2022

Investigaion of the Gamma Hurdle Model for a Single Population Mean

Alissa Jacobs
Stephen F Austin State University, lissajacobs55@gmail.com

Investigaion of the Gamma Hurdle Model for a Single Population Mean

Creative Commons License

The Gamma Hurdle Model for a Single Population Mean

by

Alissa Jacobs, B.S.

Presented to the Faculty of the Graduate School of

Stephen F. Austin State University

In Partial Fulfillment

of the Requirements

For the Degree of

Master of Science

STEPHEN F. AUSTIN STATE UNIVERSITY

December 2022

The Gamma Hurdle Model for a Single Population Mean

by

Alissa Leigh Jacobs, B.S.

APPROVED:

Jacob Turner, Ph.D., Thesis Director

Robert Henderson, Ph.D., Committee Member

Kent Riggs, Ph.D., Committee Member

Jeremy Becnel, Ph.D., Committee Member

Sheryll Jerez, Ph.D.
Interim Dean of Research and Graduate Studies

ABSTRACT

A common issue in some statistical inference problems is dealing with a high frequency of zeroes in a sample of data. For many distributions such as the gamma, optimal inference procedures do not allow for zeroes to be present. In practice, however, it is natural to observe real data sets where nonnegative distributions would make sense to model but naturally zeroes will occur. One example of this is in the analysis of cost in insurance claim studies. One common approach to deal with the presence of zeroes is using a hurdle model. Most literary work on hurdle models will focus on modeling the frequency of zeros separate from the nonnegative values. While this approach has some advantages, it doesn't typically provide an interval estimator for the global population mean of the variable of interest. In this work we developed a Wald interval for the population mean assuming the gamma hurdle model. Using simulation, we investigated our procedure along with traditional interval estimation strategies such as the t-interval and bootstrap techniques and provided some recommendations and insights. Currently, we recommend the bootstrap t-interval overall as it has better coverage properties across all scenarios we considered.

ACKNOWLEDGEMENTS

First, I would like to thank my best friend and my wife for all the love and support throughout my graduate experience. Without them, I would not be where I am today. Next, I would like to thank Dr. Jacob Turner for all his help on this project. I would also like to thank Dr. Robert Henderson, Dr. Kent Riggs, and Dr. Jeremy Becnel for being a part of my thesis committee. Last, I would like to thank my friends and family for their support through the whole process. Graduate school would not have been the same without all the support.

CONTENTS

LIST OF FIGURES

1 INTRODUCTION

When performing standard statistical inference using a parametric approach, analysts decide on an appropriate probabilistic model that governs the population of interest. Each probabilistic model typically contains 1 or more parameters that describes various characteristics of the model including its mean and variance. With a model determined and an appropriate parameter of interest defined, various techniques can be applied to a random sample of data to estimate the parameter with a confidence interval or conduct a hypothesis test for a specific value. When learning about standard statistical inferences and hypothesis tests, there are discussions about robustness of the procedure. Robustness is the ability of a statistical inference procedure to perform at is predefined significance or confidence level when the assumptions made under the procedure are suspect. For example, the t-test is relatively robust to the normal assumption when the distribution of the data is mildly skewed or symmetric for small sample sizes. The t-test is also robust to more extreme departures from normality if the sample size is large enough. Another example where robustness is discussed happens with analysis of variance (ANOVA). The error rate when performing an ANOVA F-test is said to be robust to the constant variance assumption if the sample sizes are equal across all the groups being observed.

Another statistical concern that is similar to robustness is settings where data sets exhibit a large number of zeros. For some models, it is expected that zeros would occur, but it is not consistent with the rate observed in the actual data set. Other models do not allow for zeros at all and yet they are still observed. This "inflation of zeros" problem, raises a question of robustness to any statistical procedure being utilized under an assumed model. The most common discussion of this issue, and

how to handle it, is the zero inflated Poisson model. While the Poisson model allows for zeros, the question becomes how to model the data when there are much fewer or many more zeros present in a given data set than what would be expected under the traditional Poisson model. For the Poisson model with inflated zeros, the mean-variance relationship would be broken, and the variance would be larger than the mean. This is referred to as over-dispersion [5].

A common issue in the continuous cases arises for probability models that are defined for strictly positive values such as the Gamma, F, Pareto, and Weibull. In this case, zeroes are typically not allowed due to estimation strategies such as maximum likelihood estimates. If the data is continuous, which is common in cost data for insurance companies, then there tends to be a large amount of zeroes that cause a skeweness of the data with a longer right tail for the nonzero data. The issue of data being skewed in this way is common in health care and insurance. For instance, Figure 1.1 provides an example of cost data in the hospital setting by way of a histogram. In addition to the data, there are four proposed models: the normal, three parameter gamma, lognormal, and log-lognormal distributions. After parameter estimation, their density curves are overlaid as a quick visual comparison. One can see that raw data is clearly non-normal, right skewed, and contains a portion of zeroes, and significant outliers. The ability to model the zeros and the heavy tail simultaneously is of general concern. While nonparametric intervals are available for the median, the goal in many cases such as the cost data in Figure 1.1 is to provide inference on the population mean which directly informs of us of total cost for the population.

There are generally two approaches to trying to model data with zero inflation. The first method is using the traditional zero inflated model. This is common for discrete data. Consider the zero inflated model in terms of the Poisson distribution [3]. The common problem that zero inflated data has on the Poisson distribution is that there are more zeros than the distribution can accommodate with its mass

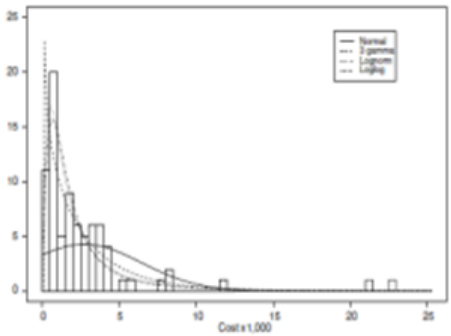

Figure 1.1: Cost data

function. The zeros that naturally occur in the Poisson model are referred to as sampling zeros. The additional zeros are seen as structural zeros from an additional random source.Hence, the zeroes are modeled using a mixture distribution. The Zero-inflated model for the Poisson Distribution:

$$P(Y=y) = \begin{cases} \pi + (1+\pi)e^{-\mu} & y = 0 \\ (1-\pi)\frac{e^{-\mu}(\mu)^y}{y!} & y = 1, 2, ... \end{cases}, \tag{1.1}$$

where μ is the mean of the Poisson distribution and π is the proportion of structural zeroes [3]. Note that the probability for $Y = 0$ is a mixture of a Bernoulli and Poisson mass function evaluated at 0. For all nonzero values of Y, the Poisson components of the mass function are scaled by $(1 - \pi)$ to ensure a proper mass function that sums to 1.

The second approach is the Hurdle Model. For this model, there is a mixed distribution as well. The difference is the 2 distributions of the mixture are on an entirely different support. Staying consistent with the previous example, a Poisson hurdle model is defined as

$$P(Y=y) = \begin{cases} \pi & y = 0 \\ \frac{(1-\pi)}{(1-e^{-\mu})}\frac{e^{-\mu}(\mu)^y}{y!} & y > 0 \end{cases}, \tag{1.2}$$

under this model, the 0 is modeled exclusively with a Bernoulli trial and a truncated

Poisson random variable models the non-zero outcomes. When applying the hurdle model to a continuous random variable, $f(y)$, with non-negative support, the observed random variable is a mixed type with a point mass at 0 and scaled continuous density function for non-negative values:

$$P(Y = y) = \begin{cases} \pi & y = 0 \\ (1 - \pi) f(y) & y > 0 \end{cases}, \tag{1.3}$$

One convenient result of this approach is that the zeroes can be modeled by themselves separate from the rest of the data. Since there are just two models, the Hurdle Model is convenient because you can use standard techniques for estimating the parameters in the model. There is also a clean interpretation for the rate of the zeroes and the parameters specified by $f(y)$. In more complex problems, such as when dealing with explanatory variables in a regression setting, the hurdle model can be extended where logistic regression is used to model the rate of zero occurrences, and a generalized linear model can be used to model the mean of the non-zero elements. This approach is quite applicable in areas such as insurance since understanding the frequency of customers who do not make a claim is of importance and can be interpreted from a logistic regression model. The generalized linear model would then be used to model the mean cost of claims made and investigating relationships here. Typical models for data such as costs use Gamma or Generalized Gamma hurdle models with a focus on the regression setting [6] [9].

