a coverage of 96.74%.

Ratio Distribution by Coverage Group
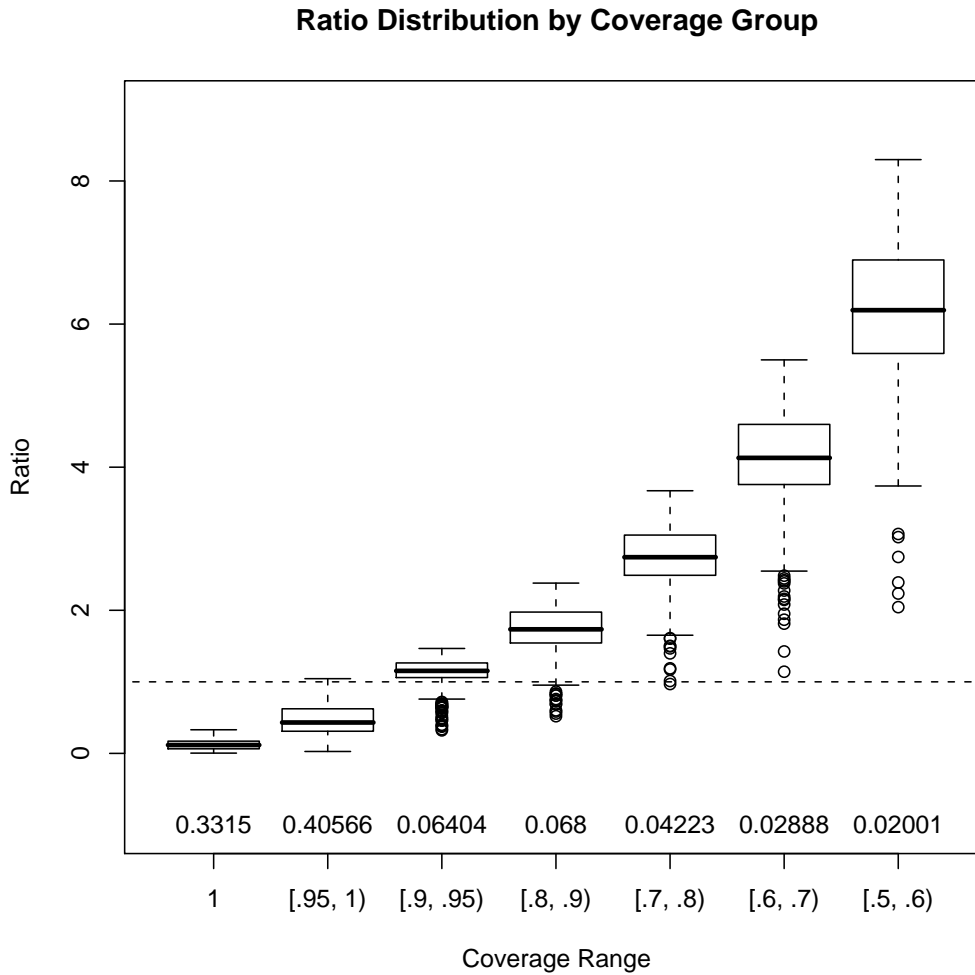
Figure 10: The distribution of ratio quantities (7) by coverage group for Chi Square 95% confidence intervals. This box plot was constructed from data collected in the simulation of Section 4.3. The dashed lined represents a ratio quantity of 1, at which the Chi Square approximation to the distribution of $\bar{X}$ is exact. The proportion of simulations falling into each category are displayed below the box plots. The remaining 4% of the simulations that produced coverages under 50% are not displayed for magnification purposes.

While refinements of the selection of b can lead to improvements over the heuristic of Equation (16), it is unclear how this selection should be adjusted as a function of sample size and the Negative Binomial parameters. However, such adjustments are considerably more straightforward than revising the estimate of variability. For a given estimate s^2 of the variance σ^2, a sufficiently large selection of the upper bound b will result in a coverage at or above the desired level. Therefore, we recommend that the researcher fine-tune the Bounded Bernstein confidence interval in a situation-dependent manner.

