1-1-2005

Confidence interval estimation for a geometric distribution

Majgan Beria
University of Nevada, Las Vegas

CONFIDENCE INTERVAL ESTIMATION

FOR A GEOMETRIC DISTRIBUTION

by

Majgan Beria

Bachelor of Science
University of Nevada, Las Vegas
2004

A thesis submitted in partial fulfillment
of the requirements for the

Master of Science Degree in Mathematical Sciences
Department of Mathematical Sciences
College of Sciences

Graduate College
University of Nevada, Las Vegas
May 2006

ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

UNLV
UNIVERSITY OF NEVADA LAS VEGAS

Thesis Approval

The Graduate College
University of Nevada, Las Vegas

April 13, 2006

The Thesis prepared by

Majgan Beria

Entitled

Confidence Interval Estimation for a Geometric Distribution

is approved in partial fulfillment of the requirements for the degree of

Master of Science in Mathematical Sciences

Ashok K. Singh
Examination Committee Chair

Dean of the Graduate College

Examination Committee Member

Examination Committee Member

Graduate College Faculty Representative

ABSTRACT

Confidence Interval Estimation for a Geometric Distribution

by

Majgan Beria

Dr. Ashok K. Singh, Examination Committee Chair
Professor, Department of Mathematical Sciences
University of Nevada, Las Vegas

A geometric random variable models the number of trials required to obtain the first success in a Bernoulli process. This distribution has been used by Merrill(2005) as a probability model for the distribution of drivers yielding to pedestrians in a traffic microsimulation investigation. The sample proportion of yielding drivers was calculated using the method of moments, and the bootstrap method was used for computing a confidence interval (CI) for the success probability. The properties of this CI for the geometric distribution, however, have not been investigated. The main objective of this thesis is to develop the performance of the bootstrap method, and then propose a Bayesian analysis for estimating a confidence interval for the population proportion when the data follow a geometric distribution.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude and appreciation to my advisor Dr. A.K. Singh. Without his guidance, this project would never have been completed so well. Being a part of this research group has been an exciting challenge and privilege.

I also want to give my gratitude to Dr. Dennis Murphy for his valuable advice on my project and for reviewing the thesis. He continuously provided me great ideas to improve both my program and my programming ability.

I also thank Dr. Rohan Dalpatadu and Dr. Tony Lucas to agree on serving my supervisory committee.

I cannot express in words how deeply I feel about the support and confidence my family and my friends have shown in me over the years. Without them, I would never have the opportunity to accomplish my education.

Finally, I am thankful to this university for giving me this chance to get into the field of mathematical science. All the staff and faculty always try their best to assist me whenever I needed them to. They also worked as hard as possible to make our learning environment better everyday.

CHAPTER 1

INTRODUCTION

The geometric distribution is a discrete probability distribution, existing only on the nonnegative integers. It models the number of failures before one success in an independent succession of Bernoulli trials where each trial results in success or failure. Merrill (2005) used this distribution as a probability model in his investigation, a practical traffic microsimulation of mid-block pedestrian crossing between two signalized signalized intersections. The model was calibrated based on field observations and findings from previous pedestrians and vehicular research. The proportion of drivers yielding to pedestrians was an essential parameter to estimate and enter into this model. The distribution of yielding drivers was represented as a geometric frequency distribution of vehicles that yields to pedestrians waiting to cross, and the proportion was estimated from the frequency of those individual occurrences. The properties of the geometric distribution will be discussed in depth in Chapter 2.

Since the true proportion in the population is an unknown parameter, Merrill tried to estimate it by constructing a confidence interval procedure. However, it has been suggested that there is no method for finding an exact confidence interval for the parameter p of a geometric distribution. So, Merrill computed an approximate confidence interval for this proportion in the population using the bootstrap method. The objective of this thesis is to investigate the performance of the above procedure and to present another interval estimator for a geometric distribution. The bootstrap method will be explained in more detail in Chapter 3. Several practical examples of this technique are discussed as an extension of the results presented in Merrill (2005). This technique will be presented by computing the classical upper and lower confidence intervals when new data sets follow a geometric distribution.

Moreover, Bayesian approaches are introduced as an alternative to the bootstrap. The roots of Bayesian philosophy are reviewed and the difference between the Bayesian interpretation of results from the classical approach is stressed. We will analyze this in more detail in Chapter 3.

The main goal of this thesis is to compare the performance of interval estimators of p by the Bootstrap method and Bayesian approach, using Monte Carlo simulation. The results of the experiment will be presented in Chapter 4. Finally, Chapter 5 summarizes the overall strategy, methods and results of the thesis.

CHAPTER 2

GEOMETRIC MODEL

Probability distributions are used to model randomness in populations; as such, statisticians usually deal with a family of distributions rather than a single distribution. There are two major types of probability distributions: discrete and continuous. A real random variable X is a function from a sample space into the real numbers, with the property that for every potential outcome there is an associated probability P[X=x] which exists for all real values of x in the sample space. A random variable X is said to have a discrete distribution if the range of X, the sample space, is countable. In most situations, the random variable has integer-valued outcomes. The second major type of distribution has a continuous random variable. In this situation, the sample space is some interval of the real line and the function used to model random behavior over the interval is called a probability density function (pdf).

The purpose of this chapter is to introduce a particular type of discrete distribution, the geometric distribution, and its relation to other common discrete distributions. For each distribution, we will give its mean and variance and some other useful statistical descriptive measures and interrelationships that may aid understanding.

2.1 Specification of geometric distribution

The geometric distribution is based on the idea of Bernoulli trial. A Bernoulli trial (named for James Bernoulli, one of the founding fathers of probability theory) is a random experiment with exactly two possible outcomes. A random variable X has a Bernoulli (p) distribution if

$$X = \begin{cases} 1 & \text{with probability p} \\ 0 & \text{with probability 1 - p} \end{cases} \quad \text{, where } \ 0 \le p \le 1.$$

The value X = 1 is often termed a "success" and p is referred to as the success probability. The value X = 0 is termed a "failure".

