Sample Size Determination in Auditing Accounts Receivable Using A Zero-Inflated Poisson Model

By
Kristen Pedersen
A Project
Submitted to the Faculty
Of
Worcester Polytechnic Institute
In partial fulfillment of the requirement for the
Degree of Master of Science
In
Applied Statistics

May 2010

APPROVED:

Dr. BalgobinNandram, Project Advisor

Dr. Huong Higgins, Co-Advisor

Sample Size Determination in Auditing Accounts Receivable Using A Zero-Inflated Poisson Model

ABSTRACT

In the practice of auditing, a sample of accounts is chosen to verify if the accounts are materially misstated, as opposed to auditing all accounts; it would be too expensive to audit all acounts. This paper seeks to find a method for choosing a sample size of accounts that will give a more accurate estimate than the current methods for sample size determination that are currently being used. A review of methods to determine sample size will be investigated under both the frequentist and Bayesian settings, and then our method using the Zero-Inflated Poisson (ZIP) model will be introduced which explicitly considers zero versus non-zero errors. This model is favorable due to the excess zeros that are present in auditing data which the standard Poisson model does not account for, and this could easily be extended to data similar to accounting populations.

Introduction

When auditing a company's accounts receivable, a sample of accounts is chosen to verify if the accounts are materially misstated. An auditor is responsible for selecting a sample of the reported book values, and making inferences about the accuracy of the company's records based on the sample. "Accounting populations are frequently very highly skewed and the error rates which the auditor is seeking to detect are often extremely low" (Knight, 1979). This paper seeks to find a method for choosing a sample size of accounts that will provide a more accurate estimate than the current methods for sample size determination that are currently being used. In this paper, all accounts receivable book values constitute the population. Book value is the value recorded for accounts or financial statements. A sample is a selection of some, but not all, of the accounts. The information gathered from the sampled accounts is used to make inferences about the population. The only way to get the total value of the accounts is to audit all accounts, but this would be very costly. Not having to verify the information on all accounts to make these inferences reduces the cost in calculating the quantity of interest. Experience has shown that a properly selected sample frequently provides results that are as good as the results from verifying all accounts, (Higgins and Nandram, 2009). Dollar unit acceptance sampling to reduce needed sample size in auditing data is thoroughly discussed in Rohrback (1986). Ponemon and Wendell studied the benefits of using statistical sampling methods in auditing data, as opposed to an auditor using his expertise to choose the sample (Ponemon and Wendell, 1995). Although cost is not specifically addressed in this paper, it provides motivation for investigating what sample size is needed to get efficient results.

The Zero-Inflated Poisson model will be introduced in section 4, it is a model to accommodate count data with excess zeros. If a company keeps accurate accounts receivable, then there would be no errors, this means that there will be more zeros in the data than would be accounted for under the standard Poisson model. Therefore, we hope get a better estimate of the appropriate sample size. In "Monetary Unit Sampling: Improving estimation of the total audit error" (Higgins and Nandram, 2009), the ZIP method is dicussed and it is shown that for accounting data and other similar data

populations, that a bound under the ZIP model is reliable and more efficient than common Monetary Unit Sampling practice.

Related Research

Kaplan (1973) conducted simulation studies based on hypothetical populations and error patterns to observe the behavior of ratio and difference estimators when the population contains a limited number of zeros. Kaplan found a strong correlation between the point estimate and the estimated standard error, and showed that the nominal confidence level implied by the normal distribution for large-sample confidence intervals was frequently far different from the proportion of correct confidence intervals.

Baker and Copeland (1979) evaluated the use of stratified regression in comparison to standard regression in account auditing data. A minimum of 20 errors for the difference of book value and audit value as an estimator was found to give superior results for the stratified regression. A minimum of 15 errors for a ratio estimator is needed for superior results in stratified regression. The usefulness of this information is questionable due to the low error rate of accounting populations.

Sahu and Smith (2006) explored a full Bayesian framework in the auditing context. They found that non-informative prior distributions lead to very small sample sizes. Specifically, if the mean of the prior distribution is far from the boundary value (or the per item material error), then the sample size required is very small. In this case, the sample size could be set by the auditor. When the prior mean is close to the material error a large sample size is required.

Berg (2006) proposed a Bayesian technique for auditing property value appraisals. A Beta-Binomial model was implemented, and their procedure required smaller sample sizes relative to those based on classical sample size determination formulas.

Data

The data comes from Lohr (1999) where the recorded values $b_1, b_2, b_3, \ldots\ldots b_n$ for a sample of n=20 accounts for a company were listed, along with all the audited (actual) values $a_1, a_2, a_3 \ldots\ldots a_n$ of the sample. The company had a total of N=87 accounts receivable. The total book value for the N=87 accounts receivable of the company was $612,824. The total book value of all accounts receivable for a company would be $B= b_1 + b_2 + \ldots\ldots + b_N$. The total audit value for a company's accounts receivable would be $A=a_1 + a_2 + \ldots\ldots + a_N$. We'll define the error to be the difference of the book value and the true audit value for each account i (i=1, 2, 3,... N) as $y_i = b_i - a_i, y_i \geq 0$. This means that the total amount of error for the accounts is $Y = \sum_1^N y_i = \sum_1^N b_i - a_i$. In the accounting context, we expect a large number of accounts to have $y_i = 0$. Let θ be the error rate per dollar. Therefore the error of each account will be $\theta = y_i / b_i$ for i=1, ... , n., and the error rate per dollar for the sample will be will be $\theta = Y / B$. Our initial estimate for θ obtained from Lohr's data is $\theta_0 = .007$ with standard deviation $\sigma_0 = .022$ these estimates are the mean and variance calculated from the sample in the Lohr text. A random sample of 20 accounts with replacement was taken from the population of 87 accounts. The book value, the audit value, and the difference between the book value and audit value for 16 of the accounts are listed in the table below. The sample of size 20 in the Lohr text was with replacement, there were 4 accounts that were repeated in the sample in the Lohr text, and these 4 accounts were removed for the purpose of this paper.

Table 1:

Account	Book Value	Audit Value	BV-AV
3	6842	6842	0
9	16350	16350	0
13	3935	3935	0
24	7090	7050	40
29	5533	5533	0
34	2163	2163	0
36	2399	2149	250
43	8941	8941	0
44	3716	3716	0
45	8663	8663	0
46	69540	69000	540
49	6881	6881	0
55	70100	70100	0
56	6467	6467	0
61	21000	21000	0
70	3847	3847	0
74	2422	2422	0
75	2291	2191	100
79	4667	4667	0
81	31257	31257	0

Initial Exploration

We first looked at the sample size required for n in a hypothesis test while controlling for a significance level of .05 ($\alpha = .05$) and power equal to .95 ($\beta = .95$). We chose both Poisson and Binomial models for our initial exploration. The poisson distribution is applied in counting the number of rare events, which in the context of auditing data we are modeling the occurrence of the book value not being equal to the audit value. A reasonable model for

initial exploration is $y_i \sim Poisson(b_i\theta)$, i=1, … ,N, which implies that the mean of the difference of the book value and audit value is the book value multiplied by the error rate. The Binomial model was also chosen for initial exploration because the binomial distribution is easily approximated by the normal distribution, and here the thought would be that the book value b_i is the number of dollars in a particular account and each dollar has probability θ of being materially misstated. This model would be expressed by $y_i \sim Binomial(b_i, \theta)$ i=1, … ,N. The binomial distribution does not allow for the large number of $y_i = 0$ that we have in our sample, but as initial exploration, the results under the binomial model can be compared with the results from the poisson model using similar methods to make comparisons and help argue that our results are reasonable. (Sahu and Smith, 2006) investigate the use of the normal distribution where the assumptions of normality are not appropriate. A thorough discussion of confidence interval criteria is given in (Jiroutek, Muller, Kupper and Stewart, 2003). Using decision theory to select and appropriate sample size is covered in two papers by Menzefricke (1983 and 1984). In sections 1 and 2 of this paper we use the poisson and binomial models respectively for a frequentist approximation (section 1) and a Bayesian method of approximation (section 2). Section 3 briefly discusses the interval for θ from the posterior distribution of the poisson model. A discussion of Bayesian model performance criteria is given in (Wang and Gelfand, 2002). In our paper, this initial exploration is moving towards the introduction of the Zero-Inflated Poisson model that will be discussed in section 4.

1. Frequentist Approximation

First exploration involves looking at the sample size required for confidence intervals under both the Poisson model and the Binomial model. Here, we chose to control for the length of the interval, we allowed L (length) to be in the interval (.001, .02). This involved using estimates from the data set. An initial estimate for the mean error per dollar θ, called θ_0 was calculated from the sample to be $\theta_0 = .007$, with a standard deviation $\sigma_0 = .022$. The formulas used to calculate the mean and the variance were

$\theta_0 = \frac{1}{16}\sum_{i=1}^{16}\frac{y_i}{b_i}$ and $\sigma_0 = \frac{1}{16-1}\sum_{i=1}^{16}\left(\frac{y_i}{b_i} - \theta_0\right)^2$, Casella and Berger (2002). The results for the sample size n that result from this initial exploration under the respective models in sections 1.1 and 1.2 below.

1.1 Frequentist Approximation for n under the Poisson Model:

$$\text{Here, } y_i \sim Poisson(b_i\theta) \quad i = 1, \ldots, n \ .$$

Equation (1) computes a $(1-\alpha)$% confidence interval for θ. Here,

$$\Pr ob\left\{\theta : \left|\frac{\bar{Y} - \bar{b}\,\theta}{\sqrt{\bar{b}\,\theta/n}}\right| \le \zeta_{\alpha/2}\right\} \ge 1-\alpha\,, \tag{1}$$

where $\zeta_{\alpha/2}$ is used to represent the z-score for the $\alpha/2$ percentile of the normal distribution. However, we are not looking to compute a confidence interval for θ, but rather we are looking to have an interval for the appropriate sample size n. The distance of the two endpoints of this interval, $\theta - \zeta_{\alpha/2} * \sqrt{\bar{b}\,\theta/n}$ to $\theta + \zeta_{\alpha/2} * \sqrt{\bar{b}\,\theta/n}$ is the length. Since our interval is centered around θ, we can use 2 multiplied by the distance of θ to $\theta + \zeta_{\alpha/2} * \sqrt{\bar{b}\,\theta/n}$ to calculate the length.

This implies that the length of the interval resulting under this model can be described by the function

$$L = \frac{2\sqrt{\bar{b}^2\left(2\bar{Y} + \frac{\zeta^2_{\alpha/2}}{n}\right)^2 - 4\bar{b}^2\bar{Y}^2}}{2\bar{b}^2}\,. \tag{2}$$

The algebra to demonstrate the intermediate steps can be found in section 1.1 of the appendix.

We also know that the expectation of $\bar{Y}$ and $\bar{Y}^2$ based on the Poisson distribution will be $E(\bar{Y}) = \bar{b}\,\theta_0$ and $E(\bar{Y}^2) = \frac{\bar{b}\,\theta_0}{n} + (\bar{b}\,\theta_0)^2$. Filling these values of the expectation in for $\bar{Y}$ and $\bar{Y}^2$ we are able to account for not having this information. After substituting these expectations into our equation for $\bar{Y}$, and solving for n, the result is

$$n = \frac{\zeta^2_{\alpha/2}\left(2\theta_0 \pm \sqrt{4\theta_0^2 + L^2}\right)}{\bar{b}L^2}. \qquad (3)$$

In Figure 1 below, the results for the sample size versus length are illustrated.

Figure 1:

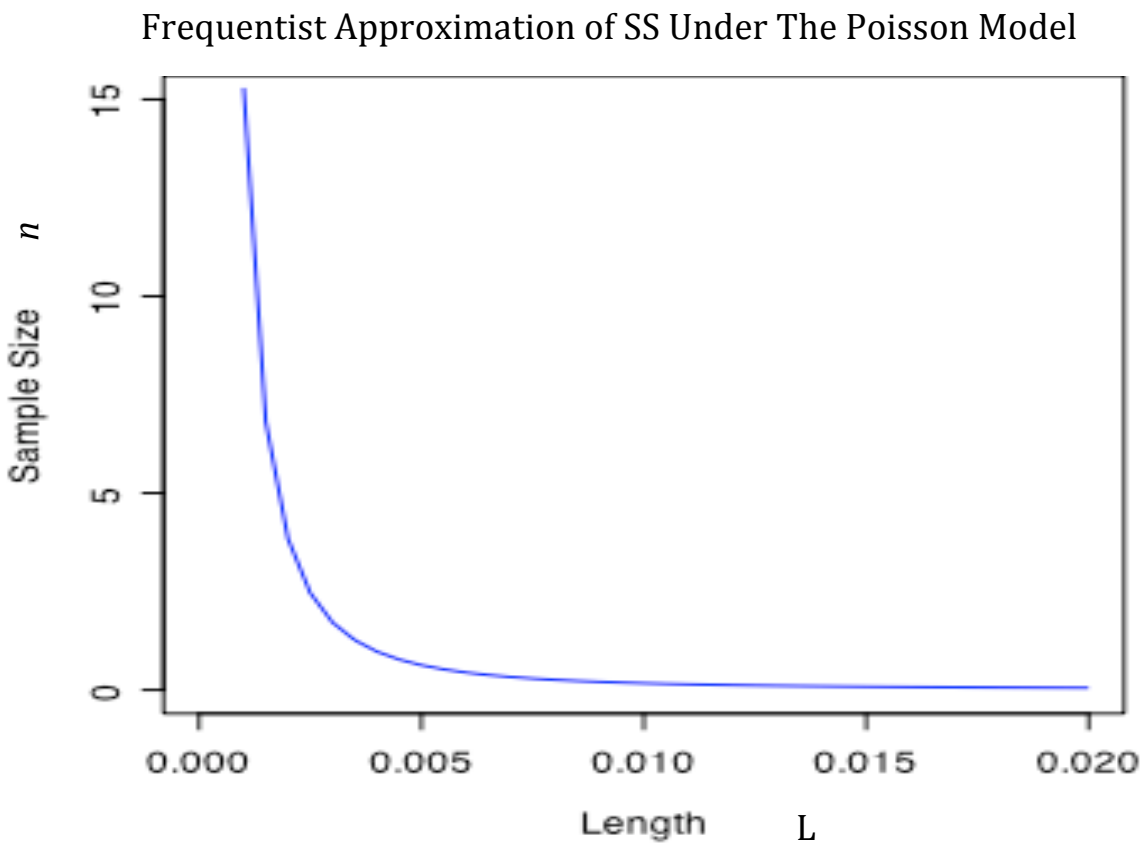


Figure 1 demonstrates that for an interval length less than 0.007 the sample size is increasing as the length becomes smaller. A Length of the interval greater than 0.007 would require a sample size of 1.