4.5 Large Sample Performance of the Gamma Method at High Dispersions

Although the Gamma method is grounded in the limit theorem of Section 3.1, its performance in the simulation studies of Section 4.1 was quite erratic. In most of these experiments, the Gamma method either lagged the Wald in coverage or produced unreliable confidence intervals. We previously speculated that this poor performance was due in part to the difficulty of accurately estimating the dispersion parameter θ at small values. However, at many combinations of μ and n, the Gamma method's coverage actually grew worse for larger values of θ. This phenomenon is easily explained by returning to the underlying Gamma assumption upon which the method is based. In particular, the Gamma approximation to the distribution of the sample mean $\bar{X}$ relies upon n growing sufficiently large and θ sufficiently small to ensure that θn is reasonably close to its limiting constant γ. Therefore, as θ grows larger, this Gamma assumption becomes less reasonable. Moreover, the Gamma method appeared to perform best in the simulation experiments corresponding to the largest values of n and smallest values of θ considered (e.g. the lower right corner of Figure 5).

With this in mind, we designed a fifth simulation to compare the performance of the Gamma and Wald methods at a variety of small values of θ and large values of n. In this setting, the Gamma assumption should be met, and it is a reasonable question as to what sample sizes are sufficient to overcome extremely small values of θ to ensure that the CLT assumption underlying the Wald method is also reasonable. The simulation was modeled after those undertaken in Section 4.1. The combinations of n, μ, and θ are displayed in Table 2. Coverages were estimated based upon the empirical proportion of confidence intervals across $10,000$ trials. The Method of Moments estimator of θ was employed with a minimum value of 10^{-6} imposed to ensure non–negativity. All coverages were left unadjusted in the case of data sets consisting of all zeros. However, because of the large sample sizes, only the
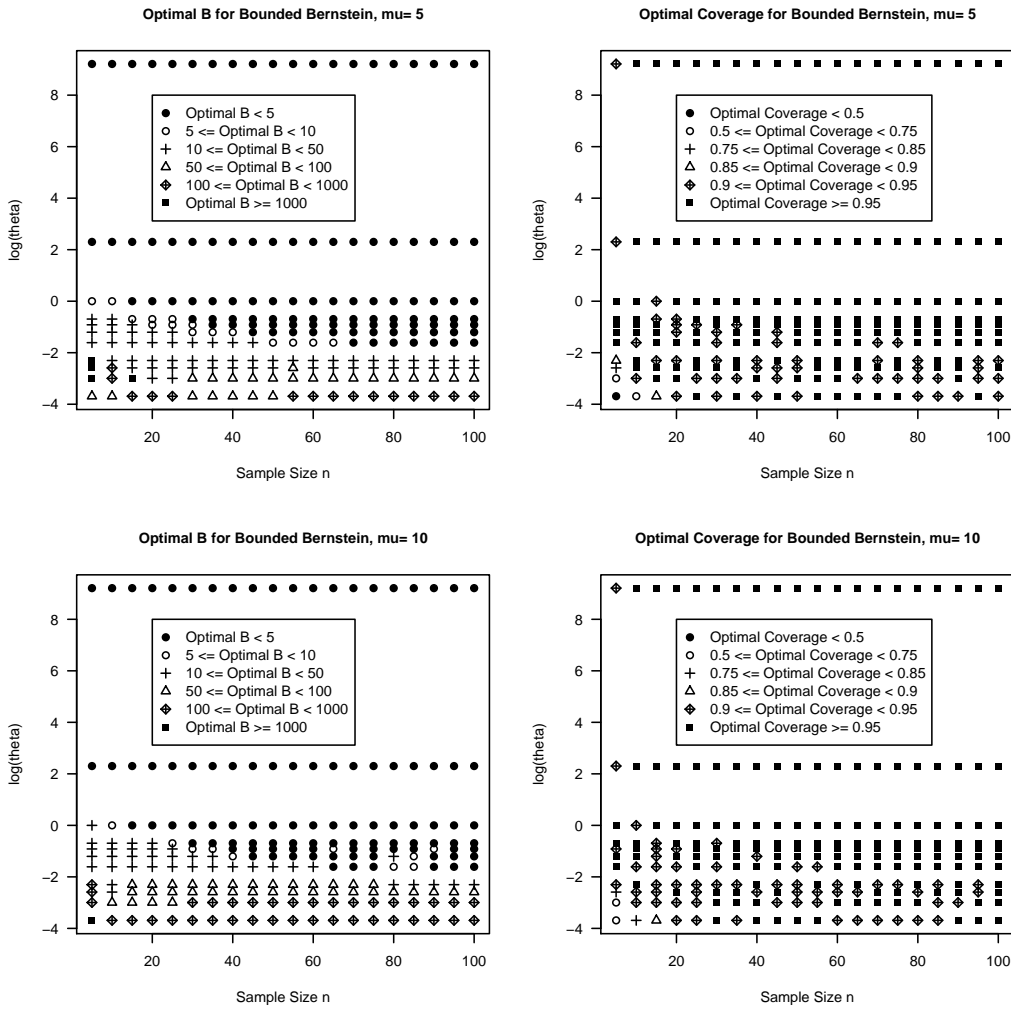
Figure 11: Refinements to the Bounded Bernstein method that can arise as a result of the proper selection of the data's upper bound b. For each value of μ in the first simulation experiment, the optimal value of b and the corresponding coverage are depicted above.

combination of $\mu = 10$, $\theta = 10^{-4}$, and $n = 500$ resulted in a sizable proportion of data sets consisting only of zeros.