If we let the random variable X with density function $f(X)$ denote the nonegative part of the hurdle model Y, the expected value of Y is

$$E(Y) = (1 - \pi) E(X). \tag{1.4}$$

1.1 Research Question

For this research, we will be investigating the Gamma Hurdle Model. While the application of the gamma hurdle model is heavily focused on the regression setting, we find it interesting that there is little to no discussion of how well the model works in simple cases as well as providing inference on the global mean rather than the mean of the nonzero part of the variable. For this reason we would like to investigate the Gamma Hurdle Model in the single population setting with no covariates. There is also very little discussion on whether a t-test procedure would be robust to a zero inflated model such as the gamma hurdle model. If it is robust, then under what degree of inflation becomes a key question as well for making decisions in practice.

The goal of this research can be summarized in two main points. The first is to develop a Wald confidence interval procedure for the global mean of the Gamma hurdle model under a single population setting. The second is to investigate the performance of the developed interval estimation assuming the Gamma hurdle model framework is true and then compare its performance against more standard techniques such as the t-interval and various bootstrap intervals.

2 Methods

In order to develop confidence intervals for the mean of a Gamma hurdle model, it will be helpful to discuss the general properties of estimating the mean of a traditional Gamma model in which there is no inflation of zeros. First, we will introduce the density function of the gamma and provide the necessary details to estimate parameters via maximum likelihood including Fisher's information for a mean parameterization of the gamma. Finally, we will introduce a Wald type interval to estimate the mean of a hurdle model and three standard based interval estimation strategies. The three strategies are two version of the bootstrap, the percentile bootstrap and the bootstrap t-interval, as well as the standard t-interval.

2.1 The Gamma Distribution

The density function of the gamma distribution along with its mean, variance, moment generating function, can be found in [8]. Under the shape and scale parameterization, the density function for the gamma random variable, X, is:

$$f(X;\alpha,\beta) = \frac{x^{\alpha-1}e^{\frac{-x}{\beta}}}{\Gamma(\alpha)\beta^{\alpha}} \tag{2.1}$$

where the Gamma function is

$$\Gamma(\alpha) = \int_0^{\infty} y^{\alpha-1}e^{-y}\,dy.$$

The mean and variance for Gamma distribution are:

$$\mathrm{E[X]} = \mu = \alpha\beta,$$

$$\mathrm{Var(X)} = \sigma^2 = \alpha\beta^2.$$

The Gamma distribution has a shape parameter, α, and a scale parameter, β. These two parameters determine what the density function will look like. As seen in Figure 2.1, when the α value increases from 1 to 2 to 10, the shape becomes less skewed. When the distribution has $\alpha = 1$, Figure 2.1 shows that there is more of an exponential shape. We see that the Gamma distribution can have a skewness to it and when the α increases, it pushes the shape of the distribution out and away from the exponential distribution to more of a bell shape.

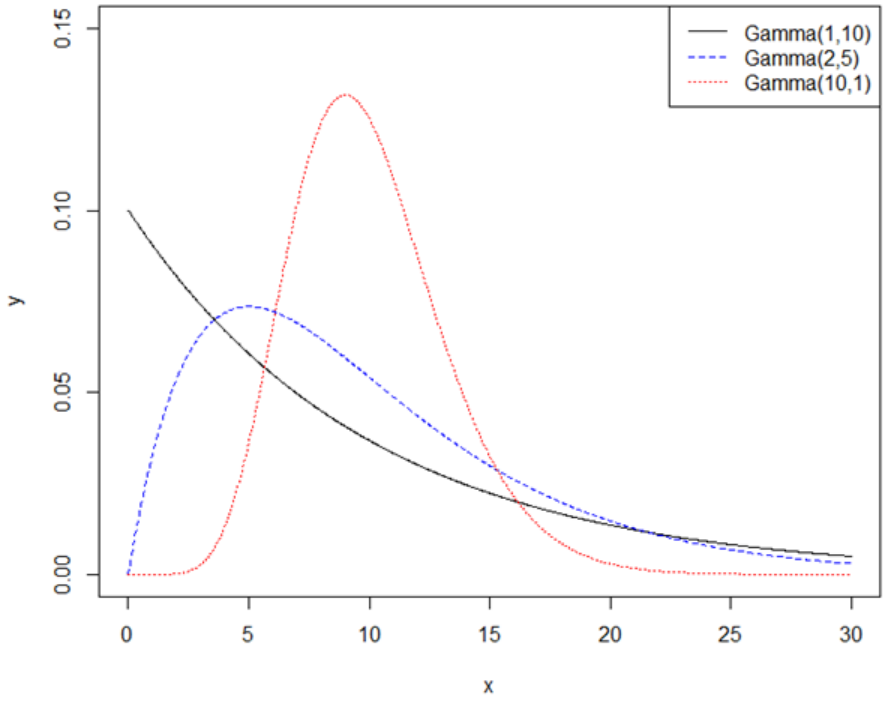

Figure 2.1: Gamma density

2.2 Maximum Likelihood Estimation

Maximum likelihood estimation (MLE) is a commonly used method for estimating parameters of a probability model from a sample. Suppose that we have a random sample $x = x_1, ..., x_n$, that are independent and identically distributed from a common probability density function $f(x; \theta)$, where $\theta = (\theta_1, \theta_2, ..., \theta_k)$ is a vector of unknown parameters. The joint density function of the random sample is thus:

$$f(x) = f(x_1, x_2, ..., x_n) = \prod_{i=1}^{n} f(x_i; \theta), \tag{2.2}$$

Once the data has been observed, the only unknown values of Equation (2.2) are the parameters θ. The joint density function, when viewed only as a function of θ, is referred to as the likelihood function, $L(\theta; x)$. To estimate the parameters for our given data set x, the general approach is to choose values of θ that maximize the likelihood function. It is also typical to maximize the log likelihood $l(\theta) = log(L(\theta))$ as the calculus is usually easier to implement, and log is an order-preserving transformation.

$$l(\theta) = logL(\theta) = \sum_{i=1}^{n} logf(x_i; \theta). \tag{2.3}$$

In general, if $l(\theta)$ is differentiable, the parameter estimates $\hat{\theta}$ are obtained by solving the system of equations:

$$\frac{\partial l}{\partial \theta_i} = 0, i = 1, ..., k. \tag{2.4}$$

For some cases, closed form solutions will exists. In situations where there is no closed form solution, numerical optimization routines such as Newton-Raphson are utilized. It should also be noted that $\hat{\theta}$ is a function of the observed values $x_1, x_2, ..., x_n$ and therefore, a random variable.

2.2.1 MLE of the Gamma Distribution.

Consider a random sample of $x_1, ..., x_n$ that has been collected for a random variable X from the gamma distribution defined by equation 2.1. Here, θ is a two parameter vector, $\theta = (\alpha, \beta)$. The log likelihood function, using Equation 2.3, for the sample is

$$\begin{aligned} l(\theta) &= -n\alpha Ln(\beta) - nLn(\Gamma(\alpha)) \\ &\quad + (\alpha - 1)\sum_{i=1}^{n} Ln(x_i) - \frac{1}{\beta}\sum_{i=1}^{n} x_i \end{aligned} \tag{2.5}$$

Taking the first derivative with respect to both α and β of equation 2.5, yields the system of equations:

$$\begin{aligned} -nLn(\beta) + \sum_{i=1}^{n} Ln(x_i) - n\Psi'(\alpha)) = 0 \\ -\frac{n\alpha}{\beta} + \frac{1}{\beta^2}\sum_{i=1}^{n}(x_i) = 0 \end{aligned} \tag{2.6}$$

where $\Psi'(\alpha) = \frac{d}{d\alpha}Ln(\Gamma(\alpha))$. There is no closed form solution for the system but can easily be solved using the R package **mle** which, by default, utilizes the routine developed by [7].

Fisher's information allows one to derive the asymptotic variance covariance matrix for maximum likelihood estimators. We will first state the definition of Fisher's information and then derive it for the Gamma model. Let X be a random variable with probability density function $f(x;\theta)$. Fisher's information matrix for a sample of size n is denoted as $I(\theta)$ where the $(i,j)^{th}$ entry, I_{θ_i,θ_j} is defined as:

$$I_{i,j} = -nE\left[\frac{\partial^2 log f(x;\theta)}{\partial\theta_i \partial\theta_j}\right]. \tag{2.7}$$

Since there are two parameters for the gamma distribution, Fisher's information is a 2×2 matrix which we will denote $I(\alpha, \beta)$. Using Equation (2.3), we have

$$I(\alpha,\beta) = -E\begin{bmatrix} \frac{\partial^2 l}{\partial \alpha^2} & \frac{\partial^2 l}{\partial \alpha \partial \beta} \\ \frac{\partial^2 l}{\partial \beta \partial \alpha} & \frac{\partial^2 l}{\partial \beta^2} \end{bmatrix} = \begin{bmatrix} I_{(\alpha,\alpha)} & I_{(\alpha,\beta)} \\ I_{(\alpha,\beta)} & I_{(\beta,\beta)} \end{bmatrix}. \tag{2.8}$$

The elements of $I(\alpha,\beta)$ are derived as follows:

$$I_{\alpha,\alpha} = -E(\frac{\partial l}{\partial \alpha^2}) = n\Psi''(\alpha),$$

$$I_{(\beta,\beta)} = -E(\frac{\partial l}{\partial \beta^2}) = \frac{-n}{\beta^2} + \frac{2\alpha\beta\sum_{i=1}^{n} x_i}{\beta^3} = \frac{n\alpha}{\beta^2},$$

and

$$I_{(\alpha,\beta)} = -E(\frac{\partial l}{\partial \alpha \partial \beta}) = \frac{n}{\beta}.$$

where $\Psi'(\alpha) = \frac{d}{d\alpha} Ln(\Gamma(\alpha))$ and $\Psi''(\alpha) = \frac{d^2}{d\alpha^2} Ln(\Gamma(\alpha))$.