1.2 Frequentist Approximation for n under the Binomial Model:

Using the same method as above, the sample size is calculated under the Binomial model. We seek to have an estimate for the sample size n that is similar to the estimate we calculated under the Poisson model. Here,

$$y_i \sim Binomial(b_i, \theta) \; i = 1, \ldots, n.$$

Under the Binomial model, equation (4) shows the interval for θ,

$$\Pr ob\left\{\theta : \left|\frac{\bar{Y} - \bar{b}\,\theta}{\sqrt{\bar{b}\,\theta(1-\theta)/n}}\right| \le \xi_{\alpha/2}\right\} \ge 1 - \alpha \; . \tag{4}$$

This implies that the length of the interval under the Binomial model is expressed as in equation (5) is

$$L = \frac{\sqrt{\bar{b}^2\left(2\bar{Y} + \frac{\xi^2_{\alpha/2}}{n}\right)^2 - 4\left(\bar{b}^2 + \xi^2_{\alpha/2}\frac{\bar{b}}{n}\right)\bar{Y}^2}}{\bar{b}^2 + \xi^2_{\alpha/2}\frac{\bar{b}}{n}}. \tag{5}$$

Intermediate steps are found in section 1.2 of the Appendix. We also know that $E(\bar{Y}) = \bar{b}\,\theta_0$ and $E(\bar{Y}^2) = Var(\bar{Y}) + E(\bar{Y})^2 = \bar{b}\,\theta_0(1-\theta_0)/n + (\bar{b}\,\theta_0)^2$, then we substitute these expectations into our equation for $\bar{Y}$ and $\bar{Y}^2$. As before, intermediate algebraic steps can be found in

section 1.2 of the Appendix. After solving for n, we get the following equation for n, where a is represented by

$$a = \sqrt{L^2 + \frac{4n\theta_0\left(1 + \theta_0\left(n\bar{b} - 1\right)\right)}{\bar{b}n + \zeta^2_{\alpha/2}}} \quad , \tag{6}$$

and,

$$n = \left(\frac{a-1}{2\bar{b}\,\theta_0 - a\bar{b}}\right)\zeta^2_{\alpha/2} \quad . \tag{7}$$

Our work was verified by using a similar, simpler method for the binomial model above, because the algebra to arrive at this result was extensive. Section 1.3 is the general outline of another approach under the binomial model and this method also resulted in a similar estimated sample size.

1.3 Frequentist Approximation to the simpler Binomial model:

If $x_i \sim Binomial(b_i, \theta)$, i =1, ... ,n, this implies that the sum of the x_i will be distributed as $X \sim Binomial(n\bar{b}, \theta)$.

We have, $x_i \sim Binomial(b_i, \theta) \Rightarrow X = \sum_1^N x_i \sim Binomial\left(\sum_1^N b_i, \theta\right) \Rightarrow X \sim Binomial(n\bar{b}, \theta)$.

A confidence interval is first set up for θ. However, to simplify calculations we choose to let $n\bar{b} = n^*$. The representation of this is given as,

$$\Pr ob\left\{\theta : \left|\frac{X - n\bar{b}\,\theta}{\sqrt{n\bar{b}\,\theta(1-\theta)}}\right| \le \zeta_{\alpha/2}\right\} \ge 1 - \alpha \quad , \tag{8}$$

where $n\bar{b} = n^*$, then

$$\Pr ob\left\{\theta : \left|\frac{X - n^*\theta}{\sqrt{n^*\theta(1-\theta)}}\right| \le \zeta_{\alpha/2}\right\} \ge 1-\alpha \ . \tag{9}$$

Calculations similar to those of the intervals for n result in the formula for length given in equation (10),

$$L = \frac{2\zeta_{\alpha/2}\sqrt{\left(\frac{X(n^* - X)}{n^*}\right) + \frac{1}{4}\zeta^2_{\alpha/2}}}{n^* - \zeta^2_{\alpha/2}} \ . \tag{10}$$

Since we know that $X \sim Binomial(n\bar{b}, \theta) \Rightarrow E(X) = n^*\theta$ we can continue similarly to the previous poisson and binomial intervals, by substituting in the expectation for X. The results are similar to the previous estimation of sample size in the models above. Also included is Figure 3 that displays the ratio of estimated sample size versus the length of the interval.

Figure 2:

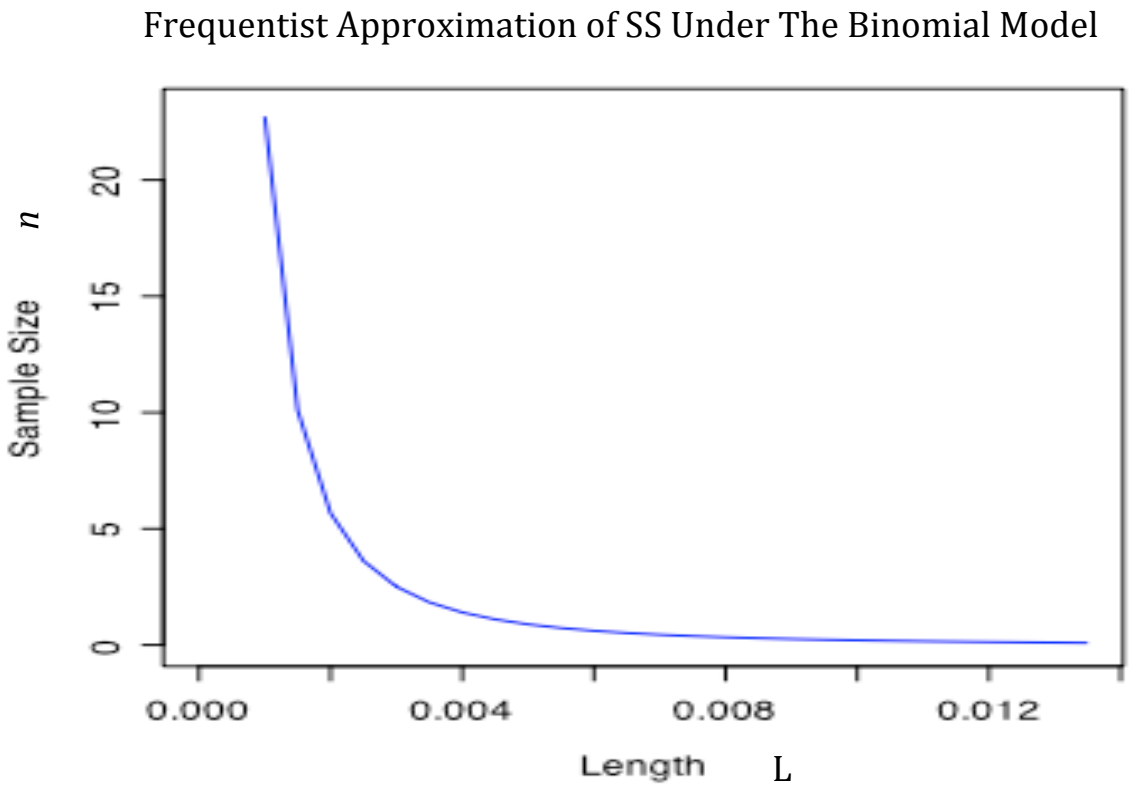
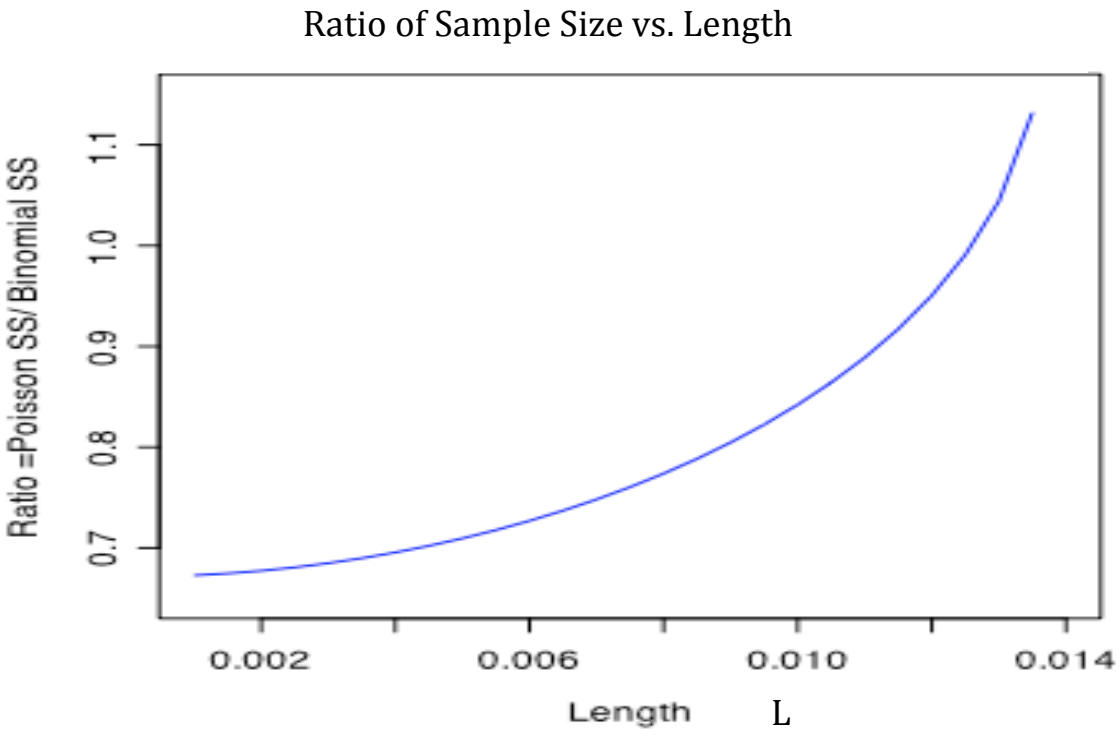


Figure 2 has values of sample size that are slightly greater than the values of sample size in Figure 1 for a given length. However these values are similar, and this is more easily expressed in the ratio plot Figure 3. It would be expected that the values of sample size for corresponding length would be somewhat similar under the poisson and binomial models.

Figure 3:



It can be noticed by comparison of Figure 1 and Figure 2, that the estimated length of the Poisson interval is smaller than the Binomial interval for any sample size, until a length of .013. The ratio of sample size, displayed at Figure 3 of the two models is about 1 after a length of .013, which means that at a length of .013 the required sample size is the same.

2. Bayesian Methods for Sample Size Determination

Next, we will take a look at Bayesian methods of approximation. Let $\mathbf{Y^{(n)}} = (X_1, X_2, \ldots..X_n)$ be a random sample of size *n,* we will look at the highest posterior density region under two models. In part 2.1 assuming a population with density $Y^{(n)}|\theta \sim Poisson(b_i\theta)$, $i = 1, \ldots, n$, and prior distribution $\theta \sim Gamma(\eta, \delta)$ of unknown θ, we will calculate the Posterior

distribution for θ and then continue to find an estimate for the sample size n using a normal approximation to the posterior distribution. McCray (1984) proposed a Bayesian model for evaluating dollar unit samples, even if an informative prior probability distribution on the expected total error is not available. In section 2.2 the model assumed a population density $Y^{(n)}|\theta \sim Binomial(b_i, \theta)$ i=1, ... ,n, with prior $\theta \sim Beta(\phi, \omega)$. Under the binomial model we will proceed by again finding the posterior distribution for θ and then using a normal approximation to find an interval for n.

2.1 Calculating the Posterior Distribution for θ under the Poisson Model:

In Bayesian statistics, the posterior distribution is equal to the prior distribution times the likelihood. We have already stated that the likelihood under the Poisson model is $Y^{(n)}|\theta \sim Poisson(b_i\theta)$, i=1, ... ,n, and we are using prior $\theta \sim Gamma(\eta, \delta)$. This implies that the posterior distribution for θ will therefore be $\theta|Y^{(n)} \sim Gamma(\eta + n\bar{y}, \delta + n\bar{b})$. Given that we have an estimate for θ, $\theta_0 = .007$ from our data, with standard deviation estimate $\sigma_0 = .022$, we can use this information to solve for our initial estimates of η and δ.

Since the prior follows a gamma distribution, we can set up equations for the mean and variance as $E(\theta_0) = \frac{\eta}{\delta} = .007$, and $Var(\theta_0) = \frac{\eta}{\delta^2} = .022^2$. After solving for η and δ, $\eta = .0996$, and $\delta = 14.28$. Intermediate steps are found in section 2.1 of the Appendix. After filling these values for η and δ, the posterior distribution will be distributed as $\theta|Y^{(n)} \sim Gamma(.0996 + n\bar{y}, 14.28 + n\bar{b})$. However, we will still need information for $\bar{y}$, and $\bar{y}$ is unknown.