Now let X count the number of failures in a Bernoulli sequence before the first success. This is the waiting time to the first success. The event {X = x} is the event of x consecutive failures followed by a success. For a Bernoulli sequence, with probability p of success on any independent and identical trial, the event of x -1 failures followed by a success has probability

$$P(X = x) = p\ (1-p)^{x-1}, \qquad 0 \le p \le 1, \quad x = 1,2,...$$

Such a random variable X has the geometric distribution with parameter (p). Note that some authors (e.g., Beyer 1987) consider a slightly different variable, Y, defined as the number of failures that occur before the first success. Thus, Y = X – 1, and

$$P(Y = y) = p\ (1-p)^{y}\ , \qquad 0 \le p \le 1, \quad y = 0,1,2,..$$

2.2 Expectation and Variance

The mean or expected value of X, where X~GEO(p) is given by

$$E(X)=\begin{cases}\sum\limits_{x=0}^{\infty} xp(1-p)^{x} & \text{if } x=0,1,\ldots\\ \sum\limits_{x=1}^{\infty} xp(1-p)^{x} & \text{if } x=1,2,\ldots\end{cases}$$

Taking the derivative of the geometric series, we obtain

$$E(X)=\begin{cases}\frac{1}{p}-1 & \text{if } x=0,1,\ldots\\ \frac{1}{p} & \text{if } x=1,2,..\end{cases}$$

A similar calculation will show that the variance of X is

$$\text{Var}(X)=\frac{1-p}{p^{2}}$$

2.3 Properties of geometric distribution

2.3.1 Memorylessness Property

The geometric distribution has an interesting property, known as the "memoryless" property; that is,

$$P\left(X\geq n+x\,|\,X\geq n\right)=P(X\geq x)$$

This is an expression of the independence of successive events in a Bernoulli sequence. If no success has been observed by trial n, then the (conditional) probability of waiting at least x more trials is the same as the probability of waiting for x or more trials at the beginning of the sequence. The sequence essentially starts over at each trial.

2.3.2 Relation to other distribution

The geometric distribution is related to the family of binomial distributions. The binomial distribution is one of the more useful discrete distributions, also based on the idea of a Bernoulli trial. If n independent Bernoulli trials with probability of success p on each trial are performed, define the random variables $X_1,\ldots,X_n$ by

$$X_i = \begin{cases} 1 & \text{with probability p} \\ 0 & \text{with probability 1 - p .} \end{cases}$$

The random variable $Y = \sum_{i=1}^{n} X_i$ has the binomial (n, p) distribution. The probability that a random variable *X* with binomial distribution B(n, p) is equal to the value k, where $k = 0, 1, \ldots., n,$ is given by

$$P(X=k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad \text{where } \binom{n}{k} = \frac{n!}{k!(n-k)!}$$

with mean $E(X) = np$ and variance $Var(X) = np(1-p)$.

Suppose that, instead of counting the number of successes in a fixed number of Bernoulli trials, which generates the binomial distribution, we count the number of Bernoulli trials required to get a fixed number of successes. This latter formulation leads to the negative binomial distribution.The negative binomial distribution is used when the number of successes is fixed and we are interested in the number of failures before reaching the fixed number of successes. A random variable *X* which follows a negative binomial distribution is denoted $X \sim NB(r,p)$. Its probabilities are computed with the formula

$$P(X=x)=\binom{x+r-1}{r-1}p^r(1-p)^x \qquad \text{for } 0\le p\le 1,\quad x=0,1,2,....$$

The geometric distribution is a special case of the negative binomial distribution, where r = 1.

2.4 Applications of geometric distribution

Many applications can be modeled by a geometric distribution; for example, runs of one species with respect to another in transects through plant populations (Pielou, 1962), a ticket control problem (Jagers, 1973), a surveillance system for congenital malformations (Chen, 1978), and estimation of animal abundance (Seber, 1982).The distribution is used in Markov chain models; for example, in meteorological models of weather cycles and precipitation amounts, developed in 1962 by Gabriel and Neumann. Many other applications in queueing theory and applied stochastic models are discussed in Taylor and Karlin (1984) and Bhat (1984). Daniels (1962) has investigated the representation of a class of discrete distributions as a mixture of geometric distributions and has applied this to busy-period distributions in equilibrium queueing systems. There are many other applications of the geometric distribution, but we are particularly interested in Merrill's (2005) research. Merrill used this distribution in developing a computer simulation model for a mid-block pedestrian crossing between two signalized intersections, and calibrated this model based on field observations and findings from previous pedestrian and vehicular research. An essential parameter to estimate for this model was the proportion of drivers yielding to pedestrians. Yielding drivers would yield for a pedestrian, except when the vehicle was further from the pedestrians than the stopping sight distance. Non-yielding drivers, however, would not yield unless the pedestrian was blocking the travel lane. Distinguishing between drivers that would and would not yield could only occur when a pedestrian was present. Since the condition no longer existed once a vehicle stopped, the vehicles upstream of the first yielding driver could not be distinguished. Since the pedestrian and vehicular arrivals are somewhat random, this situation fits the description of a geometric distribution. The sample data for the proportion of yielding drivers were collected during the midday peak period at a certain location. The distribution of yielding drivers could be represented as a geometric frequency distribution of vehicles that yield to a pedestrian waiting to cross, and the population proportion, which is unknown, could be estimated from the frequency of these individual occurrences.

Since the sample mean is $\bar{x} = \frac{1}{p}$ for a geometric distribution, then $\hat{p} = \frac{1}{\bar{x}}$. To estimate p, the population proportion, method of moments or maximum likelihood could be used to find the sample proportion. For a geometric distribution, both the method of moment estimator (mme) and the maximum likelihood estimator (mle) are equal to $\hat{p} = \frac{1}{\bar{x}}$. The Method of Moments was then used to calculate the sample proportion.

$$\hat{p} = \frac{1}{\bar{x}} = \frac{m}{\sum\left(\hat{f}(x_i) \times x_i\right)},$$

where $\hat{f}(x)$ is the observed frequency for each category x and m is the total number of observations. Calculations show that the estimated proportion of yielding vehicles is $\hat{p} = 0.435$. Table 1 shows the number of vehicles that failed to yield to the pedestrians in the crosswalk, the observed frequency and the estimated probability.

Table 1 Observed frequency and probability distribution for sample yielding vehicle proportion

Number of vehicles failed	Number of vehicles observed	Frequency (observed)	Probability (observed) $= \frac{f(x_i)}{N}$
x – 1	x	f(x)	
0	1	61	0.4919
1	2	26	0.2097
2	3	17	0.1371
3	4	6	0.0484
4	5	5	0.0403
5	6	3	0.0242
6	7	4	0.0323
7	8	0	0.000
8	9	0	0.000
9	10	0	0.000
10	11	0	0.000
11	12	1	0.008
12	13	0	0.000
13	14	1	0.008
m = 124		124	1.000

A chi-square goodness-of-fit test was used to verify that the sample proportion was plausibly represented by a geometric distribution. With 95% confidence, the interval (0.348, 0.522) contains the true proportion p. It was found that the standard error $SE(\hat{p})=0.0445$, and the sample proportion was within ± 0.087 of the proportion in the population. One of the methods developed for computing CIs for this geometric model was the bootstrap simulation technique. This method will be discussed in depth in the next Chapter and then will be compared with a Bayesian interval estimation approach.