Figure 12 displays a comparison of the Wald and Gamma methods across the simulation parameters. In general, it appears that the Gamma Assumption is validated in these large sample sizes and high dispersions, and the coverage of the Gamma method improves as n grows larger and θ smaller. Moreover,

the Gamma method appears to improve upon the Wald method in nearly all of the examples considered. There are many cases in which the Gamma method both covers better and produces a shorter average interval length than the Wald method. Therefore, it appears to be the case that the Gamma approximation converges more quickly than Normal theory in terms of sample size when the dispersion is high. This may be in part due to the fact that the Gamma method is guaranteed to produce a non–negative interval and also lacks the Normal method's symmetry requirement to approximate a skewed distribution. Therefore, the Gamma method appears to be a more suitable method than the Wald for large–sample inference when the dispersion is high.

5 Data Analysis

The Negative Binomial model is particularly applicable as a generalization of the Poisson random variable that allows for the variance parameter to differ from the mean. In this section we will consider examples from the serial

Methods	Wald, Gamma
μ	10
$\log_{10}(\theta)$	$\{-4, -3.75, -3.5, -3.25, -3, -2.75, -2, -1\}$
n	$\{500, 1000, 1500, \ldots, 4500, 5000, 10000\}$
Trials	10000
$\min \hat{\theta}$	10^{-6}

Table 2: Parameter values for the simulation experiment of Section 4.5.

analysis of gene expression (SAGE) and network traffic flow data and explore the utility of the proposed methods as alternatives to the Wald and bootstrap confidence intervals. In doing so, we seek to better elucidate the strengths and weaknesses of the candidate procedures.

5.1 SAGE Data

A serial analysis of gene expression (SAGE) is used in molecular biology to estimate the relative abundance of messenger ribonucleic acid (mRNA) molecules based upon the frequency of corresponding 14 base pair *tag* sequences that are extracted from a cell (Velculescu et al., 1995). Because the cost of sequencing can be prohibitive, the sample size is often limited to a small quantity. Robinson and Smyth (2008) propose a Negative Binomial model for the tag
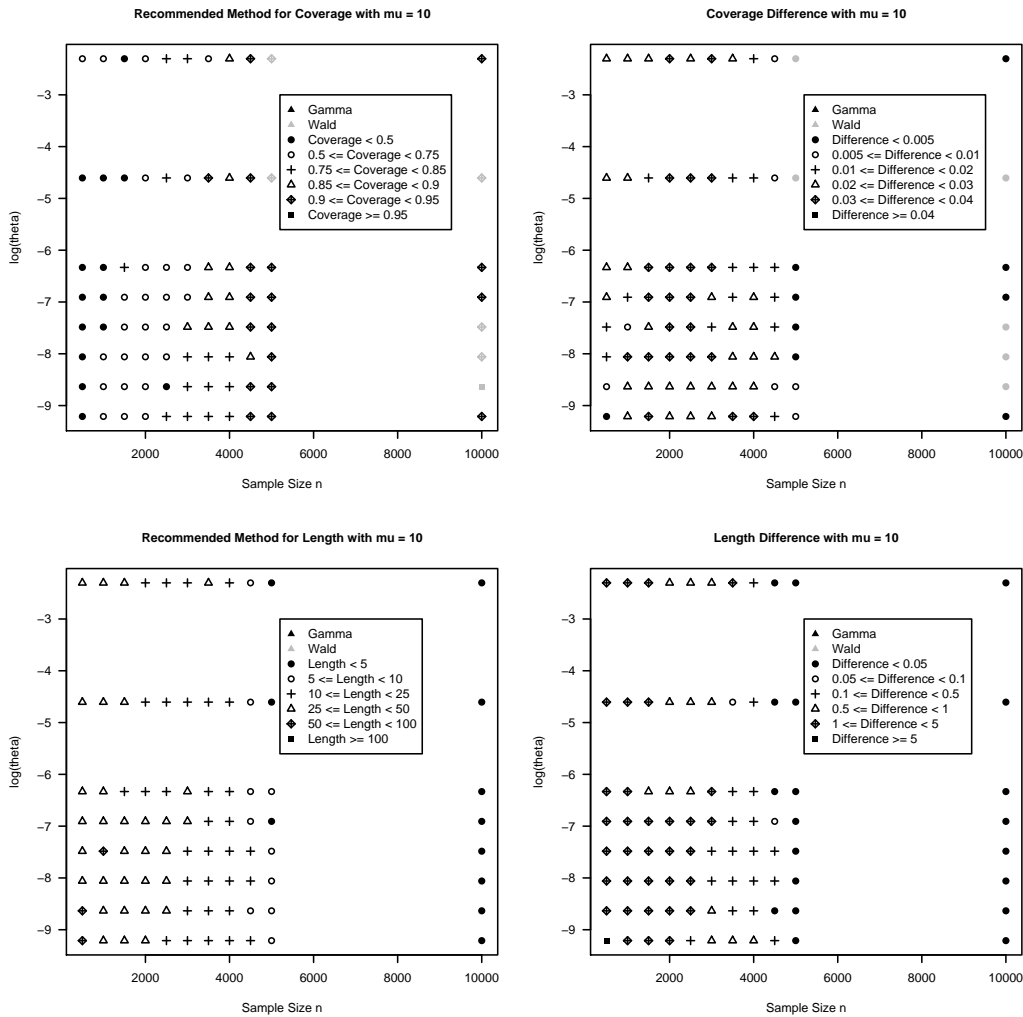
Figure 12: Simulation results comparing the Gamma and Wald method for coverage and length at large sample sizes and high dispersions.