2.2 Solving for n, using the Normal Approximation for the Posterior of θ under the Poisson Model:

Under the poisson model, the posterior distribution for θ given y is approximately normal with mean $\frac{.0996+\sum_{i=1}^{n} y_i}{14.28+n\bar{b}}$ and variance $\frac{.0996+\sum_{i=1}^{n} y_i}{\left(14.28+n\bar{b}\,\theta\right)^2}$. The mean and variance were easily derived by using the formula for the mean and variance of a Gamma distribution. Thus,

$$\theta|y \sim aN\left(\frac{.0996+\sum_{i=1}^{n} y_i}{14.28+n\bar{b}}, \frac{.0996+\sum_{i=1}^{n} y_i}{\left(14.28+n\bar{b}\,\theta\right)^2}\right), \quad i=1, \ldots, n. \tag{11}$$

We want to choose a sample size for n in $\sum_{i=1}^{n} y_i$ so that we have at least $100(1-\alpha)\%$ confidence for the true difference between the book value and the audited value, given a specified length and total book value. The interval for θ is from $\frac{.0996+\sum_{i=1}^{n} y_i}{14.28+n\bar{b}} - \frac{L}{2}$ to $\frac{.0996+\sum_{i=1}^{n} y_i}{14.28+n\bar{b}} + \frac{L}{2}$, because L is the total length of the interval, and the normal distribution is symmetric. The following equation is for finding the smallest area under our model that is at least $100(1-\alpha)$ % confident for θ. The integration involves averaging over $y^{(n)}$ so that we are no longer dealing with the posterior distribution $\theta|y$, but rather a function of θ. The formula representation for averaging over $y^{(n)}$ is given in equation (12),

$$\int_{y} \int_{\frac{.0996+\sum_{i=1}^{n} y_i}{14.28+n\bar{b}}-\frac{L}{2}}^{\frac{.0996+\sum_{i=1}^{n} y_i}{14.28+n\bar{b}}+\frac{L}{2}} \pi(\theta|y^{(n)})\,d\theta\, p(y^{(n)})\,dy^{(n)} \geq 1-\alpha \ . \tag{12}$$

In section 3, the appropriate sample size n is calculated for the posterior distribution of θ given the information for y. However, in this part of the paper we are averaging over y, so that this interval is calculated for n using similar information to the interval given in section 1. The posterior distribution of $\theta|y$ is approximately normal, so we can make use of the properties of the normal distribution to simplify the calculations. Thus,

$$\int_y \left\{ \Phi\left(\frac{\frac{.0996+\sum_{i=1}^n y_i}{14.28+n\bar{b}} + \frac{L}{2} - \frac{.0996+\sum_{i=1}^n y_i}{14.28+n\bar{b}}}{\sqrt{\frac{.0996+\sum_{i=1}^n y_i}{(14.28+n\bar{b})^2}}} \right) - \Phi\left(\frac{\frac{.0996+\sum_{i=1}^n y_i}{14.28+n\bar{b}} - \frac{L}{2} - \frac{.0996+\sum_{i=1}^n y_i}{14.28+n\bar{b}}}{\sqrt{\frac{.0996+\sum_{i=1}^n y_i}{(14.28+n\bar{b})^2}}} \right) \right\} p(y^{(n)})dy^{(n)} \geq 1-\alpha. \quad (13)$$

Here the mean cancels in the numerator, which leads to the much more simplified looking equation (14),

$$\int_y \left(2\Phi\left(\frac{L}{2}\left(\frac{14.28+n\bar{y}}{\sqrt{.0996+\sum_{i=1}^n y_i}} \right) \right) - 1 \right) p(y^{(n)})dy^{(n)} \geq 1-\alpha \ . \quad (14)$$

Monte Carlo integration was used to get the estimate for $\sum_{i=1}^n y_i$, using the fact that $n\bar{y} \sim Poisson(n\bar{b}\,\theta)$, i=1, … ,n, sample sizes between 100 and 500 were used, and with each sample size n, 10,000 simulations were drawn. The optimal value for n was found to be n=16. Similar results were found under the posterior distribution of θ under the Binomial model, and are calculated and compared in section 2.4.

2.3 Calculating the Posterior Distribution for θ under the Binomial Model:

For the $Y^{(n)}|\theta \sim Binomial(b_i,\theta)$, i=1, … ,n, model with prior $\theta \sim Beta(\phi,\omega)$, the posterior distribution is again the prior times the likelihood and results in a $\theta|Y^{(n)} \sim Beta(\phi+n\bar{y},\omega+n\bar{b}+n\bar{y})$ distribution. Again, the posterior distribution depends on parameters that we can estimate. To solve for the parameters of the beta prior we will use the known equations for the mean and variance, along with our initial estimates for θ and σ. The mean of the beta prior is $E(\theta_0)=\frac{\phi}{\phi+\omega}=.007$ and the variance of the beta distribution is given by $Var(\theta_0)=\frac{\phi\omega}{(\phi+\omega)^2(\phi+\omega+1)}=.022^2$. Solving this system of equations

the resulting parameter estimates are $\phi = .09271$ and $\omega = 13.1514$. Detailed calculation of mean and variance are given in section 2.3 of the Appendix. After filling in these estimates, the resulting posterior distribution is $\theta \left| Y^{(n)} \sim Beta\left(0.09271 + n\bar{y}, 13.1514 + n\bar{b} + n\bar{y}\right)\right.$. Due to the unknown information for $\bar{y}$, looking at the normal approximation to the beta posterior is again of interest.

2.4 Solving for n, using the Normal Approximation for the Posterior of θ under the Binomial Model:

Under the binomial model, the posterior distribution for θ given y is approximately normal with mean $\frac{0.09271 + \sum_{i=1}^{n} y_i}{612{,}754.244}$ and variance $\frac{\left(0.09271 + \sum_{i=1}^{n} y_i\right)\left(612{,}754.244 - \sum_{i=1}^{n} y_i\right)}{(612{,}754.244)^2 (612{,}755.244)}$. The mean and variance were easily derived by using the formula for the mean and variance of a beta distribution. Thus,

$$\theta \left| y \sim aN\left(\frac{\phi + \sum_{i=1}^{n} y_i}{\phi + \omega + n\bar{b}}, \frac{\left(\phi + \sum_{i=1}^{n} y_i\right)\left(\omega + n\bar{b} - \sum_{i=1}^{n} y_i\right)}{\left(\phi + \omega + n\bar{b}\right)^2 \left(\phi + \omega + n\bar{b} + 1\right)}\right)\right., \quad i = 1, \ldots, n. \tag{15}$$

Again, proceeding as we did in section 2.1 above, we want to choose a sample size for n in $\sum_{i=1}^{n} y_i$ so that we have at least $100(1-\alpha)\%$ confidence for the true difference between the book value and the audited value, given a specified length and total book value. The interval for θ is from $\frac{0.09271 + \sum_{i=1}^{n} y_i}{612{,}754.244} - \frac{L}{2}$ to $\frac{0.09271 + \sum_{i=1}^{n} y_i}{612{,}754.244} + \frac{L}{2}$. Thus,

$$\theta \left| y \sim aN\left(\frac{0.09271 + \sum_{i=1}^{n} y_i}{612{,}754.244}, \frac{\left(0.09271 + \sum_{i=1}^{n} y_i\right)\left(612{,}754.244 - \sum_{i=1}^{n} y_i\right)}{(612{,}754.244)^2 (612{,}755.244)}\right)\right., \quad i = 1, \ldots, n. \tag{16}$$

Again, we are proceeding with the same methods as used above for the gamma posterior distribution, by using known properties of the normal distribution. Here,

$$\int_{y}\int_{\frac{0.09271+\sum_{i=1}^{n}y_i}{612,754.244}-\frac{L}{2}}^{\frac{0.09271+\sum_{i=1}^{n}y_i}{612,754.244}+\frac{L}{2}}\pi(\theta|y^{(n)})d\theta p(y^{(n)})dy^{(n)} \geq 1-\alpha \ , \tag{17}$$

can be written as,

$$\int_{y}\left\{\Phi\left(\frac{\frac{0.09271+\sum_{i=1}^{n}y_i}{612,754.244}+\frac{L}{2}-\frac{0.09271+\sum_{i=1}^{n}y_i}{612,754.244}}{\sqrt{\frac{\left(0.09271+\sum_{i=1}^{n}y_i\right)\left(612,754.244-\sum_{i=1}^{n}y_i\right)}{(612,754.244)^2(612,755.244)}}}\right)-\Phi\left(\frac{\frac{0.09271+\sum_{i=1}^{n}y_i}{612,754.244}-\frac{L}{2}-\frac{0.09271+\sum_{i=1}^{n}y_i}{612,754.244}}{\sqrt{\frac{\left(0.09271+\sum_{i=1}^{n}y_i\right)\left(612,754.244-\sum_{i=1}^{n}y_i\right)}{(612,754.244)^2(612,755.244)}}}\right)\right\}p(y^{(n)})dy^{(n)} \geq 1-\alpha . \tag{18}$$

Here the mean cancels in the numerator just as previously seen in the gamma posterior distribution. This cancelation leads to the much more simplified looking equation (19),

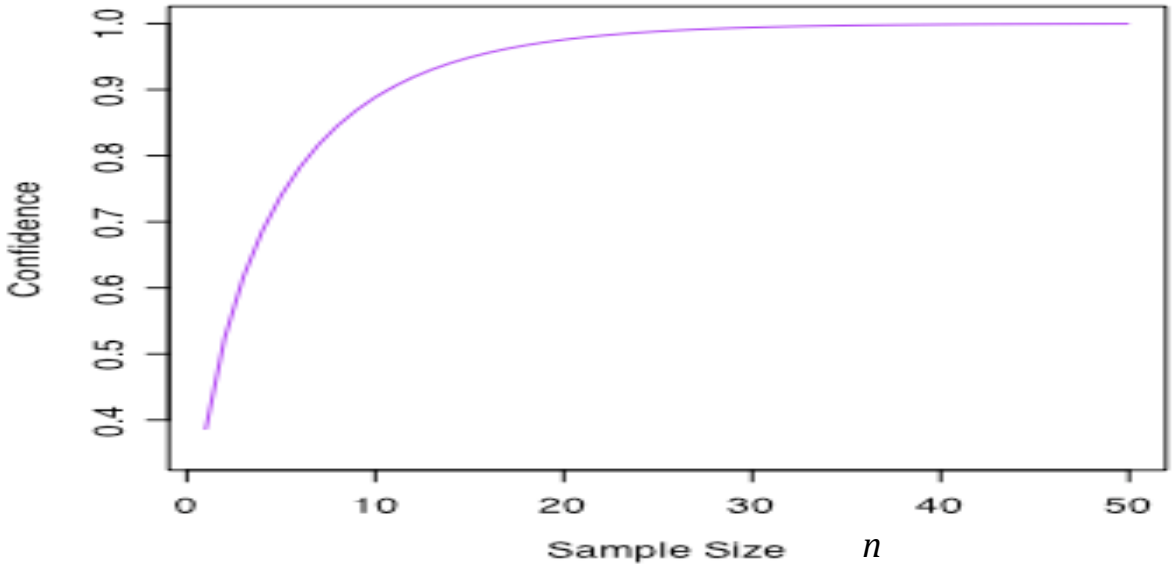
$$\int_{y}\left\{2\Phi\left(\frac{L}{2}\left(\frac{(612,754.244)\sqrt{(612,755.244)}}{\sqrt{\left(0.09271+\sum_{i=1}^{n}y_i\right)\left(612,754.244-\sum_{i=1}^{n}y_i\right)}}\right)\right)-1\right\}p(y^{(n)})dy^{(n)} \geq 1-\alpha . \tag{19}$$

Using Monte Carlo integration to get the estimate for $\sum_{i=1}^{n}y_i$, using the fact that $\sum_{i=1}^{n}y_i \sim Binomial\left(\sum_{i=1}^{n}b_i,\theta\right)$, i=1, … , n, sample sizes between 200 and 600 were used, and with each sample size n, 10,000 simulations were drawn. The optimal value for n was also found to be n=16, this was the same sample size that was found to be optimal under the poisson model.

Figures 4 and 5 below are plots of sample size vs. varying $1-\alpha$ confidence levels. This was found using the formula for the normal approximation $\int_{y}\int_{\mu-\frac{L}{2}}^{\mu+\frac{L}{2}}\pi(\theta|y^{(n)})d\theta p(y^{(n)})dy^{(n)} \geq 1-\alpha$, under the two different models respectively. Both Figures 4 and 5 are for fixed L, L=0.001 and $\bar{b}$, $\bar{b}$ =7043.

Figure 4:

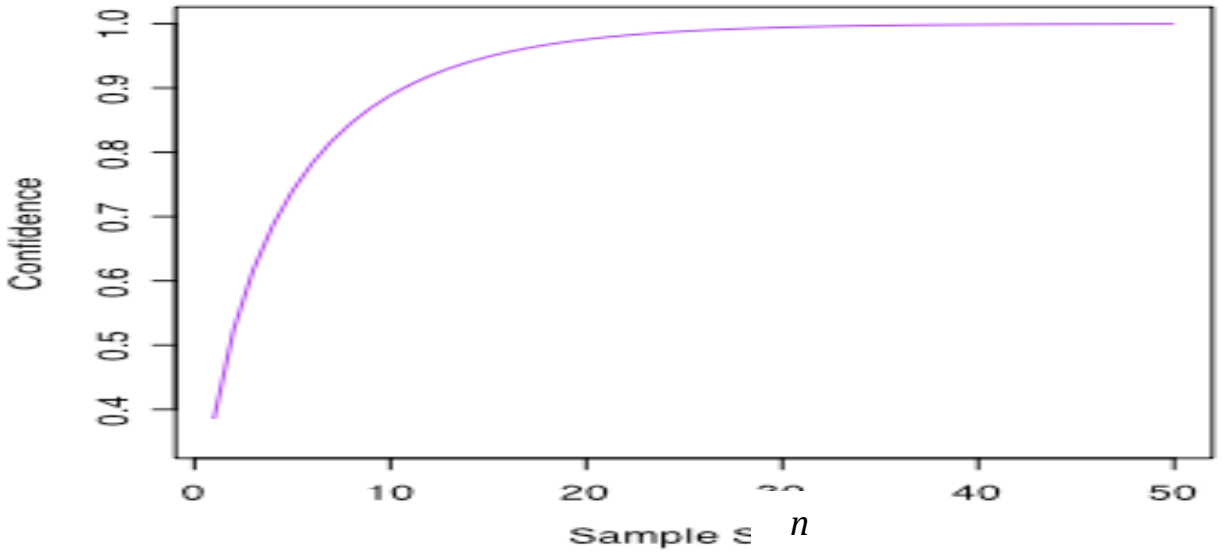
SSD For Normal Approximation Under The Poisson Model Posterior



Figures 4 and 5 demonstrate that the 95% confidence level is achieved for an integer sample size of 16 under both the normal approximation to the poisson model and the binomial model with the length held at a constant .001 and the average book value, $\bar{b}$ = 7043.