CHAPTER 3

INTERVAL ESTIMATION

3.1 Classical Interval Estimation

When sampling from a population described by a probability mass function f(x|p), knowledge of p implies that f(x|p) is strictly a function of x. Hence, it is a natural to seek a means of finding a good estimator of the unknown parameter p.

For the geometric distribution, where a random variable X~GEO(p), we want to estimate the unknown parameter p based on n independent observations of X, that is, $X_1, X_2, \ldots, X_n$. As we mentioned in Chapter 2, both the method of moments estimator and the maximum likelihood estimator are equal to $\hat{p} = \frac{1}{\overline{X}}$, where $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Along with this estimate of the value of the parameter p, we want to have some understanding of how close we can expect our estimate to be to the true value. One approach would be to consider confidence interval estimators of p, which are rules for establishing the lower and upper bounds of an interval that is expected to contain the true value of p. The confidence level associated with an interval estimator is the percentage of the time in repeated sampling that the realized intervals will contain the true value of the unknown parameter p.

To construct a confidence interval procedure for a given parameter θ, we typically need the sampling distribution of a statistic T that has θ as one of its parameters. In large samples, asymptotic normal distributions are used to approximate the sampling distributions, as justified by the central limit theorem. However, it is expected that the uncertainty of the estimate can be large if the sample size is small. Moreover, such methods are not always efficient or applicable if we are generating a confidence interval procedure for an unknown parameter of some discrete distributions. For example, in the geometric distribution, the estimator of the parameter p has a sampling distribution with no closed form. Therefore, a sensible way to construct a confidence interval for the parameter of a geometric distribution is via the bootstrap technique.

3.1.1 Bootstrap Simulation Method

The bootstrap is a type of resampling method applied to observed data (Efron and Tibshirani 1993; Mooney and Duval 1993). It was introduced and popularized by Efron (1979, 1982) and has been discussed in greater detail with many variations by other authors (Davison and Hinkley, 1997; Chernick, 1999). Bootstrap methods are computer-intensive methods of statistical analysis that use repeated resampling of the original data, with replacement, to calculate confidence intervals. This method assures that if the input is a random sample generated from a probability distribution and the resampling process is repeated a large number of times, characteristics of the population will emerge. The bootstrap is a powerful tool for testing or avoiding parametric assumptions when computing confidence intervals and can be applied to almost any problem and any data set.

The steps in the parametric bootstrap simulation experiment used in this thesis are described below:

1. Generate a random sample of a specific sample size (m), from a geometric probability distribution f(x;p), where p represents the true population proportion.

2. Generate a resample $\{x_1^*, x_2^*, ..., x_m^*\}$, with replacement, from the input sample produced in step 1.

3. Compute the sample proportion $\hat{p}$, $\hat{p} = \frac{1}{\bar{X}}$, for a geometric distribution.

4. Repeat steps 2-3 a large number of times (B). This generates B estimates of the population proportion $\left\{\hat{p}_1, \hat{p}_2, ..., \hat{p}_B\right\}$. This process simulates the sampling distribution of $\hat{p}$ from the repeated values of $\hat{p}_i$, i = 1, …, B. Sort the B estimates of the population proportion in ascending order and extract the upper and lower $\alpha/2$ quantiles. This gives us the upper and lower limits of a percentile-based $100(1-\alpha)\%$ confidence interval for p.

3.1.2 Bootstrap Experiment

For the problem at hand, we must know the true value of the parameter so that the performance of bootstrap method can be investigated when data follow a geometric distribution. For this reason, we simulate data from the geometric distribution with known parameter p.

To illustrate the steps of our bootstrap simulation experiment, we will use the dataset generated from the geometric distribution with known parameter p = 0.4.

3.1.2.1 Bootstrap Simulation Step 1

A dataset of size m = 100 is generated from a geometric distribution with p =0.4. Its frequency table x is shown below, where

Table 2 Computer-Generated Data Set

X	Frequency
0	47
1	19
2	15
3	10
4	3
5	4
7	1
10	1
m = 100	

x represents the number of failures prior to the first success. Figure 1 shows the relative frequency bar chart of X, where X~GEO(0.4) for a sample size of 100.

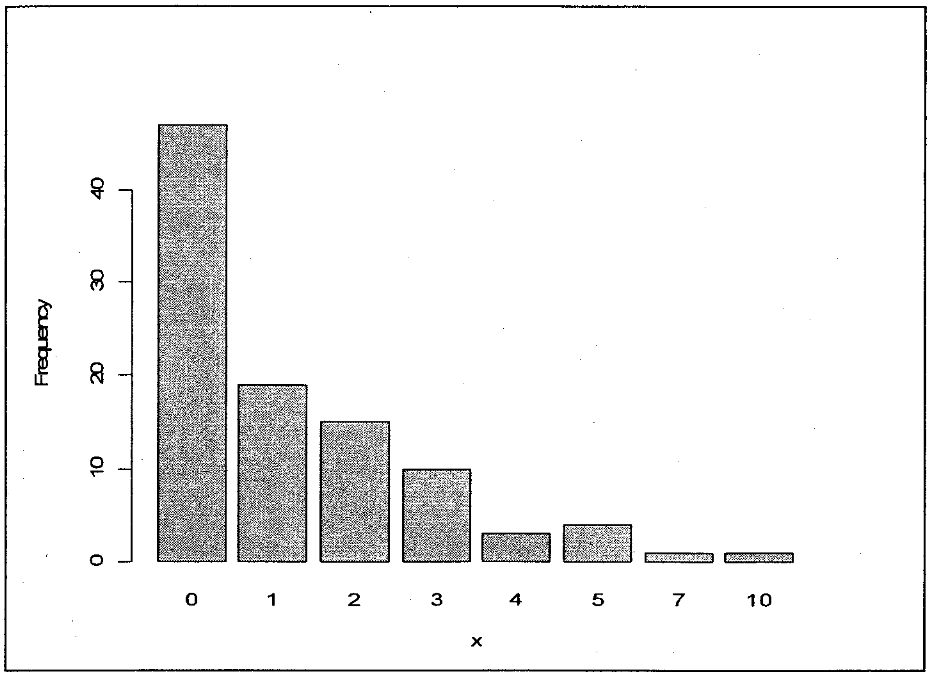

Figure 1 Barplot of Sample Geometric Distribution with p = 0.4

3.1.2.2 Bootstrap Simulation Step 2

From the computer-generated data set shown in Table 3, a bootstrap resample of X of the same size is created; its frequency table is shown below:

Table 3 Sample 1 Data Set

Boot Sample 1	
x	Frequency
0	49
1	16
2	15
3	10
4	3
5	5
6	0
10	2
	N = 100

3.1.2.3 Bootstrap Simulation Step 3

The sample proportion $\hat{p}$, an estimate of the population proportion, is computed by taking the inverse of $(1+\overline{X})$, where $\overline{X}$ is the mean of the bootstrap resample:.

$$\hat{p}_1 = \frac{1}{1+\overline{X}} = \frac{1}{1+1.33} = 0.429$$

3.1.2.4 Bootstrap Simulation Step 4

Repeat Sections 3.2.2.2 through 3.2.2.3 1000 times. This generates 1000 estimates of the population proportion $\left\{\hat{p}_1,...,\hat{p}_{1000}\right\}$. Compute the mean of those 1000 estimates to get an estimate of the true the population proportion. Figure 2 shows that the sampling distribution of the simulated values $\hat{p}_1,...,\hat{p}_{1000}$ is approximately normally distributed, as predicted by central limit theorem (CLT).