counts of SAGE data and consider the problem of small sample estimation of the dispersion parameter. In this model, the tag counts are assumed to be independent Negative Binomial random variables with common dispersion for the purposes of estimation in spite of the possibility of related biological functions and expression co-regulation (Robinson and Smyth, 2008). We consider this Negative Binomial model in the context of Sample GSM15034 of SAGE data stored at the National Center for Biotechnology Information website for the United States' National Institutes of Health:

(`http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM15034`).

The data, which are shown in Table 3, depict the $n = 20$ most frequent tags and their corresponding counts in a sample taken from the cells of *mus musculus*. The sample mean and standard deviation are $\bar{X} = 306.1$ and $s = 786.15$, respectively. We estimated the value of θ to be 0.6269 with a standard error of 0.1676. These estimates were obtained by applying an iterative maximum likelihood estimation (MLE) procedure within the **glm.nb** method of the **R** statistical programming language to the data. It should also be noted that MLE procedures typically underestimate variance parameters (and therefore the dispersion) (Robinson and Smyth, 2008), so θ may in fact be smaller than 0.6269.

However, even if the tag counts in the SAGE data may be assumed to be independent with a common dispersion, it is not at all clear that they are identically distributed. Robinson and Smyth (2008) consider a model in which each tag count has its own mean parameter, and so the data are only i.i.d. if each tag has the same mean. The Chi Square, Gamma, Wald, and bootstrap confidence intervals may not necessarily be applicable when the data are not i.i.d., but both Bernstein methods only require independent data. Furthermore, the Bounded Bernstein method only applies if the data are uniformly bounded. Such an assumption seems reasonable in this context because the tag counts cannot exceed the length of the mRNA sequence. Therefore, the Negative Binomial model may be seen as merely an approximation to the true distribution of the tag counts, and the Bounded Bernstein method seems reasonable in this context. Based upon the underlying assumptions of each technique, Table 4 displays 95% confidence intervals for the mean μ tag count computed according to the Unbounded and Bounded Bernstein, Chi Square, Wald, bootstrap, and Gamma methods based upon the SAGE tag count data of Table 3.

The 95% confidence intervals of Table 4 simultaneously illustrate many of the strengths and weaknesses of each method. Both the Bounded Bernstein and Wald intervals include a range of negative numbers as possible values for the mean μ, which is unreasonable for Negative Binomial random variables because they draw from a non-negative sample space. By contrast, the Chi Square, Gamma, and bootstrap results are assured to be non-negative. One advantage of using the Gamma approximation over the Wald method when its Gamma assumption is reasonable is that it produces a confidence interval of similar width that is also guaranteed to be positive. The simulation results of Section 4 suggest that the Wald, Gamma, and bootstrap confidence intervals will under-cover μ at small values of n and θ, and so it is not surprising that the Bounded Bernstein interval is much wider. However, the Chi Square confidence

interval is considerably more narrow than those of the Wald and bootstrap. This is not surprising because the ratio quantity (7) has an estimated value of 12.21. Because this ratio is much larger than 1, we expect the Chi Square method to significantly undercover μ. The findings of Section 4.3 suggest that a ratio of 12.21 will result in a coverage well below 50%. Although the Chi Square method performs poorly in this circumstance, this interpretation of the ratio quantity provides a strong indication that the wider intervals produced by other methods are more reasonable for this context.