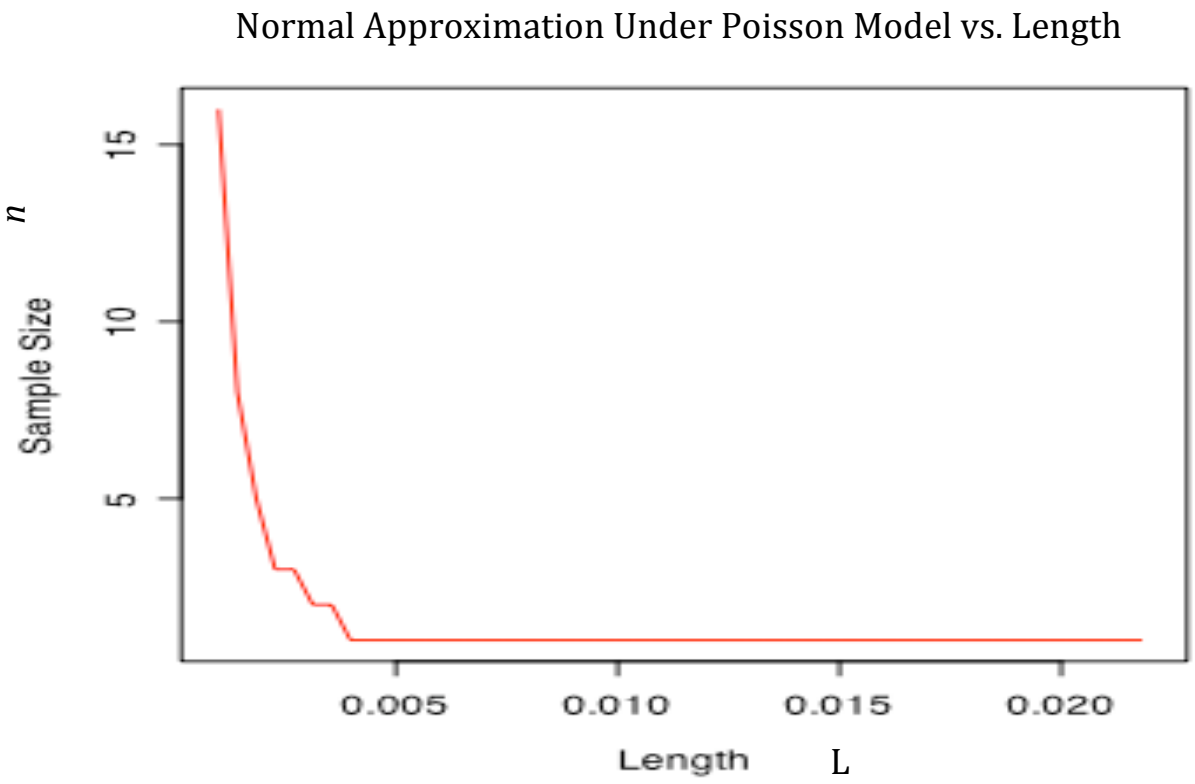
Figure 5:

SSD for Normal Approximation of SS Under The Binomial Model Posterior



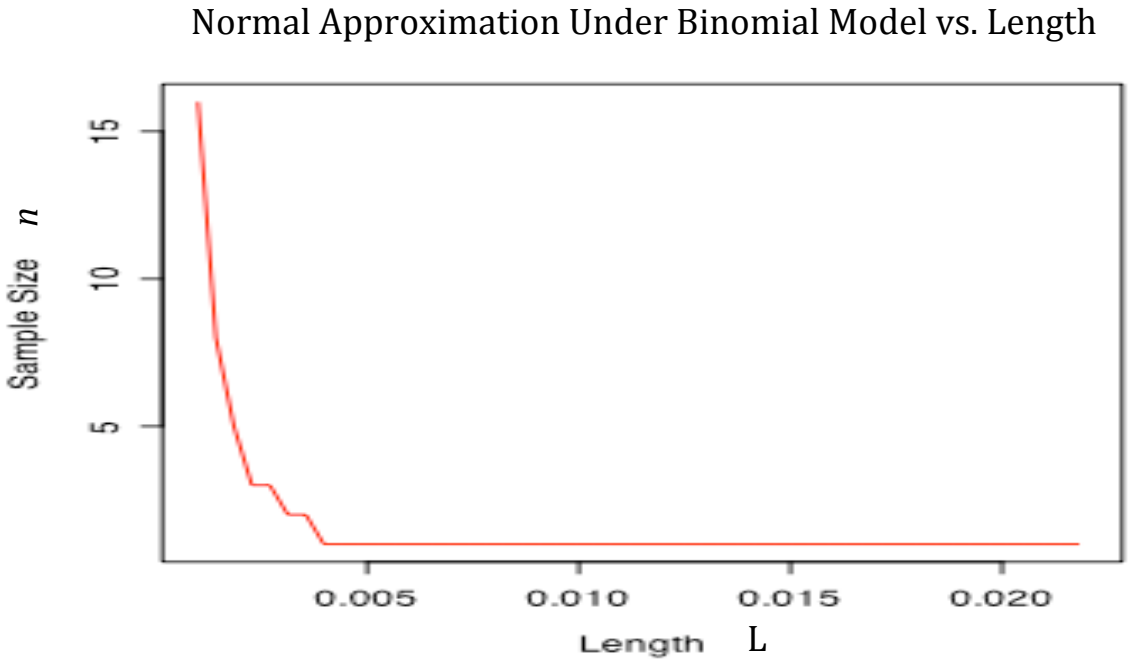
Figures 6 and 7 demonstrate a sequence of values for length varying from .001 to .022, where the sample size is the smallest integer value that solves the integral at greater than $1-\alpha$. So although the $1-\alpha$ values for each sample size versus length are not the same, the integer value that solves for the equation is the same in both Figures 6 and 7. This also means that the ratio of sample size between the two different models will be 1 for all values of length.

Figure 6:



Figures 6 and 7 have the same shape as Figures 1 and 2. This implies that the sample size vs. length under the frequentist method used and the Bayesian method are producing similar results. Figure 7 follows on the next page, however it is easily noticed that Figures 6 and 7 are quite similar, and this is due to integer values for sample size being used as input.

Figure 7:



As the length gets smaller than L=0 .002 in both Figures 6 and 7, the sample size rapidly increases.

Section 3:

Bayesian Estimates for the Posterior Distribution of $\theta|y$

Simply using the posterior distribution from the normal approximation to the poisson model in section 2.1, we are able to set up the relationship below,

$$\frac{14.28 + n\bar{b}}{\sqrt{.0996 + \sum_{i=1}^{n} y_i}} \geq \frac{\zeta_{\alpha/2}(2)}{L}.$$

The estimates .0996 and 14.29 are estimates for η and δ respectively from the prior distribution $\theta \sim Gamma(\eta, \delta)$. The estimates for η and δ are calculated in section 2.1 of the Appendix. Simple algebraic manipulation of the above formula leads to the following inequality for n,

$$n \geq \frac{\xi_{\alpha/2}(2)\sqrt{.0996 + \sum_{i=1}^{n} y_i} - 14.28L}{L\bar{b}}.$$

In Table A below, if in the above inequality for $n \geq \frac{\xi_{\alpha/2}(2)\sqrt{.0996 + \sum_{i=1}^{n} y_i} - 14.28L}{L\bar{b}}$ we allow the value for $\xi_{\alpha/2}$ to vary, while setting $\sum_{i=1}^{n} y_i$ =4226 as an initial estimate, holding $\alpha = .05$ as fixed, and L=.001 as fixed the corresponding values for sample size are given.

In Table B below, similarly if we use the inequality $n \geq \frac{\xi_{\alpha/2}(2)\sqrt{.0996 + \sum_{i=1}^{n} y_i} - 14.28L}{L\bar{b}}$ and hold $\xi_{\alpha/2}$ constant at 1.96 and allow length to vary, the corresponding values for sample size are given. For a value of $\xi_{\alpha/2}$ slightly greater than 1.96, the choice of constant value in sections 1 and 2, we have a corresponding sample size of 36.55. This value is larger than the nominal sample size under the frequentist method and previous Bayesian method where the values of $\bar{y}$ were averaged over. Also, it is noticed that as the length varies (Table B) the values closely mimic the values depicted by Figures 1 and 6, these figures correspond to the frequentist poisson approximation and the Bayesian poisson approximation respectively.

Table A:

$Z_{\alpha/2}$	Sample Size
1.28	23.63
1.38	25.47
1.48	27.32
1.58	29.16
1.68	31.01
1.78	32.86
1.88	34.7
1.98	36.55
2.08	38.4
2.18	40.24
2.28	42.09
2.38	43.93
2.48	45.78
2.58	47.63
2.68	49.47
2.78	51.32
2.88	53.16
2.98	55.01
3.08	56.86

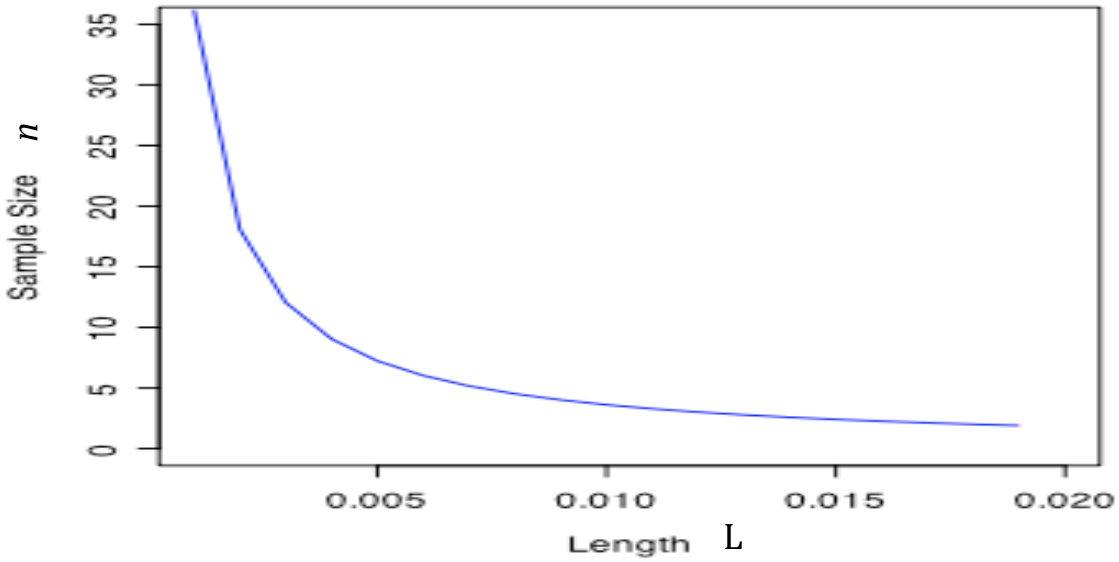
Table B:

Length	Sample Size
0.001	36.18
0.002	18.09
0.003	12.06
0.004	9.04
0.005	7.23
0.006	6.03
0.007	5.17
0.008	4.52
0.009	4.02
0.01	3.62
0.011	3.29
0.012	3.01
0.013	2.78
0.014	2.58
0.015	2.41
0.016	2.26
0.017	2.12
0.018	2.01
0.019	1.9

On the following page are the Figures 8 and 9 for varying length vs. sample size for the posterior distribution of θ under the two respective models.

Figure 8:

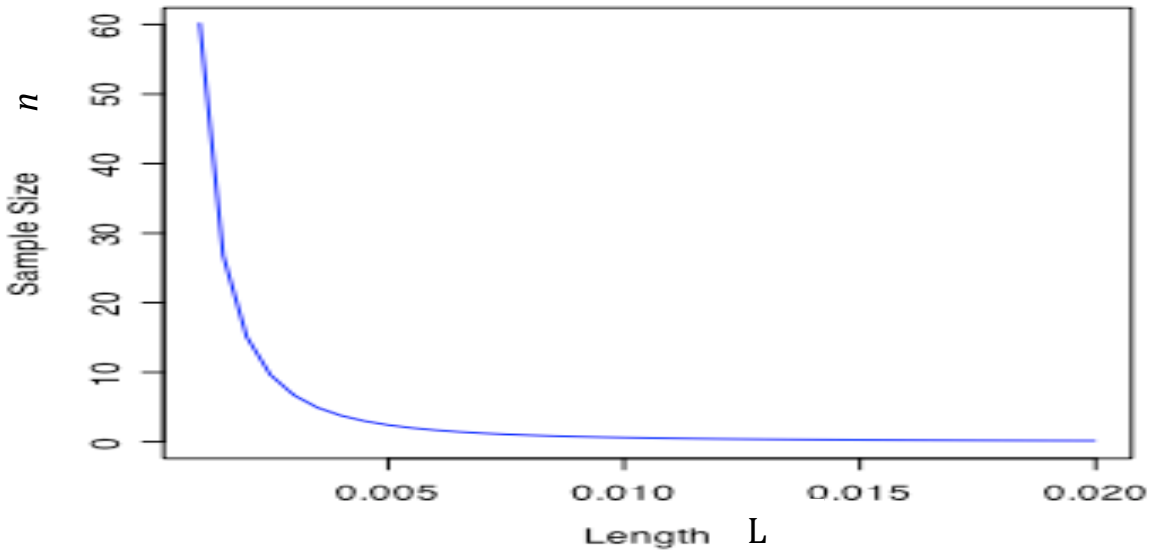
Normal Approximation to Posterior Distribution of θ in The Poisson Model



Figures 8 and 9 also have the same shape as Figures 1 and 2. Figure 8 (the Bayesian poisson method) has slightly larger values than Figure 1 (the frequentist poisson method), however, Figure 8 uses information for y that we do not have available to use before the sample is chosen. This is similarly the case if we compare Figure 9 (the Bayesian binomial method) to Figure 2 (the frequentist binomial method).

Figure 9:

Normal Approximation to Posterior Distribution of θ in The Binomial Model



4. Zero-Inflated Poisson Model

Zero-Inflated Poisson (ZIP) is a model to accommodate data with excess zeros. It assumes with probability p the only possible observation is 0 (or a true zero), and with probability $(1-p)$ a $Poisson(b_i\theta)$ $i=1, \ldots, n$, random variable is observed. If a company keeps accurate accounts receivable, there will be no errors, and this implies that there will be many zeros. Although the Poisson distribution includes zero, there are more zeros in the data than appropriate for using the standard Poisson model. The ZIP model will produce a more precise estimate. A thorough explanation of the ZIP model, fitting ZIP regression models and the simulated behavior of their properties in manufacturing defect data is given in (Lambert, 1992).