Thus, Fisher's information for the gamma distribution with parameters α and β is:

$$I(\alpha,\beta) = \begin{bmatrix} n\Psi''(\alpha) & \frac{n}{\beta} \\ \frac{n}{\beta} & \frac{n}{\beta^2} \end{bmatrix} \tag{2.9}$$

It will be beneficial to reparameterize Fisher's information using $\mu = \alpha\beta$ and β. Following the notation of [10] Fisher's information under the new paramaterization, denoted $K(\mu,\beta)$, is expressed as

$$K(\mu,\beta) = J' I(\alpha = \frac{\mu}{\beta}, \beta) J \tag{2.10}$$

where J is the Jacobian matrix from the transformation of (α,β) to (μ,β). The elements for the Jacobian matrix are defined as

$$J = \begin{bmatrix} \frac{1}{\beta} & \frac{-\mu}{\beta} \\ 0 & 1 \end{bmatrix} \tag{2.11}$$

Substituting Equation (2.11) into Equation (2.10) yields the final result:

$$K(\mu,\beta) = \begin{bmatrix} \frac{1}{\beta} & \frac{-\mu}{\beta} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} n\Psi''(\frac{\mu}{\beta}) & \frac{n}{\beta} \\ \frac{n}{\beta} & \frac{n}{\beta^2} \end{bmatrix} \begin{bmatrix} \frac{1}{\beta} & 0 \\ \frac{-\mu}{\beta} & 1 \end{bmatrix}, \tag{2.12}$$

This can then be simplified to

$$K(\mu,\beta) = \begin{bmatrix} \frac{n\beta^2\Psi''(\frac{\mu}{\beta})-2n\mu\beta+n\mu^2}{\beta^4} & \frac{n\beta-n\mu}{\beta^3} \\ \frac{n\beta-n\mu}{\beta^3} & \frac{n}{\beta^2} \end{bmatrix} \tag{2.13}$$

$$K^{-1}(\mu,\beta) = \begin{bmatrix} \frac{-\beta^2}{n-n\Psi''(\frac{\mu}{\beta})} & \frac{\beta^2-\beta\mu}{n-n\Psi''(\frac{\mu}{\beta})} \\ \frac{\beta^2-\beta\mu}{n-n\Psi''(\frac{\mu}{\beta})} & \frac{-\mu^2+2\beta\mu-\beta^2\Psi''(\frac{\mu}{\beta})}{n-n\Psi''(\frac{\mu}{\beta})} \end{bmatrix} \tag{2.14}$$

Due to the general asymptotic properties of MLEs, the joint distribution of $(\hat{\mu},\hat{\beta})$ follows a multivariate Normal distribution with mean vector (μ,β) and covariance matrix K^{-1}. Thus, the standard error of the MLE for the mean is the square root of the first row and column entry of K^{-1} and can be used for constructing a confidence interval for μ:

$$\hat{\mu} \pm z_{\alpha/2}\sqrt{\frac{\beta^2}{n\Psi''(\frac{\mu}{\beta})-n}}. \tag{2.15}$$

2.2.2 Estimating a population proportion

Consider a random sample of $y_1,...,y_n$ that has been collected from a Bernoulli distribution with mass function $f(y) = p^y(1-p)^{1-y}$ for $y = 0,1$. It is easily shown that the estimator $\hat{p} = \frac{\Sigma_{i=1}^n y_i}{n}$ maximizes the log likelihood and Fishers information is $I(p) = \frac{n}{p(1-p)}$. Similar to $\hat{\mu}$ previously, a confidence interval for p can easily be obtained and yields the widely known Wald interval:

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}. \tag{2.16}$$

2.3 Interval Estimation for the Mean of the Gamma Hurdle Model

In this section we will derive and discuss four approaches to estimating the population mean of the Gamma Hurdle Model. We will consider three standard approaches while introducing a fourth new approach that, to our knowledge, has not been investigated before. The three standard approaches are the classic student t-interval, the percentile bootstrap, and the Bootstrap t-interval. All three methods are widely used and applicable in numerous settings. While we will consider an approach that assumes the hurdle model, it will be informative to compare it to commonly used procedures that have been known to behave robustly in numerous settings. Our approach will be based on maximum likelihood estimation similar to the previous discussions when working with the gamma and Bernoulli models.

2.3.1 t-Interval

Due to the Central Limit Theorem, we know that this type of interval is robust to its own assumption of normality and can be used to estimate the population mean from any scenario given the sample size is adequate enough. For the following development, let X be a random variable and $x_1, x_2, ..., x_n$ be a random sample. let $\bar{X}$ be the sample mean and s^2 be the sample variance. When n is large, we have that

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n}).$$

This can be used to standardized the sample mean to give

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1).$$

If it is assumed that $X \sim N(\mu, \sigma^2)$, replacing σ with s in the ratio yields the t-statistic which follows a t-distribution with $n - 1$ degrees of freedom.

$$\frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t_{n-1}.$$

A $(1-\alpha)100\%$ confidence interval for μ can be constructed by writing out a probability statement for the t-statistic and algebraically manipulating the statement to get an interval estimate which contains the true mean $(1-\alpha)100\%$ of the time.

$$\bar{x} \pm t_{\alpha/2}\frac{s}{\sqrt{n}} \tag{2.17}$$

2.3.2 Bootstrap Intervals

The percentile and bootstrap t-intervals provide an approach to working with statistics in which their sampling distributions are unknown. Without this knowledge, construction of an interval would not be possible. The bootstrap procedure can also be used on well known statistics, such as the sample mean, in cases where the t-intervals robustness properties do not apply. While there are many versions of bootstrapping statistics, we will briefly discuss two bootstrap intervals that are relatively easy to implement.

Let $x_1, x_2, ..., x_n$ be a random sample for a random variable X obtained from a distribution that depends on a parameter θ. Let $\hat{\theta}$ be a statistic that is used to estimate θ. Any bootstrap interval procedure first begins by creating B bootstrap samples. These samples are obtained by sampling, with replacement, n observations from the original random sample. For each one of these samples, we can then obtain the statistic of interest $\hat{\theta}_j$ where $j = 1, 2, ..., B$. This resampling procedure allows for one to obtain an estimate of the sampling distribution of the statistic $\hat{\theta}$ essentially for free without taking any additional random samples. This is why it was given the name "bootstrap" as we are picking our self up by our own bootstraps. With the bootstrap statistics $\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_B$, a confidence interval can be constructed using different strategies.

Percentile Bootstrap

The percentile Bootstrap is a very intuitive, natural approach to producing a CI for the parameter θ. Consider ordering the bootstrapped statistics denoted as $\hat{\theta}_{(1)}, \hat{\theta}_{(2)}, ..., \hat{\theta}_{(B)}$. Denote $\hat{\theta}_{(\alpha/2)}$ and $\hat{\theta}_{(1-\alpha/2)}$ as the $\alpha/2$ and $1-\alpha/2$ percentiles from the bootstrap sampling distribution which can be easily obtained from the ordered bootstrapped statistics. These values serve as the lower and upper limits for a $(1-\alpha)100\%$ confidence interval for θ. Implementing the percentile bootstrap is straightforward for the gamma hurdle model. One simply treats the statistic of interest as the sample mean, $\hat{\theta} = \bar{X}$, and an interval is produced using the percentiles of the bootstrapped sample means.

There are some short comings to using a percentile bootstrap intervals. For some situations, the mean of the bootstrap distribution is biased away from the original sample's observed value. Additionally, the skeweness observed in a bootstrapped sampling distribution is often ill represented in the tails. Both of these issues can result in poor coverage. Additionally, the percentile bootstrap does not perform well in the presence of nuisance parameters which is common for many problems. An example of this is the Normal distribution in which we may only be interested in the population mean but we must estimate the population variance as well even though we do not care about inference on that parameter. There have been numerous modifications proposed to alleviate these shortcomings of the percentile bootstrap. One of these extensions is the bootstrap t-interval.