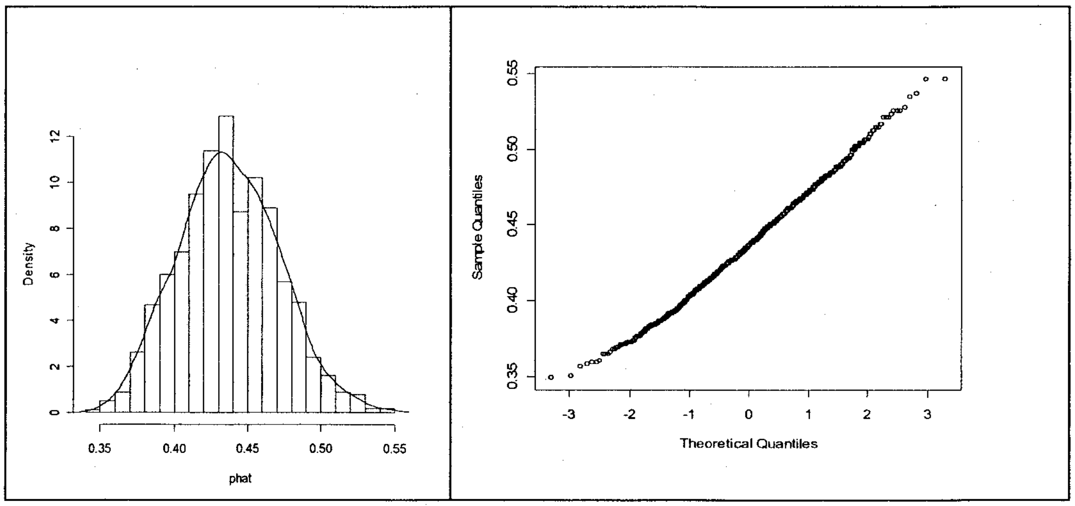

Figure 2 Histogram and Q-Q plot of $\hat{p}_i$ values based on 1000 simulations

After sorting the 1000 proportions in ascending order, the 2.5 percentile and the 97.5 percentile comprise the endpoints of a 95% bootstrap CI of the true population proportion p of the geometric distribution.

We then have $\hat{p} = 0.4372651$ as the point estimate of p, and 95% CI = (0.3734659, 0.5061224), with standard deviation SD = 0.03402014 and standard error SE = 0.003402014. We can see that this CI contains the true proportion (p = 0.4).

The bootstrap method is a powerful tool for computing CIs of the parameter of a geometric distribution, but is not the only technique available. Another interval procedure that needs to be investigated for this distribution is Bayesian in nature, and is the topic of next section.

3.2 Bayesian Estimation

3.2 1 Bayesian Method

The probability model for the distribution of drivers yielding to pedestrians, as we saw earlier when we looked at the proportion of the Geometric model, has an unknown parameter p. The classical statistical approach considers this parameter as a fixed, but unknown constant to be estimated using the sample data taken randomly from the population of interest. A confidence interval for an unknown parameter is really a frequency statement about the likelihood that numbers calculated from a sample capture the true parameter value in repeated sampling. So the classical statistical approach cannot say there is a 95% probability that the true proportion is in any single interval, because it is either already in, or it is not. This is because under the classical approach, the true proportion is a fixed unknown constant, so it does not have a distribution; however, the sample proportion $\hat{p}$ does. Thus, we can say that there is a 95% chance the random interval contains p, in repeated samples of size m from the same population.

The Bayesian approach, on the other hand, treats the population model parameter as random instead of fixed. Actually, the data are treated as fixed realizations of a random process, accounted for by the likelihood function. Before looking at the current data, we use past information to construct a prior distribution model for the parameter. The prior distribution is chosen to reflect one's prior knowledge of p, which may vary from one person to the next. As a result, the mathematical form of a prior distribution is quite flexible. In particular, conjugate priors are a natural and popular choice of Bayesian prior distribution model, due to their mathematical convenience. The prior distribution of a parameter may be noninformative or informative. Noninformative priors are locally uniform in a certain range of parameter values. The range of possible values may be fixed or may be infinite. An informative prior distribution specifies a particular nonuniform shape for the distribution of the parameter. When new data are gathered, they are used to update the prior distribution. We then take the weighted average of the prior and data, expressed through the likelihood function, to derive what is called the posterior distribution model for the population model parameter. Point estimates, along with interval estimates (known as credibility intervals), are calculated directly from the posterior distribution. Credibility intervals are legitimate probability statements about the unknown parameter, since the parameter now is considered random. Under the Bayesian point of view, we can say that there is a 95% probability that the interval contains the population proportion.

The posterior distribution model is based on Bayes' theorem, which expresses the conditional probability of an event A, given that the event B has occurred, in terms of unconditional probabilities and the probability the event B has occurred, given that A has occurred. It is defined as

$$P(A \mid B) = \frac{P(A,B)}{P(B)} = \frac{P(A) \times P(B \mid A)}{P(B)}$$

In terms of probability density functions, the theorem takes the form

$$g(p \mid x) = \frac{f(x \mid p)g(p)}{\int_0^1 f(x \mid p)g(p)dp}$$

This is known as the posterior density of x, where f(x|p) is the likelihood function of the observed data x given the unknown parameter p, g(p) is the prior density of p and the denominator represents the marginal density of x. When $g(p \mid x)$ and g(p) both belong to the same family of distribution, g(p) and f(x|p) are called conjugate distributions and g(p) is called the conjugate prior for f(x|p). For example, the Beta distribution model is a conjugate prior for the proportion of successes p when samples have a binomial distribution. Since the geometric distribution is a special case of the negative binomial distribution, we will use the Beta as our conjugate prior distribution for the proportion p.