In the ZIP model, the responses Z_i are independent and $Z_i = \begin{cases} 0 & P(Z_i = 0) = p \\ 1 & P(Z_i = 1) = 1-p \end{cases}$ $i=1, \ldots, n$.

Our model for y_i under the zero inflation is given by $y_i = p + (1-p)b_i\theta\frac{e^{-b_i\theta}}{y_i!}$. Below is a table for the joint distribution for Z_i and y_i. The joint distribution defines the probability of events in terms of both Z_i and y_i.

	$y_i = 0$	$y_i = 1$
$Z_i = 0$	p	0
$Z_i = 1$	$(1-p)b_i\theta e^{-b_i\theta}$	$(1-p)b_i\theta\frac{e^{-b_i\theta}}{y_i!}$

The likelihood function under the Zero-Inflated Poisson model is therefore given by equation (20),

$$L(f(y)) = (1-p)^{n-n_0}\prod_{y_i=0}\left(p + (1-p)e^{-\theta b_i}\right)\prod_{y_i>0}\frac{(\theta b_i)^{y_i}}{y_i!}e^{-\theta b_i},\ i=1, \ldots, n. \qquad (20)$$

Under this model, p and θ are nuisance parameters and need to be estimated. Estimates for p and θ, namely $\hat{p}$ and $\hat{\theta}$ can be estimated by $\begin{pmatrix}\hat{\theta}\\ \hat{p}\end{pmatrix} \sim N\left(\begin{pmatrix}\theta\\ p\end{pmatrix}, \sum_{\theta,p}\right)$ where $\sum_{\theta,p}$ is the covariance matrix of $\hat{p}$ and $\hat{\theta}$, and is calculated based on the second partial derivatives with respect to p and θ. The second partial derivatives are given as equations (21), (22) and (23). Detailed steps which led to these results are given in section 4 of the Appendix. The second partial derivatives are given by,

$$\frac{\partial}{\partial p^2} = -\frac{n-n_0}{(1-p)^2} - \sum_{y_i=0}\frac{\left(1-e^{-\theta b_i}\right)^2}{\left(p+(1-p)e^{-\theta b_i}\right)^2}, \tag{21}$$

$$\frac{\partial}{\partial \theta^2} = \frac{p(1-p)b_i^2 e^{-\theta b_i}}{\left[p+(1-p)b_i e^{-\theta b_i}\right]^2} - \sum_{y_i=0}\frac{y_i}{\theta^2}, \tag{22}$$

and

$$\frac{\partial}{\partial \theta \partial p} = \sum_{y_i=0}\frac{b_i e^{-\theta b_i}}{\left[p+(1-p)b_i e^{-\theta b_i}\right]^2}. \tag{23}$$

These are the elements of the Hessian matrix, and these will be used to solve for the covariance matrix $\hat{p}$ and $\hat{\theta}$ by taking the opposite of the Hessian matrix and then solving for the inverse. Once we have determined the covariance matrix, only the (1,1) element of the matrix is of interest, because it is the variance of theta. Detailed steps can be found in section 4 of the Appendix. The estimate of θ is approximately normal, and is given by,

$$\hat{\theta}_{ZIP} \sim aN\left(\theta, \frac{1}{n} * \frac{n\left[1-e^{-\theta\bar{b}}\right]\left(p+(1-p)e^{-\theta\bar{b}}\right)+n\left(1-e^{-\theta\bar{b}}\right)(1-p)^2}{\left(\frac{n\bar{y}}{\theta^2}-np(1-p)\bar{b}^2e^{-\theta\bar{b}}\right)\left(n\left[1-e^{-\theta\bar{b}}\right]\left(p+(1-p)e^{-\theta\bar{b}}\right)+n\left(1-e^{-\theta\bar{b}}\right)(1-p)^2\right)-\frac{\left(\bar{b}e^{-\theta\bar{b}}\right)^2}{\left[p+(1-p)e^{-\theta\bar{b}}\right]^2}}\right),$$

$$\text{where } n - n_0 = n\left[1 - \sum_{i=1}^{n} \frac{e^{-\theta b_i}}{n}\right]. \tag{24}$$

We need to get estimates $\hat{p}$ and $\hat{\theta}$ for the nuisance parameters p and θ before we can continue. Numerical methods were used due to the complexity of the problem. The EM-algorithm yielded estimates of $\hat{p}$ =.78882 and $\hat{\theta}$ =.01113, and the Nelder-Mead method yielded estimates of $\hat{p}$ =.79991 and $\hat{\theta}$ =.01143. This was calculated by first finding estimates for the Z_i 's. Finding estimates for the Z_i 's was necessary because Z_i is an indicator variable therefore the sequence would not converge well if it were not first estimated by the equations,

$$\hat{p} = \frac{\sum_{i=1}^{n} Z_i}{n} \text{ and } \hat{\theta} = \frac{\sum_{i=1}^{n} (1 - Z_i) y_i}{\sum_{i=1}^{n} (1 - Z_i) b_i}.$$

The Z_i 's follow a Bernoulli distribution $Z_i \sim Ber\left\{\frac{p}{p + (1-p)e^{-b_i\theta}}\right\}$, i =1, ... , n.

The expectation for the Z_i 's are given by the equations, $E\left[Z_i = 1 | y_i = 0\right] = \frac{p}{p + (1-p)e^{-b_i\theta}}$

and $E\left[Z_i = 1 | y_i > 0\right] = 0$.

The expectation for Z_i can then be plugged into the equations for $\hat{p}$ and $\hat{\theta}$, and then the numerical methods were performed and resulted in the estimates previously stated. After obtaining estimates for the nuisance parameters, the EM method is preferred to the Nelder-Mead method so the EM estimates were used in calculating the variance. $\hat{\theta}_{ZIP}$ is distributed as

$$\hat{\theta}_{ZIP} \sim aN\left(\theta, \frac{1}{n} * \frac{n\left[1-e^{-\theta\bar{b}}\right]\left(p+(1-p)e^{-\theta\bar{b}}\right)+n\left(1-e^{-\theta\bar{b}}\right)(1-p)^2}{\left(\frac{n\bar{y}}{\theta^2}-np(1-p)\bar{b}^2e^{-\theta\bar{b}}\right)\left(n\left[1-e^{-\theta\bar{b}}\right]\left(p+(1-p)e^{-\theta\bar{b}}\right)+n\left(1-e^{-\theta\bar{b}}\right)(1-p)^2\right)-\frac{\left(\bar{b}e^{-\theta\bar{b}}\right)^2}{\left[p+(1-p)e^{-\theta\bar{b}}\right]^2}}\right).$$

We call this variance "a" and use our estimates $\hat{\theta} = .01113$, $\hat{p} = .78882$, $\bar{b} = 7043$, $\bar{y} = 46.5$ to solve for $a = .00000167$, we use this factor in $a^* = \frac{a\bar{b}}{\theta}$ =1.687012 to investigate the sample size given in the ZIP model in comparison to the sample size under the standard poisson model. This way we will have results similar to those that were given under the frequentist poisson model, but they will be scaled by a factor a^* to make for ease in comparison of the two models. Supplementary calculations can be found in section 4 of the Appendix. The calculations for the sample size n under the ZIP model follow similarly to the frequentist method used in section 1, but beginning with

$$\hat{\theta}_{ZIP} \sim N\left(\theta, \frac{a}{n}\right). \tag{25}$$

Multiplying and dividing through by $a^* = \frac{a\bar{b}}{\theta}$ so that we will have a sample size that is easily compared to the Poisson frequentist model results in

$$\hat{\theta}_{ZIP} \sim aN\left(\theta, \frac{\theta a^*}{\bar{b}n}\right). \tag{26}$$

An interval for $\hat{\theta}_{ZIP}$ will be

$$\Pr ob\left\{\theta : \left|\frac{\hat{\theta}_{ZIP}-\theta}{\sqrt{\frac{\theta a^*}{\bar{b}n}}}\right| \le \zeta_{\alpha/2}\right\} \ge 1-\alpha. \tag{27}$$

After some algebraic manipulation, with appropriate intermediate steps found in section 4 of the Appendix, the length of the interval under the ZIP model is

$$L = \sqrt{\left(2\hat{\theta}_{ZIP} + \frac{\zeta^2_{\alpha/2} a^*}{\bar{b} n}\right)^2 - 4\hat{\theta}^2_{ZIP}} \,. \tag{28}$$

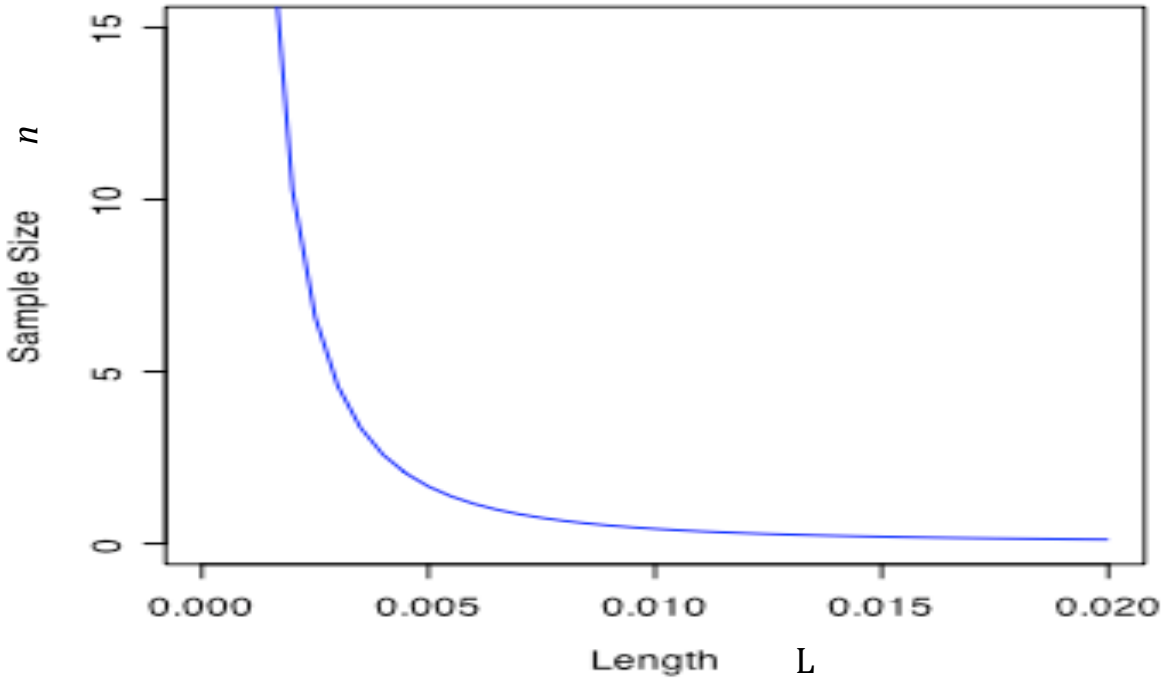
After rearranging the length equation to solve for n, we are left with equation (29),

$$n : \frac{a^* \zeta^2_{\alpha/2} \left(2\hat{\theta}_{ZIP} \pm \sqrt{4\hat{\theta}_{ZIP} + L^2}\right)}{L^2 \bar{b}} \,. \tag{29}$$

The supporting algebra is found in section 4 of the Appendix. It is noticed that the interval for n under the Zero-Inflated Poisson model is very similar to the interval for n under the frequentist Poisson model. The only difference between the two equations for n is a factor a*. Figure 10 below displays the Zero-Inflated Poisson model comparing the sample size vs. length.

Figure 10:

SSD Under The Zero-Inflated Poisson Model



The graph for the ZIP model (Figure 10) has a slightly higher sample size for any given length than under the poisson frequentist model (Figure 1). Under the ZIP model we expect there to be increased precision, however, this increased precision requires a larger sample size to attain.

Conclusion:

We proposed a method using the Zero-Inflated Poisson (ZIP) model which explicitly considers zero versus non-zero errors. This model is favorable due to the excess zeros that are present in auditing data that the standard poisson model does not account for. This method could easily be extended for data similar to accounting populations. The estimated necessary sample size was larger under the ZIP model than under the standard poisson model. However, due to the excess zeros in the data set, it would be reasonable to assume that the larger sample size is necessary for increased precision. Further research and investigation is needed to examine more precisely the benefits under the ZIP model. Below is a table summarizing the results of the methods from the four sections of this paper when the length is L=0.001, α is fixed at $\alpha = 0.05$, $\theta_0 = .007$, and $\bar{b} = 7043$.

	Poisson	Binomial	ZIP
Frequentist	15.29	22.72	41.02
Bayesian θ	16	16	
Bayesian $\theta \mid y$	36.18		

The ratio of the poisson frequentist sample size to the binomial frequentist sample size is 0.67. Figure 3 in section 1 of the paper depicted the frequentist binomial sample size being larger, until the length of the interval reached 0.013. The posterior distribution of $\theta \mid y$ for the poisson model had a nominal sample size of $n = 36.18$, this was the value closest to our ZIP sample size estimate of $n = 41.02$. However, as stated in section 3 of this paper, information for y was used that would not be available before deciding the number of accounts to select for the sample. The estimates for the sample size under the Bayesian method where Monte Carlo integration was implemented resulted in a sample size of 16 under both the poisson and binomial models. This was because we

chose the first integer sample size with a length equal to 0.001 to have greater than .95 confidence as our optimal sample size.

REFERENCES

Baker, Robert L., Copeland, Ronald M., (1979). Evaluation of the Stratified Regression Estimator for Auditing Accounting Populations*. Journal of Accounting Research,* **17**: 606-617.

Berg, Nathan. (2006). A Simple Bayesian Procedure for Sample Size Determination in an Audit of Property Value Appraisals. *Real Estate Economics,* **34**:133-155.

Casella, G., Berger, R. L., (2002). Statistical Inference 2nd Ed. Duxbury Press.