Bootstrap t-interval

The bootstrap t-interval builds on the approach of the traditional t-statistic discussed in the previous section. Recall that the traditional t-statistic is a ratio of the random variable $\bar{X}$ and its estimated standard error $SE(\bar{X}) = \frac{s}{\sqrt{n}}$.

$$t = \frac{\bar{X} - \mu}{SE(\bar{X})}$$

When samples are taken from a normal distribution, we know that the t-statistic follows a t-distribution, and thus the theoretical percentiles of the t-distribution are used to create the upper and lower limits. When we are not dealing with normally distributed data or the sample mean as a statistic, we could still construct a t-statistic but we would no longer know the distribution of this t-statistic. The bootrap t-interval suggest that we create a bootstrap distribution of t-statistics to approximate what the sampling distribution is. From this distribution, we can estimate the $\alpha/2$ and $1 - \alpha/2$ percentiles and use these to construct a "t-like" interval.

Formally, let $\hat{\theta}$ be the sample statistic and let $SE(\hat{\theta})$ be the estimated standard error obtained on a sample $x_1, x_2, ..., x_n$. Note that the standard error computation can be obtained in a multitude of ways depending on how $\hat{\theta}$ is defined. For example, if we are dealing with the sample mean statistic or an MLE, we have derived formulas for the standard errors that can be used. If the standard error formula is not known then the bootstrap samples of the statistic can be used to obtain a standard error estimate as all one needs to do is calculate the standard deviation of the $\hat{\theta}_j$'s.

Using the bootstrap statistics, one can create the sampling distribution of the analogous t-statistic by calculating, for $j = 1, 2, ..., B$,

$$T_j = \frac{\hat{\theta}_j - \hat{\theta}}{SE(\hat{\theta}_j)} \tag{2.18}$$

The simplicity of the expression above should not be taken for granted. As stated previously, the standard error of the statistic can be computed in different ways depending on the statistic of interest. When computing $SE(\hat{\theta}_j)$, we must obtain the standard error of the statistic for the j^{th} bootstrap sample. So in situations where there is no theoretical form of the standard error, an additional bootstrap procedure on the bootstrap sample, must be implemented. This highlights the fact that the

bootstrap t-interval in some cases is more computationally expensive to run.

Letting $T_{(\alpha/2)}$ and $T_{(1-\alpha/2)}$ be the percentiles of the T_j's, a confidence interval is obtained by obtaining the upper (U) and lower (L) limits as follows:

$$\begin{aligned} L &= \hat{\theta} - T_{(\alpha/2)} SE(\hat{\theta}) \\ U &= \hat{\theta} + T_{(1-\alpha/2)} SE(\hat{\theta}) \end{aligned} \tag{2.19}$$

For the gamma hurdle model, we will be performing the bootstrap t-interval where $\hat{\theta} = \bar{X}$ and thus an estimate of the standard error is easily obtained for each bootstrap sample.

2.3.3 A Wald type interval for $(1-\pi)\mu$

Under the gamma hurdle model, estimation of the proportion of zeros, π, can be obtained via maximum likelihood as described in Section 2.2 treating the observed value of zero as $Y = 1$ in the Bernoulli model. Similarly, for the nonzero values within the sample, a maximum likelihood estimate for the mean can be obtained. Denoting these two MLES as $\hat{\pi}$ and $\hat{\mu}$ respectively, the MLE for mean of the hurdle model is simply $(1-\hat{\pi})\hat{\mu}$. The purpose of this section is to derive an approximate standard error for the MLE, then a Wald interval can be constructed utilizing the large sample properties of the MLE.

To derive an appropriate standard error, we will utilize the multivariate delta method [4]. For our case we will simply introduce this when working with two variables. Suppose that X_1 and X_2 converge in distribution to a multivariate normal distribution with mean vector (μ_1, μ_2) and covariance matrix Σ. Denote the entries of Σ as σ_{11}, the variance of X_1. The variance of X_2 is σ_{22} and the covariance between X_1 and X_2 is $\sigma_{12} = \sigma_{21}$. Let $g(x_1, x_2)$ be a scalar function for which we would like to derive a distribution for the new random variable $Y = g(X_1, X_2)$.

Letting Δ be the gradient of g, $\Delta = (\frac{\partial g}{\partial x_1}, \frac{\partial g}{\partial x_2})$, the delta method states that

$Y = g(X_1, X_2)$ converges in distribution to a Normal distribution,

$$g(X_1, X_2) \xrightarrow{D} N\Big(g(\mu_1, \mu_2), \Delta\Sigma\Delta^T\Big) \tag{2.20}$$

For the Gamma hurdle model, the joint distribution of $\hat{\pi}$ and $\hat{\mu}$ converges to a multivariate normal distribution with mean vector (π, μ) and covariance matrix Σ. The elements of Σ are obtained from Sections 2.1 and 2.2.1 yielding $\sigma_{11} = \frac{\pi(1-\pi)}{n}$ and $\sigma_{22} = \frac{\beta^2}{n\Psi''(\frac{\mu}{\beta})-n}$. The off diagonal element, σ_{12}, is assumed to be zero, since the proportion of zeros and the mean of the nonzero values are modeled independently of each other.

With the above information, we wish to derive an asymptotic results for $Y = (1-\hat{\pi})\hat{\mu}$. Note in this case, $g(x_1, x_2) = (1-x_1)x_2$ and the gradient is $\Delta = (-x_2, 1-x_1)$ and thus $\Delta\Sigma\Delta^T = x_2^2\sigma_{11} + (1-x_1)^2\sigma_{22}$. Using the delta method result in Equation 2.20, we have the following asymptotic result:

$$(1-\hat{\pi})\hat{\mu} \xrightarrow{D} N\Big((1-\pi)\mu, \mu^2\sigma_{11} + (1-\pi)^2\sigma_{22}\Big). \tag{2.21}$$

A $(1-\alpha)100\%$ confidence interval for the mean of the gamma hurdle model can be obtained using Slutsky's theorem, and replacing the estimates of π and μ in the the variance covariance matrix Σ.

$$(1-\hat{\pi})\hat{\mu} \pm z_{\alpha/2}\sqrt{\hat{\mu}^2\hat{\sigma}_{11} + (1-\hat{\pi})^2\hat{\sigma}_{22}} \tag{2.22}$$

3 Simulation Studies

The simulation studies will be conducted assuming data are observed from the gamma hurdle model previously defined. For a given sample size, a Binomial random variable is drawn with parameter π to determine the number of 0's present in the sample. The remaining nonzero observations will be randomly sampled from a gamma distribution with parameters α and β ($\mu = \alpha\beta$). For a given scenario, 10000 simulations were conducted and an estimate of the coverage and average interval width was obtained for each of the four interval approaches described in Chapter 2. We also recorded the parameter estimates of the MLE's to investigate their point estimate properties such as bias and mean square error.

3.1 Simulation Overview

In terms of simulation scenarios, we varied the total sample size $n = 50, 100, 200, 400$. For each sample size, α and β were set as defined in Table 3.1. The choice of parameters were chosen because we wanted there to be a mean that was close to the inflated zeroes and one that is further out. So, we varied α and β values, keeping the mean fixed at either 10 and 100. Doing so allowed us to investigate situations where the gamma distribution part of the model is strictly decreasing ($\alpha = 1$) while others behaved far less skewed. We also varied the proportion of zeros, π. The π value ranged from 0.0 to 0.8, by increments of 0.2.

α	β	$\mu = \alpha\beta$
1	10	10
1	100	100
2	5	10
2	50	100
10	10	100

Table 3.1: Selected α, β, and μ scenarios

Pseudo code is provided below on the general flow of the simulation conducted in R. The following is a description of one scenario for a chosen choice of n,π, and $\mu = \alpha\beta$:

- For 10000 iterations
  - Generate data from hurdle model
  - Compute CI for the various techniques
  - Store intervals and parameter estimates
  - Count if interval was correct
- Compute Overall Coverage and Average Widths
- Compute parameter estimate properties (Bias)

3.2 Results

To help clearly show the results of the simulations, graphs were created to examine the coverage performance across the scenarios. For reach of the scenarios, there will be a graph showing the coverage of the four interval estimation techniques discussed in Chapter 2 followed by a graph of the estimated widths of the intervals. Both

graphs are stratified over varying sample sizes. There is also a legend on each graph that distinguishes the different interval methods. The label "boot" is the bootstrap method, "boot2" is the bootstrap-t method, "t" is the t-interval method, and "wald" is the Wald interval method. For each of the graphs, there are 4 panels. For the coverage, each panel has the labeled sample size above the graph and then the coverage was plotted by the π values. For the widths, the layout and legends are the same, except the widths are being plotted by sample size and π values.