In selecting among the candidate confidence intervals of Table 4, we recommend choosing the Bounded Bernstein result for two reasons: first, the simulation studies suggest that this method improves upon the coverage of the Wald, Gamma, and bootstrap procedures at small values of n and θ. Second, the assumption of independent data is more reasonable than that of i.i.d. data, and only the Unbounded and Bounded Bernstein methods remain robust

Tag	Count
TGAGCAAAGCCACC	3581
CATGTCGACCAGCC	657
TGAGCGCATGGGTC	428
CAGGCAGTGACAGC	170
CAGGTCGCGAAGGG	143
CTGGAGGACCCATG	138
CCACCAGGCAGCTC	122
CGAGCAAATGCCAG	116
CCATGCCAGGCAAT	98
CCCAGCCATCCCAT	78
CCAAAGGAGAGGGC	74
CACCTGGCGTCATG	74
TTAAACGGCGGCTG	66
TGGCCTGAAGAGCA	65
CTAACGGCCGAGAT	62
TGACCTTGCATGTA	54
CTACCGATGGCTGT	53
CAGGACACCACATC	50
CTGGGAGGTCAGGC	48
CTGCCCAATTTGCC	45

Table 3: SAGE Sample GSM15034 taken from *mus musculus* displaying the 20 most frequent tags and their corresponding counts in the SAGE sample.

in this setting. Although its negative left endpoint limits the interpretation of the confidence interval, the Bounded Bernstein method at least suggests that a wider range of values for μ should be considered than those reported by the Wald, Gamma, and bootstrap results. In selecting among the remaining confidence intervals, the bootstrap and Gamma results are expected to exhibit similar coverage to that of the Wald while maintaining a positive left endpoint. By contrast, the Chi Square confidence interval does not appear to be a good fit for this combination of n, μ, and θ.

	Lower Limit	Upper Limit	Other Quantities
Unbounded Bernstein	182.14	430.06	ϵ quality = 6.90e-07
Bounded Bernstein	-455.02	1067.22	–
Chi Square	259.53	356.46	$\hat{\mu}/(2n\hat{\theta}) = 12.21$
Gamma	160.82	497.35	–
Wald	-38.00	650.20	–
Bootstrap	111.95	1015.75	–

Table 4: 95% confidence intervals for the mean tag count based upon the SAGE data of Table 3. The Chi Square, Wald, bootstrap, and Gamma confidence intervals were computed under the assumption of i.i.d. data, whereas the Unbounded and Bounded Bernstein intervals only assume that the data are independent. The estimated ratio quantity of 12.21 indicates that the Chi Square method is likely to significantly undercover μ. The ϵ quality metric shows that the selected value of ϵ in the Unbounded Bernstein method solves Equation (12) to within 6.90e-07.

5.2   Traffic Flow Data

We now consider an example arising from the analysis of traffic flow data in an internet communications network. Sanchez and He (2005) seek to estimate the mean packets per second (PPS) flowing through the network and propose a Negative Binomial model for the packet counts. The data, which are available at the Lawrence Berkeley National Laboratory's website:
    (`http://ita.ee.lbl.gov/html/contrib/DEC-PKT.html`),
    consist of packet counts at each of $n = 102$ consecutive seconds. It is presumed that packets arrive according to a Poisson process with dispersion, so a Negative Binomial model is suitable for this analysis. This again raises the question of whether the data are actually bounded. Communications networks typically operate under a capacity constraint that suggests bounded data; therefore, it seems reasonable to assume that the Bounded Bernstein method

is also appropriate in this scenario. The sample mean and standard deviation of the data are $\bar{X} = 310.31$ and $s = 94.54$, respectively. We again used the **glm.nb** method in **R** to estimate the dispersion as $\hat{\theta} = 10.59$ with a standard error of 1.52. Although this example has a similar sample mean to that of the SAGE data above, the values of n and θ are considerably larger in this case.