Cockburn, I. M., Puterman, M. L., and Wang, P. (1998). Analysis of Patent Data – A Mixed-Poisson-Regression-Model Approach. *Journal of Business & Economic Statistics*, **16**: 27-41

Higgins, H. N., Nandram, B., (2009). Monetary Unit Sampling: Improving Estimation of the Total Audit Error. *Advances in Accounting, Incorporating Advances in International Accounting*, 25, 2:174-182

Jiroutek, M. R., Muller, K. E., Kupper, Lawrence L., Stewart, P. W., (2003).A New Method for Choosing Sample Size for Confidence Interval-Based Inferences. *Biometrics,* **59**: 580-590.

Kaplan, Robert S., (1973). Statistical Sampling in Auditing with Auxiliary Information Estimators*. Journal of Accounting Research,* **11**: 238-258.

Knight, P., (1979). Statistical Sampling in Auditing: An Auditor's Viewpoint. *Journal of the Royal Statistical Society,* **28**: 253-266.

Lambert, D., (1992).Zero-Inflated Poisson Regression, With an Application to Defects in Manufacturing.*Technometrics,* **34**: 1-14

Lohr, S. L., (1999). Sampling: Design and Analysis. New York: Duxbury Press.

McCray, John H., (1984). A Quasi-Bayesian Audit Risk Model for Dollar Unit Sampling.*The Accounting Review,* **LIX, No. 1:** 35-41.

Menzefricke, U., (1983). On Sampling Plan Selection with Dollar-Unit Sampling.*Journal of Accounting Research,* **21:** 96-105.

Menzefricke, U., (1984). Using Decision Theory for Planning Audit Sample Size with Dollar Unit Sampling. *Journal of Accounting Research,* **22:** 570-587.

Ponemon, Lawrence A., Wendell, John P., (1995).Judgemental Versus Random Sampling in Auditing: An Experimental Investigation. *Auditing: A Journal of Practice and Theory,* **14**: 17-34.

Rohrback, Kermit J., (1986). Monetary Unit Acceptance Sampling. *Journal of Accounting Research.* **24:** 127-150.

Sahu, S. K., Smith, T. M. F. (2006). A Bayesian Method of Sample Size Determination with Practical Applications. *Journal of the Royal Statistical Society,* **169**: 235-253.

Wang, F., and Gelfand, A. E., (2002).A Simulation-Based Approach to Bayesian Sample Size Determination for Performance Under a Given Model and for SepartingModels. *Statistical Science,* **17**: 193-208

APPENDIX

Section 1:

1.1 Frequentist Approximation to the Poisson model:

$y_i \sim Poisson(b_i\theta)$, i=1, … , n.

(1) $\Pr ob\left\{\theta : \left|\frac{\bar{Y} - \bar{b}\,\theta}{\sqrt{\bar{b}\,\theta/n}}\right| \le \xi_{\alpha/2}\right\} \ge 1-\alpha$

$$\Pr ob\left\{\theta : \bar{b}^2\theta^2 - \theta\left(2\bar{b}\,\bar{Y} + \xi^2_{\alpha/2}\frac{\bar{b}}{n}\right) + \bar{Y}^2 \le 0\right\} \ge 1-\alpha$$

The length of the interval for theta will be:

(2) $$L = \frac{2\sqrt{\bar{b}^2\left(2\bar{Y} + \frac{\xi^2_{\alpha/2}}{n}\right)^2 - 4\bar{b}^2\bar{Y}^2}}{2\bar{b}^2}$$

$$L^2 = \frac{\left(2\bar{Y} + \frac{\xi^2_{\alpha/2}}{n}\right)^2 - 4\bar{Y}^2}{\bar{b}^2}$$

$$L^2 = \frac{4\bar{Y}^2 + 4\bar{Y}\frac{\xi^2_{\alpha/2}}{n} + \frac{\xi^4_{\alpha/2}}{n^2} - 4\bar{Y}^2}{\bar{b}^2}$$

$$L^2\bar{b}^2n^2 - 4\bar{Y}\xi^2_{\alpha/2}n - \xi^4_{\alpha/2} = 0$$

$\bar{Y}$ is estimated by $E(\bar{Y}) = \bar{b}\,\theta_0$

$$L^2\bar{b}^2n^2 - 4\bar{b}\,\theta_0\xi^2_{\alpha/2}n - \xi^4_{\alpha/2} = 0$$

After substituting in for $\bar{Y}$, we can solve for n:

$$n = \frac{4\bar{b}\,\theta_0 \xi^2_{\alpha/2} \pm \sqrt{16\bar{b}^2\theta_0^2\xi^4_{\alpha/2} + 4L^2\bar{b}^2\xi^4_{\alpha/2}}}{2\bar{b}^2L^2}$$

(3) $$n = \frac{\xi^2_{\alpha/2}\left(2\theta_0 \pm \sqrt{4\theta_0^2 + L^2}\right)}{\bar{b}L^2}$$

1.2 Frequentist Approximation to the Binomial model:

$y_i \sim Binomial(b_i, \theta)$, i=1, ... , n.

(4) $$\Pr ob\left\{\theta : \left|\frac{\bar{Y} - \bar{b}\,\theta}{\sqrt{\bar{b}\,\theta(1-\theta)/n}}\right| \le \xi_{\alpha/2}\right\} \ge 1-\alpha$$

$$\Pr ob\left\{\theta : \theta^2\left(\bar{b}^2 + \xi^2_{\alpha/2}\frac{\bar{b}}{n}\right) - \theta\left(2\bar{b}\bar{Y} + \xi^2_{\alpha/2}\frac{\bar{b}}{n}\right) + \bar{Y}^2 \le 0\right\} \ge 1-\alpha$$

The length of the interval for theta will be:

(5) $$L = \frac{\sqrt{\bar{b}^2\left(2\bar{Y} + \frac{\xi^2_{\alpha/2}}{n}\right)^2 - 4\left(\bar{b}^2 + \xi^2_{\alpha/2}\frac{\bar{b}}{n}\right)\bar{Y}^2}}{\bar{b}^2 + \xi^2_{\alpha/2}\frac{\bar{b}}{n}}$$

$$L^2 = \frac{\bar{b}^2\left(2\bar{Y} + \frac{\xi^2_{\alpha/2}}{n}\right)^2 - 4\left(\bar{b}^2 + \xi^2_{\alpha/2}\frac{\bar{b}}{n}\right)\bar{Y}^2}{\bar{b}^2\left(\bar{b} + \frac{\xi^2_{\alpha/2}}{n}\right)^2}$$

$E(\bar{Y}) = \bar{b}\,\theta_0$ And the expectation of $E(\bar{Y}^2) = Var(\bar{Y}) + E(\bar{Y})^2 = \bar{b}\,\theta_0(1-\theta_0)/n + (\bar{b}\,\theta_0)^2$

This implies after filling in for $E(\bar{Y})$ and $E(\bar{Y}^2)$ that:

$$L^2 = \frac{\bar{b}^2\left(2\bar{b}\theta_0 + \frac{\zeta^2_{\alpha/2}}{n}\right)^2 - 4\left(\bar{b}^2 + \zeta^2_{\alpha/2}\frac{\bar{b}}{n}\right)\left(\frac{\bar{b}\theta_0(1-\theta_0)}{n} + \left(\bar{b}\theta_0\right)^2\right)}{\bar{b}^2\left(\bar{b} + \frac{\zeta^2_{\alpha/2}}{n}\right)^2}$$

$$L^2 = \frac{\bar{b}^2\left(2\bar{b}\theta_0 + \frac{\zeta^2_{\alpha/2}}{n}\right)^2 - 4\left(\bar{b}^2 + \zeta^2_{\alpha/2}\frac{\bar{b}}{n}\right)\left(\frac{\bar{b}\theta_0(1-\theta_0)}{n} + \left(\bar{b}\theta_0\right)^2\right)}{\bar{b}^2\left(\bar{b} + \frac{\zeta^2_{\alpha/2}}{n}\right)^2}$$

After multiplying through top and bottom by n^2 the resulting equation is:

$$L^2 = \frac{\left(2\bar{b}\theta_0 n + \zeta^2_{\alpha/2}\right)^2}{\left(\bar{b}n + \zeta^2_{\alpha/2}\right)^2} - \frac{4n\left(\bar{b}\theta_0(1-\theta_0) + n\left(\bar{b}\theta_0\right)^2\right)}{\bar{b}\left(\bar{b} + \frac{\zeta^2_{\alpha/2}}{n}\right)}$$

$$L^2 = \frac{\left(2\bar{b}\theta_0 n + \zeta^2_{\alpha/2}\right)^2}{\left(\bar{b}n + \zeta^2_{\alpha/2}\right)^2} - \frac{4n\theta_0\left(1 + \theta_0\left(n\bar{b} - 1\right)\right)}{\bar{b}n + \zeta^2_{\alpha/2}}$$

Simply rearranging the equation and taking the square root of both sides results in:

$$\frac{2\bar{b}\theta_0 n + \zeta^2_{\alpha/2}}{\bar{b}n + \zeta^2_{\alpha/2}} = \sqrt{L^2 + \frac{4n\theta_0\left(1 + \theta_0\left(n\bar{b} - 1\right)\right)}{\bar{b}n + \zeta^2_{\alpha/2}}}$$

We will call the square root term “a”

$$(6)\quad a = \sqrt{L^2 + \frac{4n\theta_0\left(1 + \theta_0\left(n\bar{b} - 1\right)\right)}{\bar{b}n + \zeta^2_{\alpha/2}}}$$

And we continue solving the equation for n:

$$\frac{2\bar{b}\,\theta_0 n + \zeta^2_{\alpha/2}}{\bar{b}n + \zeta^2_{\alpha/2}} = a$$

$$2\bar{b}\,\theta_0 n - a\bar{b}n = a\zeta^2_{\alpha/2} - \zeta^2_{\alpha/2}$$

$$(7)\quad n = \left(\frac{a-1}{2\bar{b}\,\theta_0 - a\bar{b}}\right)\zeta^2_{\alpha/2}$$

And proceed using a simple iterative algorithm to get values for n for varying lengths.

1.3 Frequentist Approximation to the simpler Binomial model:

$$(8)\quad \Pr ob\left\{\theta : \left|\frac{X - n\bar{b}\,\theta}{\sqrt{n\bar{b}\,\theta(1-\theta)}}\right| \le \zeta_{\alpha/2}\right\} \ge 1-\alpha$$

Let $n\bar{b} = n^*$

$$(9)\quad \Pr ob\left\{\theta : \left|\frac{X - n^*\theta}{\sqrt{n^*\theta(1-\theta)}}\right| \le \zeta_{\alpha/2}\right\} \ge 1-\alpha$$

$$\Pr ob\left\{\theta : \theta^2\left(n^{*2} - \zeta^2_{\alpha/2}n^*\right) - \theta\left(2Xn^* - \zeta^2_{\alpha/2}n^*\right) + X^2 \le 0\right\} \ge 1-\alpha$$

The length of the interval for theta will be:

$$L = \frac{2\sqrt{\left(2Xn^* + \zeta^2_{\alpha/2}n^*\right)^2 - 4X^2\left(n^{*2} - \zeta^2_{\alpha/2}n^*\right)}}{2\left(n^{*2} - \zeta^2_{\alpha/2}n^*\right)}$$

$$L = \frac{\sqrt{4X^2n^{*2} + 4Xn^{*2}\zeta^2_{\alpha/2} + \zeta^4_{\alpha/2}n^{*2} - 4X^2n^{*2} + 4X^2n^*\zeta^2_{\alpha/2}}}{n^*\left(n^* - \zeta^2_{\alpha/2}\right)}$$

$$L = \frac{\sqrt{4Xn^{*2}\zeta^2_{\alpha/2} + \zeta^4_{\alpha/2}n^{*2} + 4X^2 n^* \zeta^2_{\alpha/2}}}{n^*\left(n^* - \zeta^2_{\alpha/2}\right)}$$

Take out 2n*Z from under the radical, the n* in the numerator then cancels with the n* in the denominator.

$$(10)\ L = \frac{2\zeta_{\alpha/2}\sqrt{\left(\frac{X\left(n^* - X\right)}{n^*}\right) + \frac{1}{4}\zeta^2_{\alpha/2}}}{n^* - \zeta^2_{\alpha/2}}$$

Now to use $E(X)$ as an estimate for X, we know that $X \sim Binomial(n^*, \theta) \Rightarrow E(X) = n^*\theta$

And after substituting in we would have $n^* - \zeta^2_{\alpha/2} = \frac{2\zeta_{\alpha/2}}{L}\sqrt{\left(\frac{n^*\theta\left(n^* - n^*\theta\right)}{n^*}\right) + \frac{1}{4}\zeta^2_{\alpha/2}}$

However, a better approximation would involve:

$$E\left(X\left(n^* - X\right)\right) = n^*E(X) - E\left(X^2\right) = n^{*2}\theta - \left[n^*\theta(1-\theta) - n^{*2}\theta^2\right] = n^{*2}\theta(1-\theta) - n^*\theta(1-\theta) = \left(n^{*2} - n^*\right)\theta(1-\theta))$$

And after substitution we are left with $n^* - \zeta^2_{\alpha/2} = \frac{2\zeta_{\alpha/2}}{L}\sqrt{\left(\frac{\left(n^{*2} - n^*\right)\theta(1-\theta))}{n^*}\right) + \frac{1}{4}\zeta^2_{\alpha/2}}$

$$n^* - \zeta^2_{\alpha/2} = \frac{2\zeta_{\alpha/2}}{L}\sqrt{\left(n^* - 1\right)\theta(1-\theta)) + \frac{1}{4}\zeta^2_{\alpha/2}}$$

Let $n^* - 1 = n^{**}$, and add and subtract 1 from the left-hand side of the equation:

$$\left(n^* - 1\right) - \zeta^2_{\alpha/2} + 1 = \frac{2\zeta_{\alpha/2}}{L}\sqrt{\left(n^* - 1\right)\theta(1-\theta)) + \frac{1}{4}\zeta^2_{\alpha/2}}$$

$$n^{**} - \zeta^2_{\alpha/2} + 1 = \frac{2\zeta_{\alpha/2}}{L}\sqrt{n^{**}(\theta(1-\theta)) + \frac{1}{4}\zeta^2_{\alpha/2}}$$