Our first scenario sets $\alpha = 1$ and $\beta = 10$. Looking at Figure 3.1, we see that the bootstrap t-interval method generally performed better for the sample size of 50, but only included the desired 95% coverage when $\pi = .2, .4$. When the π value is .80, this means that the proportion of zeroes is much higher. Even though the bootstrap-t does not quite reach the desired coverage, it does perform significantly better than the other intervals. Also, recall Figure 2.1, when $\alpha = 1$ the gamma distribution is an exponential distribution, which is strictly decreasing, and not the case when $\alpha > 1$. So, when the distribution was strictly decreasing and the π value increases, the bootstrap-t seems to perform better, even though it still does not contain the desired coverage level. When the sample size was 100, the bootstrap-t interval still performed better than the others, but only included the desired 95% when $\pi = .2, .4, .6$. When the sample size was 200 and 400, all the methods seemed to perform more similar to each other, but there is still slight variations. When the sample size was 400, all of the methods include the desired 95% coverage.

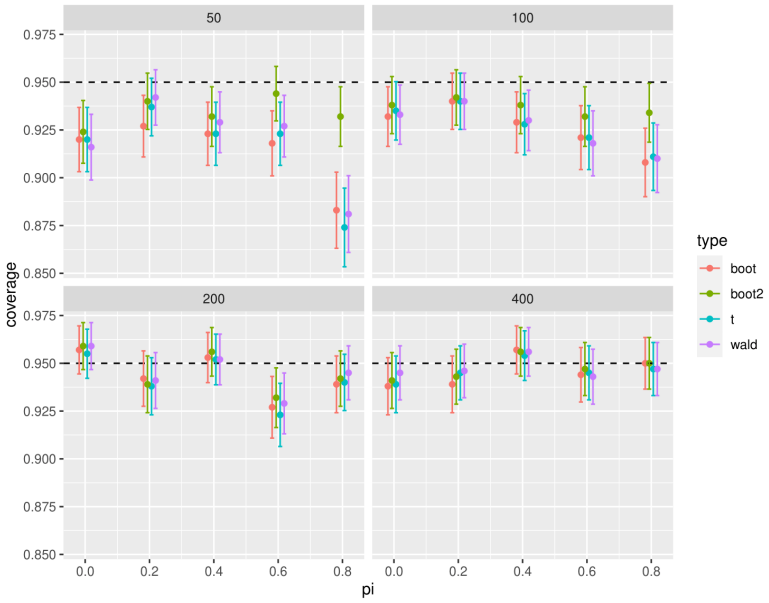

Figure 3.1: Interval coverage for all types across varying sample sizes and π values when $\alpha = 1, \beta = 10$

When looking at Figure 3.2, the widths seem to follow a similar pattern. The widths tend to behave similarly across methods. However, when the sample size was 50, the regular bootstrap method shows that the width was noticeably smaller than the width of the other intervals. The width of the bootstrap t interval is wider than the other intervals as well when the sample size is 50 and corresponds to the only procedure close to a nominal coverage rate. As the π values increases, we can see that the width of the intervals becomes narrower as expected. For the standard methods this makes a great deal of sense. The large number of zeroes, shrink the sample standard deviation in most generated samples. For the Wald interval, seeing a tighter interval was at first counter intuitive. Having a smaller number of nonnegative values to estimate the parameters of the gamma part of the model intuitively would suggest that a wider interval would be reflected. We will discuss some issues with the Wald interval in greater detail in the next chapter.

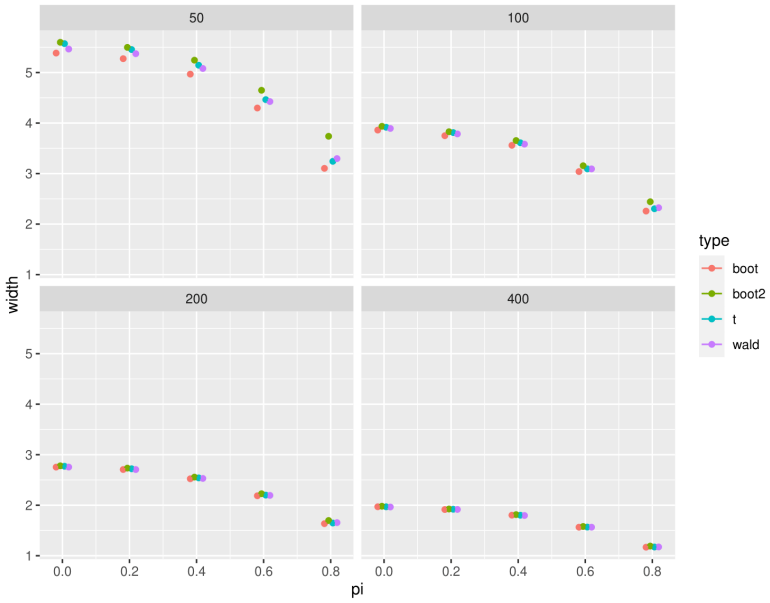

Figure 3.2: Interval widths for all types across varying sample sizes and π values when $\alpha = 1, \beta = 10$

Next, we consider the coverage and widths when $\alpha = 1$ and $\beta = 100$. When looking at Figure 3.3, we can see that the methods perform in a similar fashion when the sample size was 50 as they did when β was 10. The coverage seems to under perform for all intervals as π gets closer to 1. The bootstrap-t method still performs better than the other methods for the small sample size, but the difference is that now all the intervals include the desired 95% coverage when the sample size was 50 and $\pi = 0.0, 0.2, 0.4$, and the bootstrap-t includes the desired coverage when $\pi = .8$ as well. When the sample size was 100, all of the methods either include the desired 95% or almost do for $\pi = 0.0, 0.2, 0.4$, and 0.6. When π was 0.8, only the bootstrap-t has the desired coverage. For the sample size of 200 and 400, the methods perform better than they did when β was 10 because the 200 sample size shows all methods except the t interval includes the desired coverage for all π values. When π is 0.8, the t interval shows slight under performance. When the sample size was 400, all

intervals include the desired coverage at all π values.

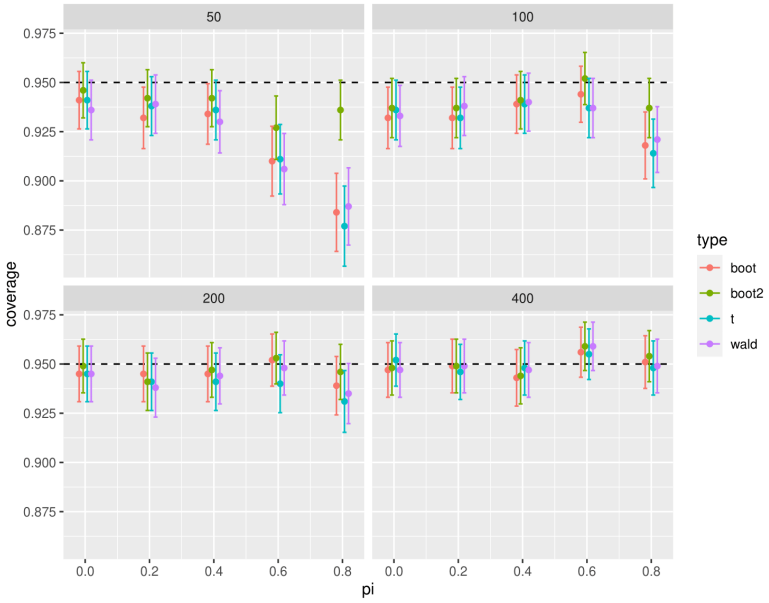

Figure 3.3: Interval coverage for all types across varying sample sizes and π values when $\alpha = 1, \beta = 100$

When examining the widths of when $\alpha = 1$ and $\beta = 100$ in Figure 3.4, we see that the widths tend to perform similarly as the previous scenario. The sample size of 50 shows that the regular bootstrap method is slightly tighter than the widths of the other intervals. The other sample sizes show that all of the methods seem to have similar widths, and there is the slight pattern when the widths become tighter as the π value approaches 1. We also see that as the sample size gets larger, the widths for all the intervals gets smaller.