In the simulation studies of Section 4, the most similar case to the current example is that of $n = 100$ and $\theta = 10$. At larger values of n and θ, the simulation results generally suggest that the Wald and bootstrap methods perform well in terms of coverage, whereas the Bounded Bernstein and Chi Square techniques generally over–cover the mean μ. Meanwhile, it is unclear whether the Gamma assumption is reasonable at this combination of n and θ. Table 5 displays 95% confidence intervals for the mean PPS. The Wald and bootstrap methods result in very similar confidence intervals. Here the Bounded Bernstein and Chi Square intervals are considerably wider than those of the Wald, Gamma, and bootstrap, which all offer similar results. The Unbounded Bernstein method actually produces the most narrow interval. While its selected value of ϵ appears to almost exactly solve Equation (12), the simulation results suggest that this Unbounded Bernstein method tends to undercover while the Wald and bootstrap results are reasonably accurate. With a ratio quantity of 0.14, the Chi Square method is expected to severely overcover because its approximation is not sufficiently close to the Gamma model, which suggests that shorter intervals are more appropriate.

The simulation results of Section 4 suggest that the Wald and bootstrap methods cover μ with a probability very close to the desired 95% for larger values of n and θ while the Bounded Bernstein and Chi Square methods tend to over–cover the mean. Because the Wald and bootstrap results are similar, we recommend selecting either as the preferred confidence interval in this setting. Therefore, it seems reasonable to believe that the mean traffic flow of the network is somewhere between approximately 292 and 329 packets per second. The question remains as to why the Gamma method also produced a similar interval but exhibited poor coverage in the simulations. It is possible that the Gamma approximation is sensitive to the estimated value of θ and performs poorly in some circumstances on account of this estimator's variability.

6   Discussion

The two data analysis examples presented in the previous section are an important reminder that the question of which confidence interval to select should be addressed in the context of the problem at hand. In constructing confidence intervals for the mean of Negative Binomial random variables, a careful

investigation of the dispersion and sample size must be considered. The simulation studies of Section 4 identify a variety of scenarios in which the proposed methods improve upon the standard techniques. Interestingly, the methods largely prove to be complementary. Table 6 provide general guidelines for selecting a method based upon the dispersion and sample size. When both n and θ are small, the Bounded Bernstein method generally improves upon the standard techniques, and Section 4.4 shows that a refined selection of the

	Lower Limit	Upper Limit	Other Quantities
Unbounded Bernstein	306.97	313.66	ϵ quality = 2.29e-08
Bounded Bernstein	276.02	344.60	–
Chi Square	263.41	361.01	$\hat{\mu}/(2n\hat{\theta}) = 0.14$
Gamma	292.08	329.09	–
Wald	291.97	328.66	–
Bootstrap	292.91	329.49	–

Table 5: 95% confidence intervals for the mean packets per second (PPS) flowing through a communications network. The data were collected from $n = 102$ seconds of traffic flow. The ϵ quality metric shows that the selected value of ϵ in the Unbounded Bernstein method solves Equation (12) to within 2.29e-08. Similarly, the estimated ratio value of 0.14 suggests that the Chi Square method is likely to significantly overcover the mean μ.

upper bound b has the potential to greatly improve the coverage even at the smallest values of n and θ considered. When θ is small but the sample size n is large, the Gamma assumption becomes reasonable, and the simulation of Section 4.5 shows that the Gamma method can often provide both more reliable coverage and a shorter average length than the Wald method. When both n and θ are large, the Wald and bootstrap methods generally perform well. Meanwhile, the Chi Square method best applies at the combination of μ, θ, and n values that produce a ratio quantity (7) close to 1. In practice, this subspace of values will be in between the extremes at which the other methods are recommended. Finally, the simulation results suggest that the case of small n and large θ represent a decision point at which it's unclear whether to prefer the Bounded Bernstein method or rely upon the standard techniques. Likewise, the boundaries at which each technique overtakes another are not clearly demarcated, and some investigation of a study's context should be considered in selecting between methods.

A number of other considerations apply in selecting among candidate procedures for constructing $1 - \alpha$ confidence intervals. These concerns are summarized in Table 6. In terms of computational speed, the bootstrap method requires B bootstrap resamplings of the data and corresponding sample mean calculations plus a final sort of the results, which leads to a computational complexity of $O\left(Bn + B\log(B)\right)$. In practice, the value of B should be a reasonably large number such as 10,000, which renders the bootstrap method significantly more costly than the alternatives. However, in many cases a single bootstrap confidence interval may be computed in no more than a minute. Furthermore, compared to the time required to design and gather data in a scientific study, even a computation requiring several hours or days to compute a bootstrap confidence interval is reasonable. We are also concerned with

Scenario	Preferred Method
Small n, small θ	Bounded Bernstein
Large n, small θ	Gamma
Small n, large θ	Wald, Bootstrap, or Bounded Bernstein
Large n, large θ	Wald or Bootstrap
Ratio quantity close to 1	Chi Square

Table 6: General guidelines for selecting among the proposed methods according to the scenario.