With probability $1-\alpha$ level, a Bayesian credibility interval for p is given by (p_L, p_U), where p_L and p_U satisfy

$$\int_{p_L}^{p_U} g(p \mid x)dp = 1-\alpha.$$

This yields an interval estimate with probability $1-\alpha$.

3.2.2 Bayes Credibe Set For Geometric Distribution

Suppose $x_1, x_2, \ldots, x_m$ are independent random variables from the same Geometric distribution with parameter p. That is, $x_i \sim$ GEO (p) for $i \in \{1,2,\ldots,m\}$ .The likelihood function for the observed data X, given the unknown parameter p, is proportional to

$$\begin{aligned} f(x_1, x_2, \ldots, x_m \mid p) &= \prod_{i=1}^{m}\left[p(1-p)^{x_i}\right] \\ &= p^m (1-p)^{\sum_{i=1}^{m} x_i} \end{aligned}$$

For convenience, let $\sum_{i=1}^{m} x_i = S$; then,

$$f(x_1, x_2, \ldots, x_m \mid p) \propto p^m (1-p)^S.$$

We now consider posterior distributions with this likelihood using different selected prior distributions.

3.2.2.1 The Posterior Of Geometric Distribution With Uniform Prior

A reasonable prior distribution for p must be bounded between zero and one. One option is the uniform prior, p ~ Unif (0, 1), which yields equally likely values of p. The density function for this prior is then,

$$g(p) = \begin{cases} 1 & \text{where } 0 < p < 1, \\ 0 & \text{elsewhere.} \end{cases}$$

Using Bayes' formula, the posterior density function is given by

$$g(p \mid x_1, x_2, \ldots, x_m) = \frac{f(x_1, x_2, \ldots, x_m \mid p) g(p)}{\int_0^1 f(x_1, x_2, \ldots, x_m \mid p) g(p) dp};$$

that is,

$$g(p \mid x_1, x_2, \ldots, x_m) = \frac{p^m (1-p)^S \times 1}{\int_0^1 \left(p^m (1-p)^S \times 1 \right) dp}.$$

Notice that $\int_0^1 \left(p^m (1-p)^S \times 1 \right) dp$ is the normalizing constant, and $p^m (1-p)^S$ is the kernel of the beta distribution. The probability density function (pdf) for the beta distribution, which is known to be proper, is

$$\text{Beta}(x \mid \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1},$$

where $0 < x < 1$, $\alpha, \beta > 0$ and $\Gamma(k) = (k-1)!$ is the Gamma function when $k \geq 1$ is integer-valued.

If we let $x = p$, $\alpha = m+1$, $\beta = S+1$, we get the following expression:

$$g(p\,|\,x_1,x_2,\ldots,x_m)=\frac{p^m(1-p)^S}{\left(\frac{\Gamma(m+1)\Gamma(S+1)}{\Gamma(m+S+2)}\right)\times\int_0^1\frac{\Gamma(m+S+2)}{\Gamma(m+1)\Gamma(S+1)}\times p^m(1-p)^S\,dp}$$

Since $\int_0^1\frac{\Gamma(m+S+2)}{\Gamma(m+1)\Gamma(S+1)}\times p^m(1-p)^S\,dp$ is the integral of the beta pdf over the parameter space for p, this expression equals one. Thus, after simplification we have

$$g(p\,|\,x_1,x_2,\ldots,x_m)=\frac{\Gamma(m+S+2)}{\Gamma(m+1)\Gamma(S+1)}\times p^n(1-p)^S,$$

which is a Beta $(m+1,S+1)$ density.

It worked out that the posterior distribution is a form of a beta distribution. The Bayes estimator of the proportion in the population p, under squared error loss, is just the posterior mean. If $Y\sim\text{Beta}(\alpha,\beta)$, then the mean of a beta distribution is $E[Y]=\frac{\alpha}{\alpha+\beta}$.

Therefore, the Bayes estimator of p is

$$\hat{p}_{\text{Bayes}}=E[p\,|\,X]=\frac{(m+1)}{(m+S+2)}.$$

The credibility interval for the parameter p is then computed from that posterior beta distribution with parameters m+ 1, S + 1.

This uniform prior is just one of an infinite number of possible prior distributions. What other prior distribution could we use?

3.2.2.2 The Posterior of Geometric Distribution with Informative Prior

A reasonable alternative to the Unif(0,1) distribution is the beta prior distribution.

For a random variable p, where $p \sim \text{Beta}(\alpha,\beta)$, the pdf is

$$\text{Beta}(p \mid \alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}$$

In fact, we can show that the Beta (1, 1) distribution is the Unif(0,1) distribution; that is

$$\text{Beta}(1,1) = \frac{\Gamma(2)}{\Gamma(1)\Gamma(1)} p^{1-1}(1-p)^{1-1} \qquad 0 < p < 1,$$

which gives us

$$\text{Beta}(1,1) = \frac{\Gamma(2)}{\Gamma(1)\Gamma(1)} p^{0}(1-p)^{0} = 1 \qquad 0 < p < 1$$

This is the density function for the Uniform (0,1) distribution. Hence

Beta (1,1) ~ Unif(0,1) (or any Uniform(θ, θ + 1)) for any parameter p.

Figure 3 shows the probability density of 6 different Beta distributions. These six different shapes of prior densities correspond to different degrees of belief about the probability that different values of p will be observed. Since p represents the probability of success for a geometric distribution, we need to choose a prior that is bounded between 0 and 1. The plots suggest that the choice of beta distribution as prior is reasonable, in particular Beta (1, 1) and Beta (5, 5). Beta (1, 1) gives equally likely values of p; and the distribution of Beta (5, 5) is symmetric with respect to p = 0.5.