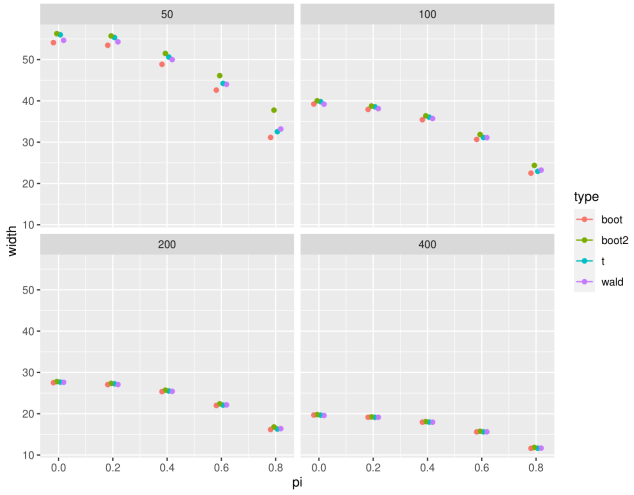

Figure 3.4: Interval widths for all types across varying sample sizes and π values when $\alpha = 1, \beta = 100$

Next consider when $\alpha = 2$ and $\beta = 5$. Upon examination of Figure 3.5, we can see that there is an obvious under performance from the Wald interval at all π values and all sample sizes. When the sample size is 100, 200, and 400, the other three types of intervals perform well since we can see that at all π values the intervals include the desired 95% coverage. When the sample size is 50, the only interval that appears to include the desired coverage at all π values is the bootstrap-t method. As the π value grows, the performance of the other 3 intervals appears to be worse. The exception to this would be when π is 0.6, the t-interval performs just as well as the bootstrap-t methods, but then the t-interval under performs for the next two π values.

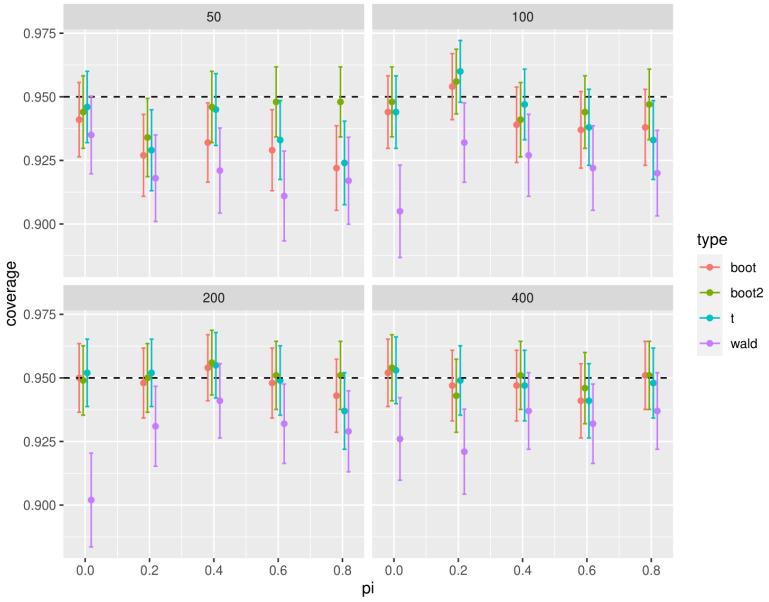

Figure 3.5: Interval coverage for all types across varying sample sizes and π values when $\alpha = 2, \beta = 5$

The widths in Figure 3.6 show that the Wald interval seems to have slightly tighter widths for all sample sizes. These tighter widths would explain the under coverage of the interval. One interesting note, the interval widths appear to be a quadratic function of π. In fact, upon further examination of all of the previous scenarios, a

quadratic relationship seems plausible. This intuitively makes sense as the margin of error for the Wald interval contains the standard error of the sample proportion of zeros. How dominant the quadratic relationship exists appears to depend on the choice of parameters for the gamma distribution as we will see in an upcoming scenario.

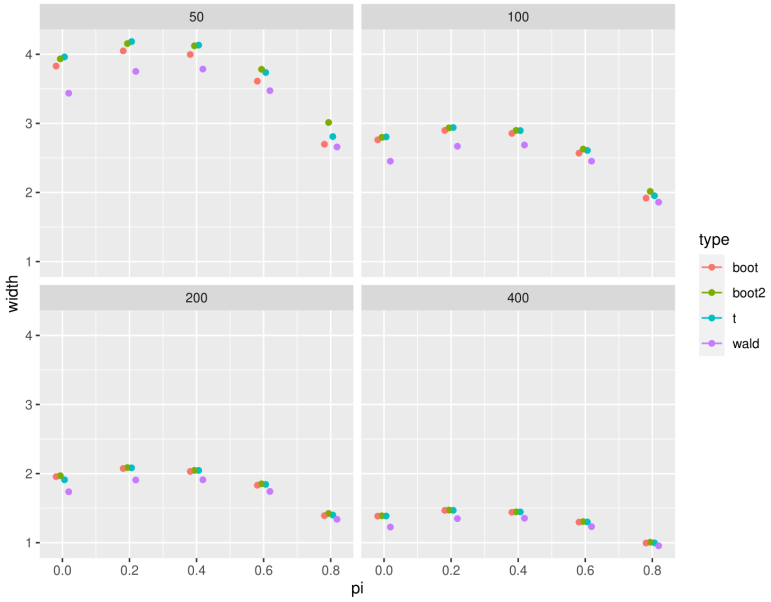

Figure 3.6: Interval widths for all types across varying sample sizes and π values when $\alpha = 2, \beta = 5$

When looking at Figure 3.7 where $\alpha = 2$ and $\beta = 50$, we can see that there is still the general under performance by the Wald interval. There is a general trend for the Wald interval where the interval gets better from $\pi = 0.0$ to 0.4, and then as π increases to 0.6 and 0.8 the Wald interval becomes further from the desired 95% coverage. For sample size of 200 and 400 the other 3 intervals contain the desired coverage at all π values. One notable difference between 3.5 and 3.7 is that for the $\beta = 50$, when the sample size was 50 we see that the bootstrap-t method does not contain the desired coverage in the interval when π is 0.6.

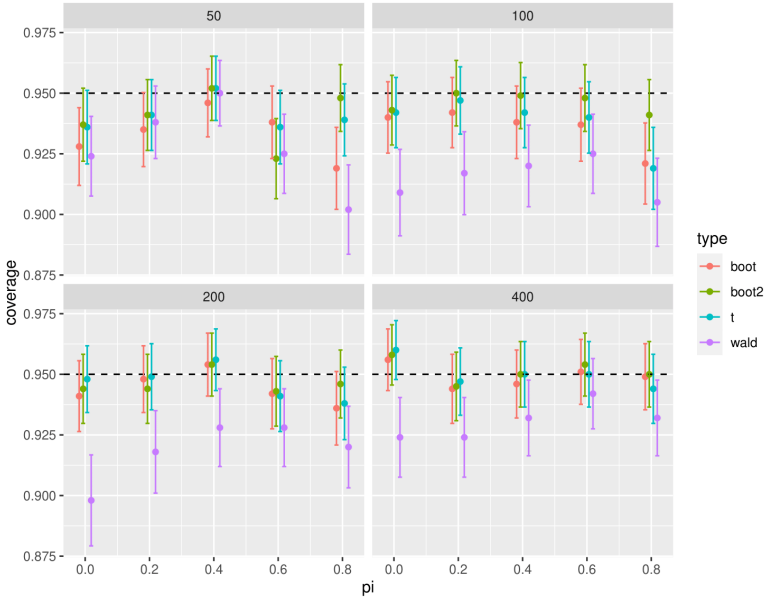

Figure 3.7: Interval coverage for all types across varying sample sizes and π values when $\alpha = 2, \beta = 50$

The widths in figure 3.8 show what would be expected based on the results from Figure 3.7. We can see that the Wald interval seems to have a more narrow width, which can explain the reason why the interval does not contain the desired coverage. We also see that as the sample size gets larger, the widths for all the intervals gets smaller. The quadratic relationship for the widths as a function of π becomes more

pronounced as well.

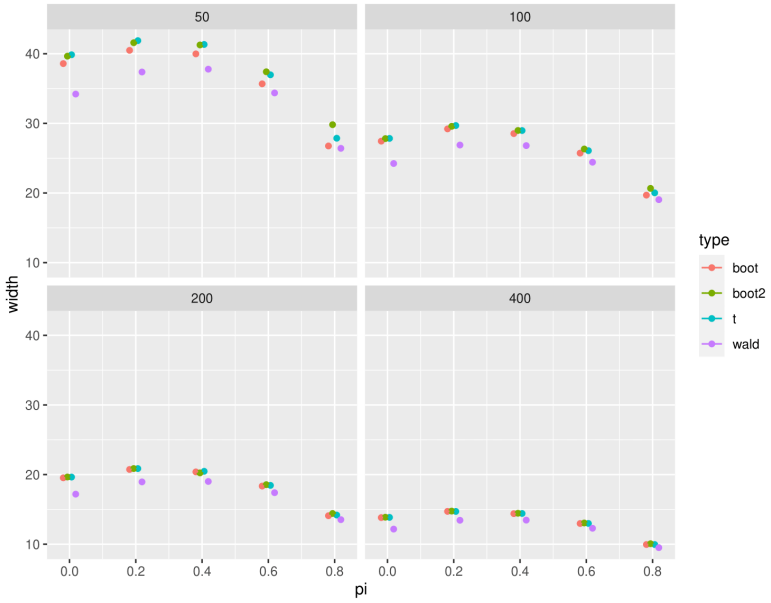

Figure 3.8: Interval width for all types across varying sample sizes and π values when $\alpha = 2, \beta = 50$

The last scenario considered was α and β were 10, which creates a highly variable gamma distribution which is much less skewed than the other gamma models considered. Here there is a clear separation from the nonzero data and the inflated zeroes when visualizing simulated data with histograms. In Figure 3.9 we see that in general, it seems that all the interval types include the desired 95% coverage. This is even true for the sample size of 50. For the sample size of 50, we can see that the Wald interval does have an obvious under performance. For all the other sample sizes, the desired coverage is reached for all the intervals.