	U.B.	B.B.	χ^2	Wald	Boot	Γ
Computationally Fast	Yes	Yes	Yes	Yes	No	Yes
Positive CIs Assured	No	No	Yes	No	Yes	Yes
Assumptions on n, μ, or θ	No	No	Yes	Yes	No	Yes
Useful at Small n and θ	No	Yes	Yes	No	No	No
Over–covers at Large n	No	Yes	Yes	No	No	No
Over–covers for High θ	No	Yes	Yes	No	No	No
Requires Independence	Yes	Yes	Yes	Yes	Yes	Yes
Requires i.i.d. Data	No	No	Yes	Yes	Yes	Yes

Table 7: A comparison of the Unbounded Bernstein (U.B.), Bounded Bernstein (B.B.), Chi Square (χ^2), Wald (Normal Approximation), bootstrap (Boot), and Gamma (Γ) methods for computing $1 - \alpha$ confidence intervals of the mean μ based upon n i.i.d. observations of a Negative Binomial random variable in terms of a variety of concerns about the applicability, feasibility, and interpretability of these methods.

the interpretability of the confidence intervals produced by each method. In the case of a small value of μ, the resulting Wald, Unbounded Bernstein, and Bounded Bernstein confidence intervals may result in a left endpoint that is less than zero; such a result is of course an implausible value of μ for non-negative data. By contrast, the Chi Square, Gamma, and bootstrap confidence intervals always result in non–negative left endpoints. In terms of applicability, the bootstrap and Unbounded and Bounded Bernstein methods only require very mild assumptions (e.g. finite parameter values and independent or i.i.d. data), although the Bounded Bernstein's supposition of uniformly bounded data is violated when the data truly follow a Negative Binomial distribution. By contrast, the Chi Square, Gamma, and Wald confidence intervals require stronger assumptions about n, μ, and θ. The Chi Square approximation is only applicable when the ratio quantity of Equation (7) is reasonably close to 1. Meanwhile, the Wald and Gamma methods are only reasonable when their underlying CLT and Gamma assumptions are respectively true. Finally, it should be emphasized that the Chi Square, Wald, Gamma, and bootstrap methods assume i.i.d. data, whereas both Bernstein confidence intervals only require that the data be independent.

The simulation results clearly demonstrate that the Bounded Bernstein, Chi Square, and Gamma methods are useful alternatives to the Wald and bootstrap under high dispersion. However, it is also important to consider whether these methods' respective coverage probabilities asymptotically converge to the desired fiduciary limit. The Wald method is well justified at large sample sizes by the Central Limit Theorem, and bootstrap confidence intervals can be shown to converge in coverage to $1 - \alpha$ as the sample size n and number of resamplings B grow large. Provided that θ is sufficiently small, the Gamma method appears to converge faster than the Wald as a function of sample size. Such a convergence cannot be expected of the Bounded Bernstein and Chi Square methods, though. The simulation results suggest that these techniques will largely over–cover μ for large sample sizes. Although the coverage probability of each technique is the most informative measure of the method's reliability, these other aspects should be considered in selecting among candidate procedures for constructing $1 - \alpha$ confidence intervals.

Future investigation in this area may explore a variety of questions raised by this study. The two Bernstein confidence intervals may be refined through improvements in probability tail bounds, improved procedures for calculating ϵ to solve Equation (12), and improved estimates of the upper limit b, variance σ^2, and the dispersion parameter θ, particularly in the case of small sample sizes. The limits of the Chi Square distribution's applicability as a probability model for $\bar{X}$ may be better substantiated through both analytical

and empirical techniques. The length of Chi Square confidence intervals may also be reduced; for instance, Tate and Klett (1959) demonstrate a variety of approaches that reduce the length of a Chi Square interval for the variance of a Normal distribution over that obtained from a procedure allocating equal probability mass to each tail. A more thorough investigation of the Gamma model for $\bar{X}$ would provide greater insight into the relationship between n and θ required to justify the Gamma assumption. Finally, the proposed techniques may be generalized to construct confidence intervals for other parameters of a sample of n i.i.d. Negative Binomial random variables using techniques based upon the data's empirical influence curve.