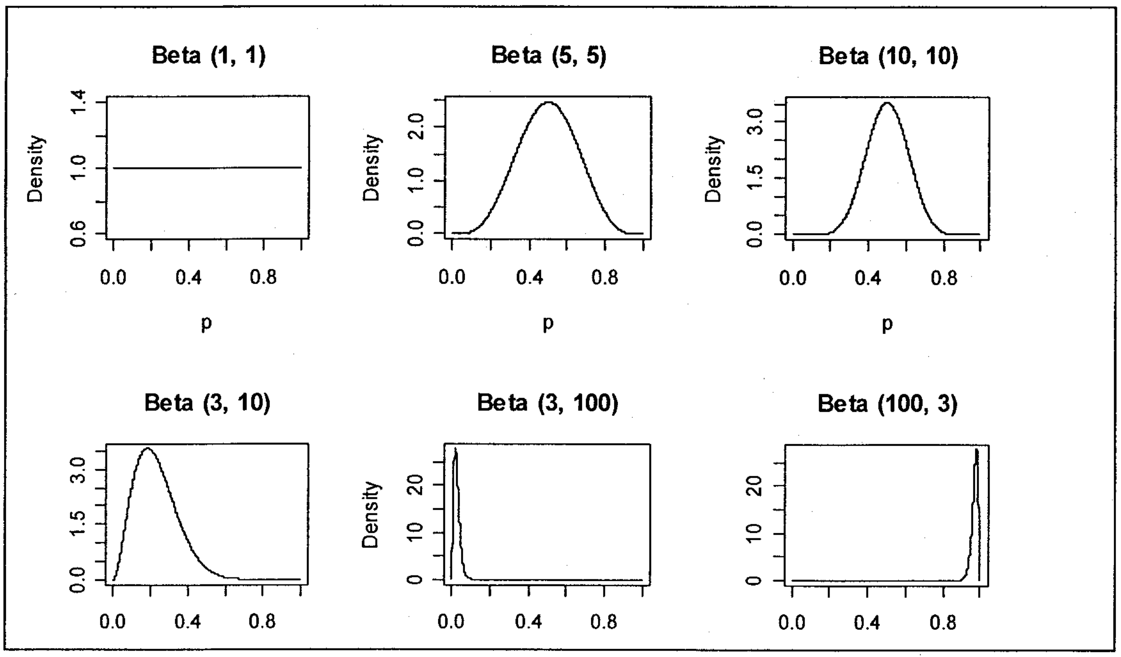

Figure 3 Prior Distributions: Plot of 6 Different Beta Distributions

Let $X = (x_1, x_2, \ldots, x_m) \sim \text{GEO}(p)$, as we defined in the previous section, and let

$p \sim \text{Beta}(\alpha, \beta)$; that is, the proportion p has beta prior distribution. Then,

$$g(p) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1} (1-p)^{\beta-1},$$

where $0 < p < 1$, $\alpha > 0$, $\beta > 0$ and α, β are known arbitrary constants. The posterior density is given by

$$g(p \mid x_1, x_2, \ldots, x_m) = \frac{f(x_1, x_2, \ldots, x_m \mid p) g(p)}{\int_0^1 f(x_1, x_2, \ldots, x_m \mid p) g(p) dp}$$

where $f(x_1, x_2, \ldots, x_m \mid p) = p^m (1-p)^S$.

We then have

$$g(p\,|\,x_1,x_2,\ldots,x_m)=\frac{p^m(1-p)^S\times\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}p^{\alpha-1}(1-p)^{\beta-1}}{\int_0^1\left(p^m(1-p)^S\times\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}p^{\alpha-1}(1-p)^{\beta-1}\right)dp}$$

$$=\frac{\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\times p^{m+\alpha-1}(1-p)^{S+\beta-1}}{\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\times\int_0^1\left(p^{m+\alpha-1}(1-p)^{S+\beta-1}\right)dp}$$

$$=\frac{p^{m+\alpha-1}(1-p)^{S+\beta-1}}{\int_0^1\left(p^{m+\alpha-1}(1-p)^{S+\beta-1}\right)dp}$$

The integral $\int_0^1\left(p^{m+\alpha-1}(1-p)^{S+\beta-1}\right)dp$ is the normalizing constant of a beta distribution. We then have the following expression,

$$g(p\,|\,x_1,x_2,\ldots,x_m)=\frac{p^{m+\alpha-1}(1-p)^{S+\beta-1}}{\frac{\Gamma(m+\alpha)\Gamma(S+\beta)}{\Gamma(m+\alpha+S+\beta)}\int_0^1\frac{\Gamma(m+\alpha+S+\beta)}{\Gamma(m+\alpha)\Gamma(S+\beta)}\times\left(p^{m+\alpha-1}(1-p)^{S+\beta-1}\right)dp}$$

Since $\int_0^1\frac{\Gamma(m+\alpha+S+\beta)}{\Gamma(m+\alpha)\Gamma(S+\beta)}\times\left(p^{m+\alpha-1}(1-p)^{S+\beta-1}\right)dp$ is the integral of the beta pdf over the parameter space for p, this expression equals one. The posterior density is then

$$g(p\,|\,x_1,x_2,\ldots,x_m)=\frac{\Gamma(m+\alpha+S+\beta)}{\Gamma(m+\alpha)\Gamma(S+\beta)}\times\left(p^{m+\alpha-1}(1-p)^{S+\beta-1}\right),$$

which is a Beta $(m+\alpha,S+\beta)$ density.

It worked out that the posterior distribution is a form of the prior distribution updated by the new data. In general, when this occurs we say the prior is conjugate.

The Bayes estimator, under squared error loss, is the mean of the posterior $\mathrm{Beta}(m+\alpha, S+\beta)$ distribution,

$$\hat{p}_{\mathrm{Bayes}} = E[p \mid X] = \frac{(m+\alpha)}{(m+\alpha+S+\beta)} .$$

The Bayes credible set for p is given by (p_L, p_U) where p_L and p_U satisfy

$$\int_{p_L}^{p_U} g(p \mid X)dp = 1-\alpha ,$$

where α is the significance level. That is, the credible interval is computed from posterior beta cumulative density function as

$$P\left[\mathrm{Beta}_{\alpha/2}(m+\alpha, S+\beta) < p < \mathrm{Beta}_{1-\alpha/2}(m+\alpha, S+\beta)\right] = 1-\alpha .$$

When then say a $1-\alpha$ Bayesian credible interval for p is (p_L, p_U) where p_L is the quantile of order $\alpha/2$ and p_U is the quantile of order $1-\alpha/2$ for the beta distribution with parameters $m+\alpha$ and $S+\beta$.

3.2.3 Application of Bayesian Interval Estimation

In this section, we compute a Bayesian credible set for the proportion of yielding drivers. In the following chapter, we will investigate Bayesian interval estimation in detail with more applications and then compare it to the bootstrap method.

For the sample proportion of yielding drivers, it was found that $\hat{p}$ = 0.435, and the sample mean $\bar{x}$ = 1.299 for a sample size of 124. Before constructing a credible set, we assume some prior distribution model for the true proportion in the population. Since we have a geometric model here, we first use the Uniform distribution, or Beta(1,1), as our prior distribution. In the previous section, we found that the posterior distribution of p given X has a beta distribution with parameters m+1, S+1, where m = 124 and S = 161.06 for our example. The 95% Bayesian interval is the quantile of order 0.025 and the quantile of order 0.975 for the Beta(125,162.06). To estimate this interval, we use the function `qbeta(c(0.025,0.975),125,162.06)` in R software. It follows that the $100(1-\alpha)\%$ Bayesian credible set for the true proportion of yielding drivers p is

$$0.3786618 \leq p \leq 0.4930884 .$$

Figure 4 represents the plot of the posterior distribution of p given X where p is believed to have an Unif(0,1) prior distribution.