Then square both sides:

$$n^{**2} + 2n^{**}\left(\zeta^2_{\alpha/2} - 1\right) + \left(\zeta^2_{\alpha/2} - 1\right)^2 = \frac{4\zeta^2_{\alpha/2}}{L^2}\left[n^{**}(\theta(1-\theta)) + \frac{1}{4}\zeta^2_{\alpha/2}\right]$$

$$n^{**2} + 2n^{**}\left[\left(\zeta^2_{\alpha/2} - 1\right) - \frac{2\zeta^2_{\alpha/2}(\theta(1-\theta))}{L^2}\right] + \left(\zeta^2_{\alpha/2} - 1\right)^2 - \frac{\zeta^4_{\alpha/2}}{L^2} = 0$$

This implies:

$$\Rightarrow A = \left(\zeta^2_{\alpha/2} - 1\right)^2 - \frac{\zeta^4_{\alpha/2}}{L^2}$$

$$\Rightarrow B = \left[\left(\zeta^2_{\alpha/2} - 1\right) - \frac{2\zeta^2_{\alpha/2}(\theta(1-\theta))}{L^2}\right]$$

$$n^{**2} + 2n^{**}B + A = 0$$

$$n^{**2} + 2n^{**}B + B^2 + A - B^2 = 0$$

$$\left(n^{**2} + B^2\right) = B^2 - A$$

$$n^{**} + B = \pm\sqrt{B^2 - A}$$

$$n^{**} = -B \pm \sqrt{B^2 - A}$$

$$n^{**} = -\left[\left(\zeta^2_{\alpha/2} - 1\right) - \frac{2\zeta^2_{\alpha/2}(\theta(1-\theta))}{L^2}\right] \pm \sqrt{\left[\left(\zeta^2_{\alpha/2} - 1\right) - \frac{2\zeta^2_{\alpha/2}(\theta(1-\theta))}{L^2}\right]^2 - \left(\zeta^2_{\alpha/2} - 1\right)^2 - \frac{\zeta^4_{\alpha/2}}{L^2}}$$

$$n^{**} = -\left(\zeta^2_{\alpha/2} - 1\right) + \frac{2\zeta^2_{\alpha/2}(\theta(1-\theta))}{L^2} \pm \frac{\zeta_{\alpha/2}}{L}\sqrt{-\left(4(\theta(1-\theta))\left(\zeta^2_{\alpha/2} - 1\right)\right) + \frac{4\zeta^2_{\alpha/2}(\theta(1-\theta))^2}{L^2} - \zeta^2_{\alpha/2}}$$

Section 2:

2.1 Bayesian Normal approximation, under the Poisson model to the Gamma Posterior Distribution:

$y_i \sim Poisson(b_i\theta)$, i=1, … ,$n$, which can be written as $Y^{(n)}|\theta \sim Poisson(b_i\theta)$ i=1, … ,n. In Bayesian statistics, this is called the likelihood.

We assume prior distribution $\theta \sim Gamma(\eta, \delta)$

The Posterior distribution $\theta \left| Y^{(n)} \right.$ is proportional to the Prior*Likelihood

$$P\left(\theta \left| Y^{(n)} \right.\right) \propto P(\theta) P\left(Y^{(n)} \left| \theta \right.\right)$$

$$P\left(\theta \left| Y^{(n)} \right.\right) \propto \theta^{\eta-1} e^{-\delta\theta} \theta^{\sum_1^N y_i} e^{-\left(\sum_1^N b_i\right)\theta}$$

$$P\left(\theta \left| Y^{(n)} \right.\right) \propto \theta^{\eta-1+\sum_1^N y_i} e^{-\left(\delta+\sum_1^N b_i\right)\theta}$$

$$\theta \left| Y^{(n)} \right. \sim Gamma\left(\eta + n\bar{y}, \delta + n\bar{b}\right)$$

Solving for η and δ given that $\theta_0 = .007$ and $\sigma_0 = .022$:

$$E(\theta_0) = \frac{\eta}{\delta} = .007$$

$$Var(\theta_0) = \frac{\eta}{\delta^2} = .022^2$$

$$\Rightarrow .007\delta - .00049\delta^2 = 0$$

$$\Rightarrow \delta : \frac{.007 \pm \sqrt{(.007)^2 - 0}}{2(.00049)}$$

$$\Rightarrow \delta = 14.28$$

now substitute in to solve for η:

$$\Rightarrow \eta = 14.28 * .007 = .09996$$

$$\theta \left| Y^{(n)} \right. \sim Gamma\left(.0096 + n\bar{y}, 14.28 + n\bar{b}\right)$$

2.2 Solving for n, using the Normal Approximation for the Posterior of θ under the Poisson Model:

(11) $\theta|y \sim aN\left(\frac{.0996+\sum_{i=1}^{n} y_i}{14.28+n\bar{b}}, \frac{.0996+\sum_{i=1}^{n} y_i}{\left(14.28+n\bar{b}\,\theta\right)^2}\right)$

$$\int_{y}\int_{\mu-\frac{L}{2}}^{\mu+\frac{L}{2}} \pi(\theta|y^{(n)})d\theta p(y^{(n)})dy^{(n)} \geq 1-\alpha$$

(12) $$\int_{y}\int_{\frac{.0996+\sum_{i=1}^{n} y_i}{14.28+n\bar{b}}-\frac{L}{2}}^{\frac{.0996+\sum_{i=1}^{n} y_i}{14.28+n\bar{b}}+\frac{L}{2}} \pi(\theta|y^{(n)})d\theta p(y^{(n)})dy^{(n)} \geq 1-\alpha$$

(13) $$\int_{y}\left\{\Phi\left(\frac{\frac{.0996+\sum_{i=1}^{n} y_i}{14.28+n\bar{b}}+\frac{L}{2}-\frac{.0996+\sum_{i=1}^{n} y_i}{14.28+n\bar{b}}}{\sqrt{\frac{.0996+\sum_{i=1}^{n} y_i}{\left(14.28+n\bar{b}\right)^2}}}\right)-\Phi\left(\frac{\frac{.0996+\sum_{i=1}^{n} y_i}{14.28+n\bar{b}}-\frac{L}{2}-\frac{.0996+\sum_{i=1}^{n} y_i}{14.28+n\bar{b}}}{\sqrt{\frac{.0996+\sum_{i=1}^{n} y_i}{\left(14.28+n\bar{b}\right)^2}}}\right)\right\} p\left(y^{(n)}\right)dy^{(n)} \geq 1-\alpha$$

(14) $$\int_{y}\left(2\Phi\left(\frac{L}{2}\left(\frac{14.28+n\bar{y}}{\sqrt{.0996+\sum_{i=1}^{n} y_i}}\right)\right)-1\right)p(y^{(n)})dy^{(n)} \geq 1-\alpha$$

2.3 Bayesian Normal approximation, under the Binomial model to the Beta Posterior Distribution:

$Y^{(n)}|\theta \sim Binomial(b_i,\theta)$, i=1, ... ,$n$.

$\theta \sim Beta(\phi,\omega)$

The Posterior distribution again is given by:

$$P\left(\theta|Y^{(n)}\right) \propto P(\theta)P\left(Y^{(n)}|\theta\right)$$

$$P\left(\theta|Y^{(n)}\right) \propto \theta^{\phi-1}(1-\theta)^{\omega-1}\theta^{\sum_1^N y_i}(1-\theta)^{\sum_1^N b_i-y_i}$$

$$P\left(\theta|Y^{(n)}\right) \propto \theta^{\phi-1+\sum_1^N y_i}(1-\theta)^{\omega-1+\sum_1^N b_i-y_i}$$

Solving for ϕ and ω given that $\theta_0 = .007$ and $\sigma_0 = .022$

$$E(\theta_0) = \frac{\phi}{\phi+\omega} = .007$$

$$\Rightarrow \phi = .007\phi + .007\omega$$

$$\Rightarrow .993\phi = +.007\omega$$

$$\Rightarrow \phi = +.00705\omega$$

$$Var(\theta_0) = \frac{\phi\omega}{(\phi+\omega)^2(\phi+\omega+1)} = .022^2$$

$$\Rightarrow \frac{\frac{\phi}{\phi+\omega} * \frac{\omega}{\phi+\omega}}{\phi+\omega+1} = .022^2$$

$$\Rightarrow \frac{(.007)(.993)}{\phi+\omega+1} = .022^2$$

$$\Rightarrow \frac{0.006951}{\phi+\omega+1} = .00048841$$

$$\Rightarrow 0.006951 = .00048841\phi + .00048841\omega + .00048841$$

$$\Rightarrow -\phi = \omega - 13.244169$$

Simply solving the system of equations results in ϕ=.09271 and ω =13.1514.

2.4 Solving for n, using the Normal Approximation for the Posterior of θ under the Binomial Model:

(15) $$\theta|y \sim aN\left(\frac{\phi+\sum_{i=1}^{n} y_i}{\phi+\omega+n\bar{b}}, \frac{\left(\phi+\sum_{i=1}^{n} y_i\right)\left(\omega+n\bar{b}-\sum_{i=1}^{n} y_i\right)}{(\phi+\omega+n\bar{b})^2(\phi+\omega+n\bar{b}+1)}\right)$$

$$\theta|y \sim aN\left(\frac{0.09271+\sum_{i=1}^{n} y_i}{0.09271+13.1514+612741}, \frac{\left(0.09271+\sum_{i=1}^{n} y_i\right)\left(13.1514+612741-\sum_{i=1}^{n} y_i\right)}{(0.09271+13.1514+612741)^2(0.09271+13.1514+612741+1)}\right)$$

(16) $$\theta|y \sim aN\left(\frac{0.09271+\sum_{i=1}^{n} y_i}{612{,}754.244}, \frac{\left(0.09271+\sum_{i=1}^{n} y_i\right)\left(612{,}754.244-\sum_{i=1}^{n} y_i\right)}{(612{,}754.244)^2(612{,}755.244)}\right)$$

$$\int_{y}\int_{\mu-\frac{L}{2}}^{\mu+\frac{L}{2}} \pi(\theta|y^{(n)})d\theta p(y^{(n)})dy^{(n)} \geq 1-\alpha$$

(17) $$\int_{y}\int_{\frac{0.09271+\sum_{i=1}^{n} y_i}{612{,}754.244}-\frac{L}{2}}^{\frac{0.09271+\sum_{i=1}^{n} y_i}{612{,}754.244}+\frac{L}{2}} \pi(\theta|y^{(n)})d\theta p(y^{(n)})dy^{(n)} \geq 1-\alpha$$

(18)

$$\int_{y}\left\{\Phi\left(\frac{\frac{0.09271+\sum_{i=1}^{n} y_i}{612{,}754.244}+\frac{L}{2}-\frac{0.09271+\sum_{i=1}^{n} y_i}{612{,}754.244}}{\sqrt{\frac{\left(0.09271+\sum_{i=1}^{n} y_i\right)\left(612{,}754.244-\sum_{i=1}^{n} y_i\right)}{(612{,}754.244)^2(612{,}755.244)}}}\right)-\Phi\left(\frac{\frac{0.09271+\sum_{i=1}^{n} y_i}{612{,}754.244}-\frac{L}{2}-\frac{0.09271+\sum_{i=1}^{n} y_i}{612{,}754.244}}{\sqrt{\frac{\left(0.09271+\sum_{i=1}^{n} y_i\right)\left(612{,}754.244-\sum_{i=1}^{n} y_i\right)}{(612{,}754.244)^2(612{,}755.244)}}}\right)\right\}p(y^{(n)})dy^{(n)} \geq 1-\alpha$$

(19) $$\int_{y}\left\{2\Phi\left(\frac{L}{2}\left(\frac{(612{,}754.244)\sqrt{(612{,}755.244)}}{\sqrt{\left(0.09271+\sum_{i=1}^{n} y_i\right)\left(612{,}754.244-\sum_{i=1}^{n} y_i\right)}}\right)\right)-1\right\}p(y^{(n)})dy^{(n)} \geq 1-\alpha$$

Section 4:

Zero-Inflated Poisson model:

First Solving for the second partial derivatives of the likelihood function:

The likelihood function is $L(f(y))=\prod_{y_i=0}\left(p+(1-p)e^{-\theta b_i}\right)\prod_{y_i>0}(1-p)\frac{(\theta b_i)^{y_i}}{y_i!}e^{-\theta b_i}$

(20) $$L(f(y))=(1-p)^{n-n_0}\prod_{y_i=0}\left(p+(1-p)e^{-\theta b_i}\right)\prod_{y_i>0}\frac{(\theta b_i)^{y_i}}{y_i!}e^{-\theta b_i}$$

$$\log(L(f(y)))=(n-n_0)\log(1-p)+\sum_{y_i=0}\log\left(p+(1-p)e^{-\theta b_i}\right)+\sum_{y_i=0}\log\left(\frac{(\theta b_i)^{y_i}}{y_i!}e^{-\theta b_i}\right)$$

$$= (n-n_0)\log(1-p) + \sum_{y_i=0}\log\left(p+(1-p)e^{-\theta b_i}\right) + \sum_{y_i=0} y_i \log(\theta b_i) - \theta b_i - \log y_i!$$

$$\frac{\partial}{\partial p} = -\frac{n-n_0}{1-p} + \sum_{y_i=0}\frac{1-e^{-\theta b_i}}{p+(1-p)e^{-\theta b_i}}$$

$$(21)\quad \frac{\partial}{\partial p^2} = -\frac{n-n_0}{(1-p)^2} - \sum_{y_i=0}\frac{\left(1-e^{-\theta b_i}\right)^2}{\left(p+(1-p)e^{-\theta b_i}\right)^2}$$

$$\frac{\partial}{\partial \theta} = \sum_{y_i=0}\frac{-(1-p)b_i e^{-\theta b_i}}{p+(1-p)e^{-\theta b_i}} + \sum_{y_i=0}\frac{y_i b_i}{\theta b_i} - \sum_{y_i=0} b_i$$

$$\frac{\partial}{\partial \theta} = \sum_{y_i=0}\frac{-(1-p)b_i e^{-\theta b_i}}{p+(1-p)e^{-\theta b_i}} + \sum_{y_i=0}\frac{y_i}{\theta} - \sum_{y_i=0} b_i$$

$$\frac{\partial}{\partial \theta^2} = \frac{\partial}{\partial \theta}\left[\sum_{y_i=0}\frac{-(1-p)b_i e^{-\theta b_i}}{p+(1-p)e^{-\theta b_i}} + \sum_{y_i=0}\frac{y_i}{\theta} - \sum_{y_i=0} b_i\right]$$