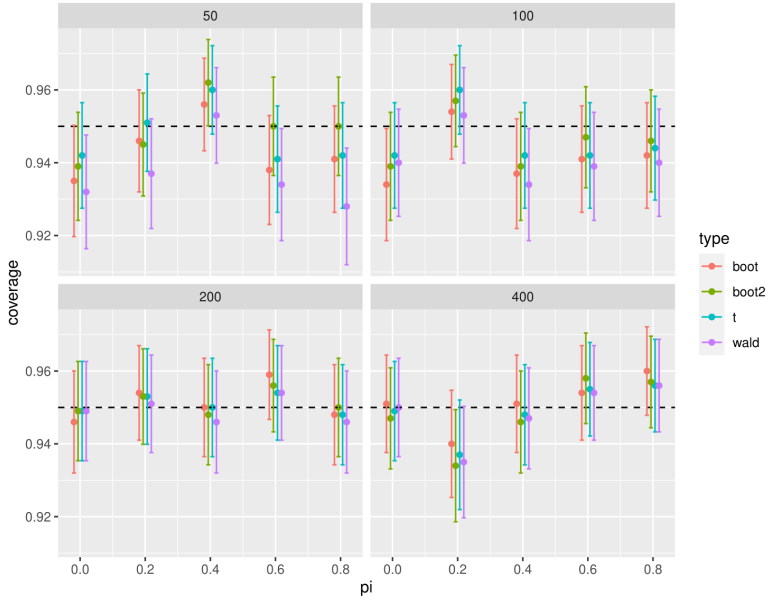

Figure 3.9: Interval coverage for all types across varying sample sizes and π values when $\alpha = 10, \beta = 10$

In Figure 3.10, we can see that there does seem to be a general trend where all the widths for the interval types are about the same, where the most noticeable difference in widths occurs when the sample size is 50. There is still the general decreasing in the width as the sample size increases and the the quadratic relationship is the most pronounced of all the scenarios.

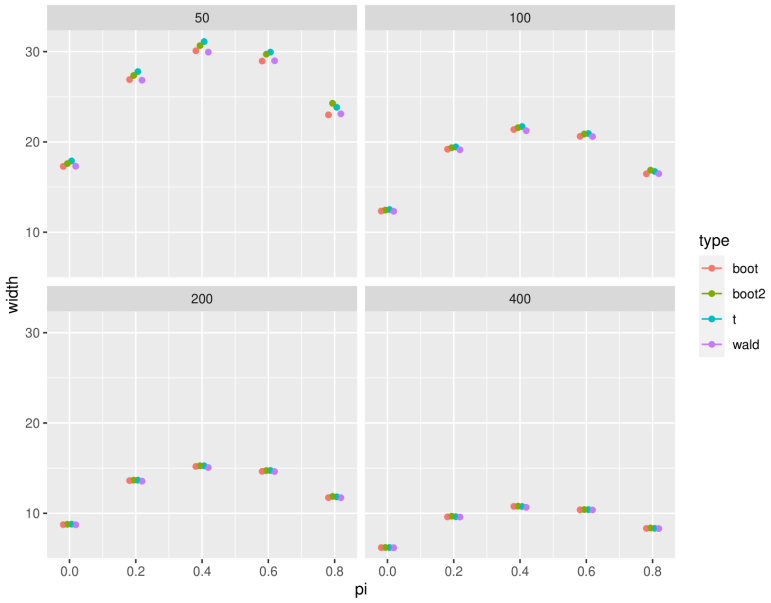

Figure 3.10: Interval widths for all types across varying sample sizes and π values when $\alpha = 10, \beta = 10$

Based on the results that are seen throughout the graphs, it seems that the interval that generally performs the best when working with a Gamma Hurdle model is the bootstrap-t interval. The Wald interval that was created using the idea of MLEs and Fisher's information is not recommended without further refinement.

4 Final Remarks and Future Work

From our simulation studies, the general conclusion is that the bootstrap t-interval procedure performs at the pre-specified coverage more consistently over the other methods we considered. This is most noticeable when the proportion of zeroes in the hurdle model are high. Our newly constructed Wald interval, along with the t-interval and percentile bootstrap under-covered in numerous situations where π is large. Additionally, in scenarios where the gamma is more symmetrically shaped, the Wald interval under-covered even when the sample sizes were quite large. The t-interval and bootstrap percentile interval for these more symmetric cases performed poorly as well when dealing with large proportion of zeroes, however they benefited more from the larger sample sizes than the Wald interval.

Perhaps a more troubling concern is that when the proportion of zeroes was set to 0, the coverage for the Wald interval performed more poorly than when zeroes were present in the data. The following discussion attempts to provide some insight as to why this was indeed the case along with some theories as to why the bootstrap t-interval did a much better job overall.

Table 4.1 below provides simulated biased estimates of the MLE for the gamma parameters μ and β when $n = 50$ under the hurdle model. The simulated biases are provided under different values of π as well. It should be noted that the only impact that π plays on estimation of μ and β is the fact that fewer and fewer observations are used in the estimation. So for example, when $\pi = 0.8$, on average there will only be $(1 - 0.8)50 = 10$ observed values that are generated from the gamma. Upon examination of the table, we can see that the bias in estimation of the β parameters is consistently negative and as π gets closer to 0, the bias is much less pronounced.

This intuitively makes sense due to the asymptotic unbiased properties of maximum likelihood estimators. There are more nonzero observations used in estimation when π is low. The bias of the mean estimate is much less severe and within simulation error of the true value.

Recall Equation 2.15. When looking just at the standard error in this confidence interval, when $\hat{\beta}$ is too small, this causes everything under the radical to decrease. Consider hypothetical values where $\hat{\beta} = 10, \hat{\mu} = 10$, and n = 400. The $\hat{\beta}$ value in the numerator is being squared, so this would make the numerator 100. Notice that if you have a slightly smaller $\hat{\beta}$ value of say 9.9 being squared, the numerator would become 98.01, causing the whole SE to decrease. One potential option for future work is to provide a biased corrected estimate using the work of Barndorff-Nielsen and Cox [1]

$E(\hat{\beta}) - \beta$	$E(\hat{\mu}) - \mu$	π
-0.5022	0.0396	0.8
-0.2331	0.005269	0.6
-0.1463	-0.02912	0.4
-0.1292	-.002218	0.2
-0.1264	-.009338	0.0

Table 4.1: Estimate bias of MLEs with $n = 50$, $\alpha = 2$, $\beta = 5$ ($\mu = 10$)

While we do not provide the results here, we should note that the sampling distributions of $1 - \hat{\pi}$ and $\hat{\mu}$ look to be approximately normal as expected. However, the product of two normally distributed random variables is not normally distributed except for some specific cases [2] It is possible that poor performance observed for the Wald interval is also due to skeweness within the sampling distribution of $(1 - \hat{\pi})\hat{\mu}$. Future work could potentially investigate these concerns as well.

The difference in confidence interval coverage between the two bootstrap procedures is an interesting one. It provides some evidence for additional considerations to

improve upon the hurdle model. The bootstrap t-interval was introduced to address some of the shortcomings of the bootstrap percentile interval. One of these situations is that the percentile bootstrap does not perform well when estimating a parameter in the presence of nuisance parameters. In our case we are solely interested in the global mean, $(1-\pi)\mu$, but the additional parameter β must still be estimated. Classic statistical inference teaches us that an optimal test or interval should be based on conditioning on statistics used to estimate these nuisance parameters. The Wald interval does not consider this approach. The t-interval does this but its derivation hinges on the fact the sample mean and sample variances are independent of each other. This is most certainly not the case when dealing with the gamma distribution. The bootstrap t-interval does allow a form of conditioning since the standard error is computed on each bootstrap sample.