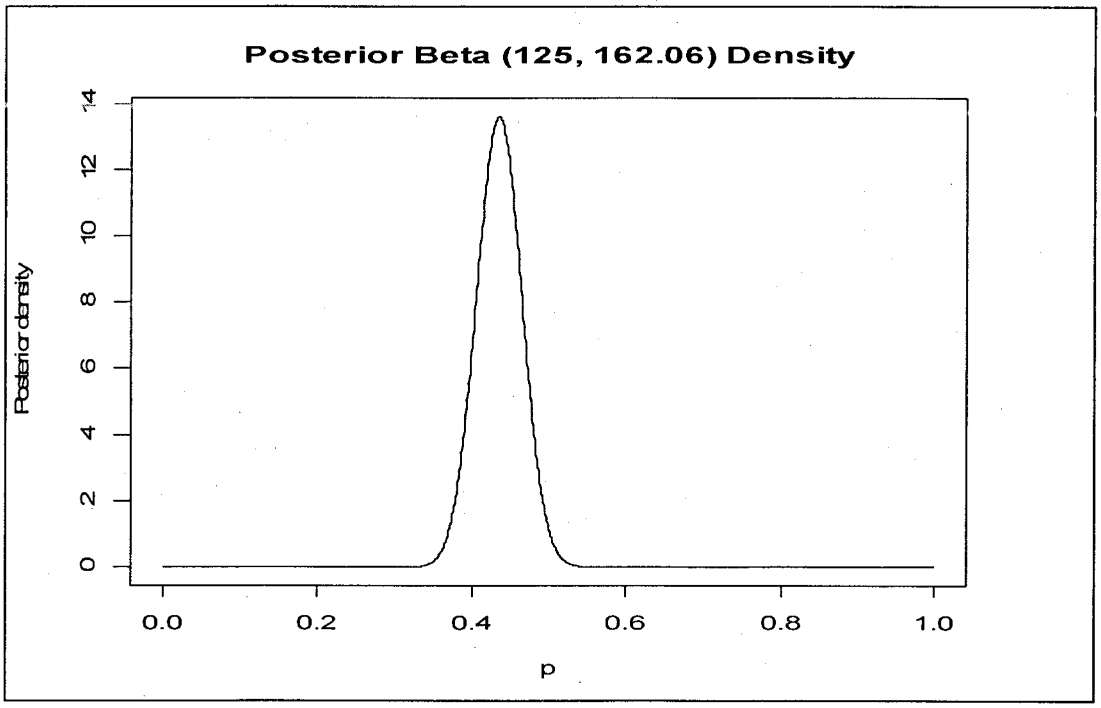

Figure 4 Plot of the Posterior Beta (125, 162.06) Distribution
for the true proportion of yielding drivers

Next, we assume that the true proportion of yielding drivers follows a Beta (α,β) distribution. The endpoints of a 95% Bayesian interval for p are then the lower and upper 0.025 quantiles of a Beta$(m+\alpha, S+\beta)$ distribution. If we let $\alpha = 5$ and β=5, then the 95% Bayesian credible set for p is

$$0.3811406 \le p \le 0.4940643$$

Figure 5 represents the plot of the posterior distribution of p given X, where p comes from Beta (5, 5) distribution.

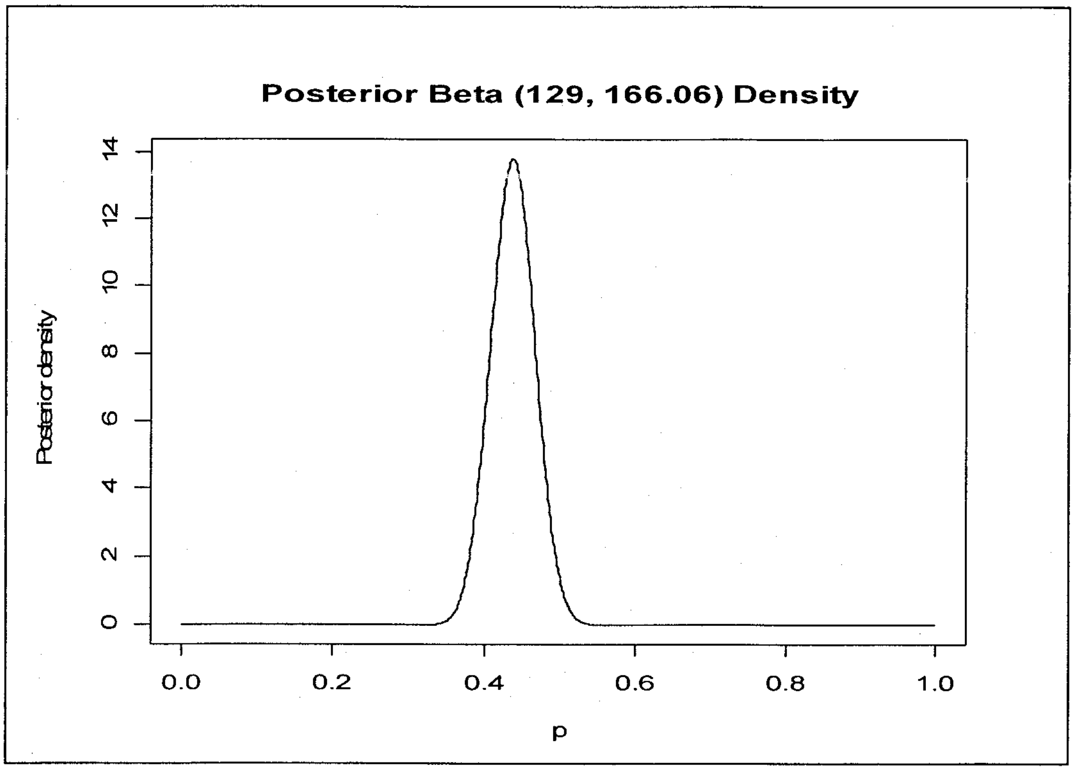

Figure 5 Plot of the Posterior Beta (129, 166.06) Distribution for the true proportion of yielding drivers

Therefore, based on the proposed Bayesian and bootstrap approaches, we have computed three different CIs for the true proportion of yielding drivers.

Table 4 shows the summary results of bootstrap CI and Bayesian intervals for the true proportion of yielding drivers.