References

Aragón, J., D. Eberly, and S. Eberly (1992). Existence and uniqueness of the maximum likelihood estimator for the two–parameter negative binomial distribution. *Statistics and Probability Letters 15*(5), 375–379.

Bennett, G. (1962). Probability inequalities for the sum of independent random variables. *Journal of the American Statistical Association 57*, 33–45.

Bennett, G. (1963). On the probability of large deviations from the expectation for sums of bounded, independent random variables. *Biometrika 50*, 528–535.

Bernstein, S. N. (1934). *Teoriya Veroiatnostei (In Russian)*. Publisher Unknown.

Berry, A. C. (1941). The accuracy of the gaussian approximation to the sum of independent variates. *Transactions of the American Mathematics Society 49*, 122–136.

Birge, L. and P. Massart (1998). Minimum contrast estimators on sieves: Exponential bounds and rates of convergence. *Bernoulli 4*(3), 329–375.

Blyth, C. R. and H. A. Still (1963). Binomial confidence intervals. *Journal of the American Statistical Association 78*, 108–116.

Casella, G. and R. Berger (1990). *Statistical Inference*. Wadsworth, Pacific Grove, CA.

Clark, S. J. and J. N. Perry (1989). Estimation of the negative binomial parameter κ by maximum quasi-likelihood. *Biometrics 45*(1), 309–316.

Clopper, C. J. and E. S. Pearson (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika 26*, 404–413.

Crow, E. L. and R. S. Gardner (1959). Confidence intervals for the expectation of a poisson variable. *Biometrika 46*, 441–453.

Efron, B. and R. Tibshirani (1994). *An Introduction to the Bootstrap.* Chapman and Hall, Boca Raton, FL.

Esseen, C. G. (1942). On the liapounoff limit of error in the theory of probability. *Ark. Mat. Astr. Fys. 28A*, 1–19.

Esseen, C. G. (1956). A moment inequality with an application to the central limit theorem. *Skand. Aktuanetidsrk. 39*, 160–170.

Ferreri, C. (1997). On the ml–estimator of the postive and negative two–parameter binomial distribution. *Statistics and Probability Letters 33*, 129–134.

Hilbe, J. M. (2007). *Negative Binomial Regression.* Cambridge University Press.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association 58*, 13–30.

Lloyd-Smith, J. O., S. J. Schreiber, P. E. Kopp, and W. M. Getz (2005). Superspreading and the effect of individual variation on disease emergence. *Nature 438*, 355–359.

Piegorsch, W. W. (1990). Maximum likelihood estimation for the negative binomial dispersion parameter. *Biometrics 46*(3), 863–867.

Pieters, E. P., C. E. Gates, J. H. Matis, and W. L. Sterling (1977). Small sample comparison of different estimators of negative binomial parameters. *Biometrics 33*(4), 718–723.

Robinson, M. and G. K. Smyth (2008). Small sample estimation of negative binomial dispersion, with applications to sage data. *Biostatistics 9*, 321–332.

Rosenblum, M. and M. J. van der Laan (2008). Confidence intervals for the population mean tailored to small sample sizes, with applications to survey sampling. Technical Report 237, Division of Biostatistics, University of California, Berkeley.

Sanchez, J. and Y. He (2005). Internet data analysis for the undergraduate statistics curriculum. *Journal of Statistics Education 13*(3).

Sterne, T. H. (1954). Some remarks on confidence or fiducial limits. *Biometrika 41*, 275–278.

Tate, F. F. and G. W. Klett (1959). Optimal confidence intervals for the variance of a normal distribution. *Journal of the American Statistical Association 54*, 674–682.

Uspensky, J. V. (1937). *Introduction to Mathematical Probability*. McGraw–Hill, New York, NY.

van Beek, P. (1972). An application of fourier methods to the problem of sharpening the berry-esseen inequality. *Z. Warsch. verw. Gebiete 23*, 187–196.

van der Laan, M. J. and D. B. Rubin (2005). Estimating function based cross-validation and learning. Technical Report 180, Division of Biostatistics, University of California, Berkeley.

Velculescu, V. E., L. Zhang, B. Vogelstein, and K. W. Kinzler (1995). Serial analysis of gene expression. *Science 270*, 484–487.

Wilcox, R. R. (2005). *Robust Estimation and Hypothesis Testing*. Elsevier Academic Press, Burlington, MA.