To illustrate this point, we performed an additional simulation study to examine the relationship between the sample mean and standard deviation when zeroes are inflating the samples. In addition to examining the Gamma hurdle model, we also performed simulations under a truncated Normal hurdle model, where $\mu = 10$ and $\sigma = 5$. The truncation was conducted at 0 to create a nonnegative Random variable. The advantage of using the truncated Normal model in this case is that when we select the mean and standard deviation of the truncated normal density to be far away from the truncation at 0 along with $\pi = 0$, we are simply looking at the Normal probability model with no inflated zeroes. Under this special case, the sample mean and standard deviation will be independent of each other. This would not be the case with the gamma hurdle model.

Figure 4.1 provides scatterplots of 10,000 randomly generated sample means of size, $n = 400$, from the truncated normal hurdle model using truncated normal parameters $\mu = 10$ and $\sigma = 5$ with π set to 0 and 0.8 respectively. When $\pi = 0$, we can see that there is essentially no (very mild) correlation between the sample mean and

standard deviation. Under this chosen setting of parameters, the truncated normal is skewed enough to create a small dependency between the two statistics. On the contrary, when we investigate the behavior at $\pi = 0.8$, the relationship between the standard deviation and mean is strongly linearly related with a Pearson's correlation coefficient of 0.9098.

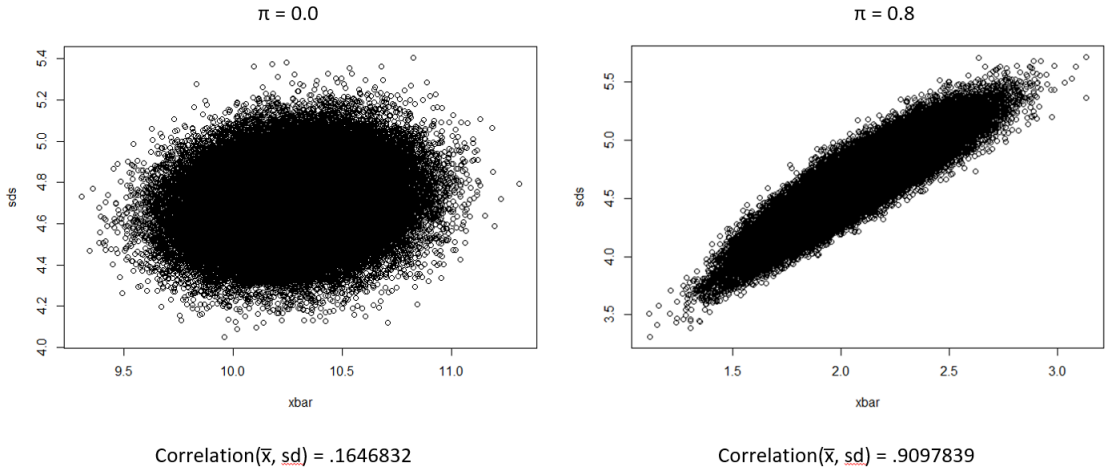

Figure 4.1: This is the truncated normal mean vs standard deviation. The mean is 10 and the sample size is 400. The left graph has $\pi = 0.0$ and the right graph has $\pi = 0.8$.

Figure 4.2 provides an additional scenario holding the parameters in the previous example fixed except for the mean parameter which was increased to 100. The truncation is so small in this case, the hurdle model we are producing is essentially a traditional normal distribution with inflated zeroes. When $\pi = 0$, the relationship between the sample mean and variance is uncorrelated again as expected. When $\pi = 0.8$, the dependency between the mean and variance is even stronger than in the first scenario.

In summary, our current investigations suggest that a Bootstrap t-interval using the mean statistic performs the best in terms of nominal coverage of the overall population mean of the Gamma hurdle model. It is the most robust across various

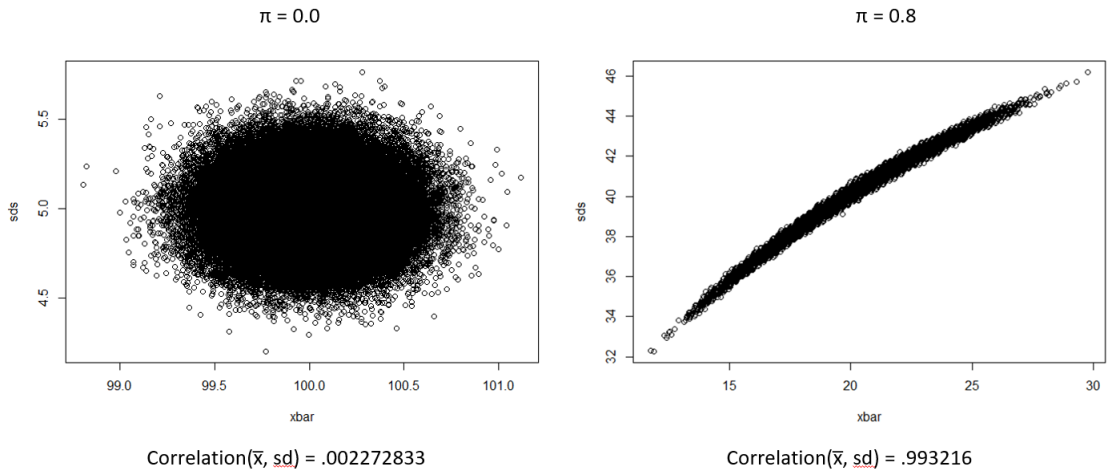

Figure 4.2: This is the truncated normal mean vs standard deviation. The mean is 100 and the sample size is 400. The left graph has $\pi = 0.0$ and the right graph has $\pi = 0.8$.

properties including distributional properties of the gamma, for high rates of zero inflation, and with moderate sample sizes of 50 and 100. The Wald interval, as we have currently derived it, is not recommended. The dependencies between the sample mean and variance are captured in the bootstrap t-interval since the bootstrapped t-statistics use the sample variance of each bootstrap sample in addition to the estimated mean. This insight leads to some additional suggestions for future work.

The first approach would be to use the bootstrap t-interval approach on the MLE of the gamma hurdle model. This should help with conditioning on the nuisance parameter β. An interesting question would be to determine if this approach would also solve the issue of biased estimates of the β parameter. If not, additional work should be considered to create a less biased estimator of β when working with the Wald interval directly. A second approach is to potentially study the dependency of the sample mean and standard deviation both analytically across a wide range of hurdle models to develop a more general approach to condition on nuisance parameters in

the hurdle model family of problems. A third approach would be to investigate a likelihood ratio type statistic and create a confidence interval through p-value inversion [10]. This approach would allow for conditioning of any nuisance parameters and uses large sample theory that is typically more accurate in finite samples relative to the normal approximation of Wald intervals. Lastly, a Bayesian approach could be used to create a credible set for the overall population mean. Additional investigations on appropriate prior structure could then be conducted.

BIBLIOGRAPHY

[1] Ole E. Barndorff-Nielsen and D. R. Cox, *Asymptotic techniques for use in statistics*, Chapman and Hall, London, 1989.

[2] C. Craig, *On the frequency function of xy*, The Annals of Mathematical Statistics **7** (1936), 1–15.

[3] Cindy Xin Feng, *A comparison of zero-inflated and hurdle models for modeling zero-inflated count data*, Journal of Statistical Distribution and Application **8** (2021).

[4] Roger L. Berger George Casella, *Statistical inference*, Brooks/cole, Cengage Learning, 2002.

[5] et al. Hu, Mei-Chen, *Zero-inflated and hurdle models of count data with extra zeros: Examples from an hiv-risk reduction intervention trial*, The American Journal of Drug and Alcohol Abuse **37** (2011), no. 5, 365–375.

[6] L. Liu, M. E. Cowen, R.L Strawderman, and Y.C.T Shih, *A flexible two-part random effects model for correlated medical costs*, Journal of Health Economics **29** (2010), 110–123.

[7] J.A. Nedler and R. Mead, *A simplex algorithm for function minimization*, Computer Journal **7** (1965), 308–313.

[8] Allen T. Craig Robert V. Hogg, Joseph W. McKean, *Introduction to mathematical statistics*, vol. 7, Pearson Education, Inc., 2013.

[9] C.E. Rose, S.W. Martin, K. A. Wannemuehler, and B.D. Plikaytis, *On the use of zero-inflated and hurdle models for modeling vaccine adverse event count data*, Journal of Bio-pharmaceutical Statistics **16** (2006), 463–481.

[10] Smith R.L. Young, G.A., *Essentials of statistical inference*, Cambridge University Press, 2005.

VITA

Alissa Jacobs received her Bachelor's degree in Mathematics from Texas Tech University in 2020. After graduating from SFA, Alissa will be taking on a data analytic position for a youth soccer club in the Dallas area.

Permanent Address: 216 Mitchell St, Apt 3
Nacogdoches, TX 75962

The style manual used in this thesis is A Manual For Authors of Mathematical Papers published by the American Mathematical Society.

This thesis was prepared by Alissa Jacobs using LaTeX.