Let $\begin{aligned} u &= -(1-p)b_i e^{-\theta b_i} \rightarrow u' = (1-p)b_i^2 e^{-\theta b_i} \\ v &= p+(1-p)e^{-\theta b_i} \rightarrow v' = -(1-p)b_i e^{-\theta b_i} \end{aligned}$ then by the quotient rule:

$$\frac{\partial}{\partial \theta^2} = \frac{u'v - uv'}{v^2} - \sum_{y_i=0}\frac{y_i}{\theta^2}$$

$$\frac{\partial}{\partial \theta^2} = \frac{(1-p)b_i^2 e^{-\theta b_i}\left(p+(1-p)e^{-\theta b_i}\right) - \left[(1-p)b_i e^{-\theta b_i}\right]^2}{\left[p+(1-p)b_i e^{-\theta b_i}\right]^2} - \sum_{y_i=0}\frac{y_i}{\theta^2}$$

$$(22)\quad \frac{\partial}{\partial \theta^2} = \frac{p(1-p)b_i^2 e^{-\theta b_i}}{\left[p+(1-p)b_i e^{-\theta b_i}\right]^2} - \sum_{y_i=0}\frac{y_i}{\theta^2}$$

$$\frac{\partial}{\partial \theta \partial p} = \frac{\partial}{\partial \theta}\left[\sum_{y_i=0}\frac{-\left(1-e^{-\theta b_i}\right)^2}{\left(p+(1-p)e^{-\theta b_i}\right)^2}\right]$$

Let $\begin{aligned} u &= 1-b_i e^{-\theta b_i} \rightarrow u' = b_i e^{-\theta b_i} \\ v &= p+(1-p)e^{-\theta b_i} \rightarrow v' = -(1-p)b_i e^{-\theta b_i} \end{aligned}$ then by the quotient rule:

$$\frac{\partial}{\partial\theta\partial p} = \sum_{y_i=0} \frac{b_i e^{-\theta b_i}\left[p + (1-p)b_i e^{-\theta b_i}\right] + (1-p)b_i e^{-\theta b_i}\left[1 - b_i e^{-\theta b_i}\right]}{\left[p + (1-p)b_i e^{-\theta b_i}\right]^2}$$

$$\frac{\partial}{\partial\theta\partial p} = \sum_{y_i=0} \frac{b_i e^{-\theta b_i}\left(\left[p + (1-p)\right] + (1-p)\left[1 - b_i e^{-\theta b_i}\right]\right)}{\left[p + (1-p)b_i e^{-\theta b_i}\right]^2}$$

$$(23)\quad \frac{\partial}{\partial\theta\partial p} = \sum_{y_i=0} \frac{b_i e^{-\theta b_i}}{\left[p + (1-p)b_i e^{-\theta b_i}\right]^2}$$

The Hessian Matrix will therefore be:

$$\begin{pmatrix} \sum_{y_i=0} \frac{p(1-p)b_i^2 e^{-\theta b_i}}{\left[p + (1-p)e^{-\theta b_i}\right]^2} - \sum_{y_i>0} \frac{y_i}{\theta^2} & \sum_{y_i=0} \frac{b_i e^{-\theta b_i}}{\left[p + (1-p)e^{-\theta b_i}\right]^2} \\ \sum_{y_i=0} \frac{b_i e^{-\theta b_i}}{\left[p + (1-p)e^{-\theta b_i}\right]^2} & -\frac{n-n_0}{(1-p)^2} - \sum_{y_i=0} \frac{\left(1-e^{-\theta b_i}\right)^2}{\left(p + (1-p)e^{-\theta b_i}\right)^2} \end{pmatrix}$$

We first take the negative of the Hessian Matrix:

$$\begin{pmatrix} \sum_{y_i>0} \frac{y_i}{\theta^2} - \sum_{y_i=0} p(1-p)b_i^2 e^{-\theta b_i} & -\sum_{y_i=0} \frac{b_i e^{-\theta b_i}}{\left[p + (1-p)e^{-\theta b_i}\right]^2} \\ -\sum_{y_i=0} \frac{b_i e^{-\theta b_i}}{\left[p + (1-p)e^{-\theta b_i}\right]^2} & \frac{n-n_0}{(1-p)^2} + \sum_{y_i=0} \frac{\left(1-e^{-\theta b_i}\right)^2}{\left(p + (1-p)e^{-\theta b_i}\right)^2} \end{pmatrix}$$

Next, to obtain the inverse of this matrix:

$$A_{2X2}^{-1} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad-bc}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

$$ad - bc = \left(\sum_{y_i>0} \frac{y_i}{\theta^2} - \sum_{y_i=0} p(1-p)b_i^2 e^{-\theta b_i}\right)\left(\frac{n-n_0}{(1-p)^2} + \sum_{y_i=0} \frac{\left(1-e^{-\theta b_i}\right)^2}{\left(p + (1-p)e^{-\theta b_i}\right)^2}\right) - \left(\sum_{y_i=0} \frac{b_i e^{-\theta b_i}}{\left[p + (1-p)e^{-\theta b_i}\right]^2}\right)^2$$

$$\frac{1}{\left(\sum_{y_i>0}\frac{y_i}{\theta^2}-\sum_{y_i=0}p(1-p)b_i^2e^{-\theta b_i}\right)\left(\frac{n-n_0}{(1-p)^2}+\sum_{y_i=0}\frac{\left(1-e^{-\theta b_i}\right)^2}{\left(p+(1-p)e^{-\theta b_i}\right)^2}\right)-\left(\sum_{y_i=0}\frac{b_i e^{-\theta b_i}}{\left[p+(1-p)e^{-\theta b_i}\right]}\right)^2}\begin{pmatrix}\frac{n-n_0}{(1-p)^2}+\sum_{y_i=0}\frac{\left(1-e^{-\theta b_i}\right)^2}{\left(p+(1-p)e^{-\theta b_i}\right)^2} & \sum_{y_i=0}\frac{b_i e^{-\theta b_i}}{\left[p+(1-p)e^{-\theta b_i}\right]}\\ \sum_{y_i=0}\frac{b_i e^{-\theta b_i}}{\left[p+(1-p)e^{-\theta b_i}\right]} & \sum_{y_i>0}\frac{y_i}{\theta^2}-\sum_{y_i=0}p(1-p)b_i^2e^{-\theta b_i}\end{pmatrix}$$

The (1,1) element of this matrix will be:

$$\frac{\frac{n-n_0}{(1-p)^2}+\sum_{y_i=0}\frac{\left(1-e^{-\theta b_i}\right)^2}{\left(p+(1-p)e^{-\theta b_i}\right)^2}}{\left(\sum_{y_i>0}\frac{y_i}{\theta^2}-\sum_{y_i=0}p(1-p)b_i^2e^{-\theta b_i}\right)\left(\frac{n-n_0}{(1-p)^2}+\sum_{y_i=0}\frac{\left(1-e^{-\theta b_i}\right)^2}{\left(p+(1-p)e^{-\theta b_i}\right)^2}\right)-\left(\sum_{y_i=0}\frac{b_i e^{-\theta b_i}}{\left[p+(1-p)e^{-\theta b_i}\right]}\right)^2}$$

(24) We can use $n-n_0=n\left[1-\sum_{i=1}^{n}\frac{e^{-\theta b_i}}{n}\right]$, after substituting this into the equation:

$$\frac{\frac{n\left[1-\sum_{i=1}^{n}\frac{e^{-\theta b_i}}{n}\right]}{(1-p)^2}+\sum_{y_i=0}\frac{\left(1-e^{-\theta b_i}\right)^2}{\left(p+(1-p)e^{-\theta b_i}\right)^2}}{\left(\sum_{y_i>0}\frac{y_i}{\theta^2}-\sum_{y_i=0}p(1-p)b_i^2e^{-\theta b_i}\right)\left(\frac{n\left[1-\sum_{i=1}^{n}\frac{e^{-\theta b_i}}{n}\right]}{(1-p)^2}+\sum_{y_i=0}\frac{\left(1-e^{-\theta b_i}\right)^2}{\left(p+(1-p)e^{-\theta b_i}\right)^2}\right)-\left(\sum_{y_i=0}\frac{b_i e^{-\theta b_i}}{\left[p+(1-p)e^{-\theta b_i}\right]}\right)^2}$$

Although $E\left(X_i^2\right)\neq E\left(X_i\right)^2$, it is approximately, so we will use $E\left(X_i\right)^2$ to substitute into our equations.

$$\frac{\frac{n\left[1-\frac{ne^{-\theta\bar{b}}}{n}\right]}{(1-p)^2}+\frac{n\left(1-e^{-\theta\bar{b}}\right)^2}{\left(p+(1-p)e^{-\theta\bar{b}}\right)^2}}{\left(\frac{n\bar{y}}{\theta^2}-np(1-p)\bar{b}^2e^{-\theta\bar{b}}\right)\left(\frac{n\left[1-\frac{ne^{-\theta\bar{b}}}{n}\right]}{(1-p)^2}+\frac{n\left(1-e^{-\theta\bar{b}}\right)^2}{\left(p+(1-p)e^{-\theta\bar{b}}\right)^2}\right)-\left(\frac{\bar{b}e^{-\theta\bar{b}}}{\left[p+(1-p)e^{-\theta\bar{b}}\right]}\right)^2}$$

Using common denominator $(1-p)^2\left(p+(1-p)e^{-\theta\bar{b}}\right)^2$

$$\frac{n\left[1-e^{-\theta\bar{b}}\right]\left\{p+(1-p)e^{-\theta\bar{b}}\right\}+n\left\{1-e^{-\theta\bar{b}}\right\}(1-p)^2}{\left(\frac{n\bar{y}}{\theta^2}-np(1-p)\bar{b}^2e^{-\theta\bar{b}}\right)\left(n\left[1-e^{-\theta\bar{b}}\right]\left\{p+(1-p)e^{-\theta\bar{b}}\right\}+n\left\{1-e^{-\theta\bar{b}}\right\}(1-p)^2\right)-\left(\frac{\bar{b}e^{-\theta\bar{b}}}{\left[p+(1-p)e^{-\theta\bar{b}}\right]^2}\right)^2}$$

$$\frac{n\left[1-e^{-\theta\bar{b}}\right]\left\{p+(1-p)e^{-\theta\bar{b}}\right\}+n\left\{1-e^{-\theta\bar{b}}\right\}(1-p)^2}{\left(\frac{n\bar{y}}{\theta^2}-np(1-p)\bar{b}^2e^{-\theta\bar{b}}\right)\left(n\left[1-e^{-\theta\bar{b}}\right]\left\{p+(1-p)e^{-\theta\bar{b}}\right\}+n\left\{1-e^{-\theta\bar{b}}\right\}(1-p)^2\right)-\frac{\left(\bar{b}e^{-\theta\bar{b}}\right)^2}{\left[p+(1-p)e^{-\theta\bar{b}}\right]^4}}$$

We call this "a" and use our estimates $\hat{\theta}=.01113$, $\hat{p}=.78882$, $\bar{b}=7043$, $\bar{y}=46.5$ to solve for $a=.00000167$, we use this factor in $a^*=\frac{a\bar{b}}{\theta}$ =1.687012 to investigate the sample size given in the ZIP model in comparison to the sample size under the standard Poisson model.

(25) $\hat{\theta}_{ZIP}\sim N\left(\theta,\frac{a}{n}\right)$

(26) $\hat{\theta}_{ZIP}\sim aN\left(\theta,\frac{\theta a^*}{\bar{b}n}\right)$

(27) $\Pr ob\left\{\theta:\left|\frac{\hat{\theta}_{ZIP}-\theta}{\sqrt{\frac{\theta a^*}{\bar{b}n}}}\right|\le\zeta_{\alpha/2}\right\}\ge 1-\alpha$

$$\Pr ob\left\{\theta:\hat{\theta}_{ZIP}^2-2\hat{\theta}_{ZIP}\theta+\theta^2-\zeta_{\alpha/2}^2\frac{\theta a^*}{\bar{b}n}\le 0\right\}\ge 1-\alpha$$

The formula for the length will be:

(28) $L=\sqrt{\left(2\hat{\theta}_{ZIP}+\frac{\zeta_{\alpha/2}^2a^*}{\bar{b}n}\right)^2-4\hat{\theta}^2}$

$$L = \sqrt{\frac{4\hat{\theta}_{ZIP} a^* \zeta^2_{\alpha/2}}{\bar{b} n} + \frac{\zeta^4_{\alpha/2} a^{*2}}{\bar{b}^2 n^2}}$$

$$L^2 \bar{b}^2 n^2 - 4\hat{\theta}_{ZIP} \bar{b} a^* \zeta^2_{\alpha/2} n - a^{*2} \zeta^4_{\alpha/2} = 0$$

The interval for n will be:

$$n : \frac{4\hat{\theta}_{ZIP} a^* \bar{b} \zeta^2_{\alpha/2} \pm \sqrt{16\hat{\theta}_{ZIP} \bar{b}^2 a^{*2} \zeta^4_{\alpha/2} + 4L^2 \bar{b}^2 a^{*2} \zeta^4_{\alpha/2}}}{2L^2 \bar{b}^2}$$

$$n : \frac{2\hat{\theta}_{ZIP} a^* \bar{b} \zeta^2_{\alpha/2} \pm \bar{b} a^* \zeta^2_{\alpha/2} \sqrt{4\hat{\theta}_{ZIP} + L^2}}{L^2 \bar{b}^2}$$

(31) $$n : \frac{a^* \zeta^2_{\alpha/2} \left(2\hat{\theta}_{ZIP} \pm \sqrt{4\hat{\theta}_{ZIP} + L^2} \right)}{L^2 \bar{b}}$$

You'll notice that this is very similar to the interval for n in the frequentist Poisson case, except is it scaled by a factor of "a".