Table 4 Bootstrap CI and Bayesian intervals for the true proportion of yielding drivers

95% Bootstrap CI	95% Bayesian Credible Set with Uniform (0, 1) prior	95% Bayesian Credible Set with Beta (5,5) prior
(0.348, 0.522)	(0.381, 0.494)	(0.378, 0.493)

As we can see there is no great difference between the bootstrap 95% CI and the Bayesian 95% credible set. But, note that the Bayesian intervals are shorter in length. Bayesian interval with Beta (5, 5) prior has the shortest length.

Based on these results, we cannot make a firm conclusion about the performance of these two methods for computing an interval estimate of the true proportion of a geometric distribution. To do so, we need to use Monte Carlo simulation in order to compare which interval procedures perform better in repeated samples from the geometric distribution with different sample sizes.

3.3 Monte Carlo Simulation Experiment

A Monte Carlo Simulation experiment was developed to investigate the performance of the bootstrap and the Bayesian method for computing an interval estimate of the true proportion for a geometric distribution. The Monte Carlo method assures that if the input of a simulation is generated from a known probability distribution and the simulation is repeated a large number of times, characteristics of the population will eventually emerge. This experiment is designed to simulate intervals by both the bootstrap and Bayesian methods, which will be compared in terms of estimated coverage (that is, we will estimate the proportion of the time that the intervals contain the true proportion p) and average length of the interval procedure.

Before we compare the two interval procedures, we will first investigate the performance of the bootstrap for the proportion of a geometric distribution.

3.3.1 Experiment 1: Investigation of Bootstrap Intervals

We have used p = 0.1, 0.2, ..., 0.9 in the following steps of a Monte Carlo Simulation experiment for the bootstrap approach used in this thesis, which are described below:

1. Generate n random samples from the geometric distribution of size m, that is sample $(x_{i1}, x_{i2}, \ldots, x_{im})$ comes from a GEO(p_i) distribution. An nxm data matrix is generated, and the row means are extracted as a vector.
2. Perform the bootstrap simulation technique mentioned in section 3.1 on each of the n samples, generating an estimate of the sampling distribution of $\hat{p}$.
3. Compute the bootstrap CI for the proportion of the geometric distribution by first arranging the $\hat{p}$ values in increasing order, and then finding the lower and upper 0.025 quantiles of the distribution in step 3.
4. Repeat steps 1-5 a large number of times (K). This generates K $100(1-\alpha)\%$ bootstrap CIs. Count the number of bootstrap CIs that contain the true proportion

p and divide by K. This generates the Estimated Coverage (%) . Then compute the average length of bootstrap intervals.

The above simulation experiment was programmed in R software using n = 1000 generated random samples of size m from a geometric distribution and K = 1000 iterations of the bootstrap process per generated sample.

For the problem at hand, we must know the true value of the population proportion of the geometric distribution so that the sampling properties of the bootstrap can be investigated. For this reason, we generated a sequence of success probabilities p between 0.1 and 0.9 (p =0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9).

One goal of this bootstrap simulation experiment is to try to find some factors that may affect our method for computing a CI for the geometric distribution. In particular, we are interested in comparing coverage probability and average length as a function of sample size and value of p.

We will generate random samples from the geometric distribution of sizes m = 10, 25, 50, 100. The results of this experiment will discussed in Chapter 4.

3.3.2 Experiment 2: Comparison of Bootstrap Intervals with Bayesian Credible Intervals

To perform this experiment, we begin by generating random samples of success probability p0 from a prior distribution. As we mentioned in section 3.2.2.2, Beta (α,β) are reasonable prior distributions for the proportion of a geometric distribution, in particular Beta (1, 1) and Beta (5, 5).

The steps in this simulation experiment are described below:

Perform N iterations of the following process:

1. Generate p0 from a Beta (α,β) prior.
2. Take a random sample of size m from a Geometric(p0) distribution.
3. Find a Bayesian credible interval from the resulting posterior.
4. Using the sample in (2) and the bootstrap method to get a bootstrap interval.
5. Check whether or not each of the Bayesian and bootstrap intervals contain p0.

After iterations are complete, find the coverage probability and average length of the set of simulated intervals.

CHAPTER 4

RESULTS

For experiment 1, the bootstrap interval procedure discussed in the previous chapter, a table of the estimated coverage (%) as a function of the true proportion p for different sample sizes (m) will be presented, as well as a table that contains the summary statistics of the length of bootstrap CIs. A Lattice plot of the estimated coverage (%) vs. p for each sample size will also be shown, along with a graph of average length vs. p and a graph of standard deviation vs. p.

For experiment 2 a table of the estimated coverage (%) and average length as a function of the sample size for each interval procedure will be presented.

4.1 Experiment 1:Bootstrap

In this section, samples are generated from a Geometric distribution with known parameter p (p = 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9), and the bootstrap as explained in detail in Chapter 3 is used. The results are summarized in Tables 5 and 6. It can be seen from Table 5 that the bootstrap method uniformly gives coverage smaller than the specified confidence (95%). However, as the sample size increases the estimated coverage gets larger and approaches the specified confidence. Observe that poor coverage occurs at large values of p (p =0.9). As p gets close to 1, the coverage decreases. Since the probability of success is high, number of failures will be very small. In these simulations, approximately 35% of the samples of size m = 10 had zero failures prior to the first success, so $\hat{p} = 1$ for all resamples and the bootstrap sampling distribution of $\hat{p}$ is degenerate at 1. For all such samples, the bootstrap CI is (1, 1), so cannot cover the true p = 0.9. As m increases, it becomes less likely that a sample consists entirely of zeros, so the coverage probability of the bootstrap intervals increases.

In term of average length (Table 6), as the sample size increases, bootstrap average length (CLT phenomenon) gets shorter. But for any sample size, average length increases as p increases but starts to get shorter as p gets closer to 1. Table 6 suggests that the lengths have more variability as p increases but less variability as sample size increases. (Again, a CLT phenomenon)

Figure 5 is a graph of the estimated coverage probability vs. parameter p shown in Table 5.

Figure 6 is a graph of the average length vs. parameter p shown in Table 6.

Figure 7 represents the graph of the standard deviation of length vs. parameter p.

Table 5 Bootstrap Estimated Coverage (%) as a function of p

Sample Size	p								
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
10	85.4	84.8	86.8	80.0	82.0	85.0	88.8	89.2	65.0
25	90.2	90.8	91.8	91.8	92.0	89.4	91.2	88.8	89.6
50	91.8	94.2	92.2	90.2	93.4	94.0	92.8	91.8	89.4
100	94.2	94.0	93.0	94.4	91.4	93.2	93.6	94.0	90.0

Table 6 Summary Statistics of Bootstrap Interval Length

Sample Size	Statistic	p								
		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
10	Mean	0.133	0.237	0.316	0.372	0.392	0.398	0.376	0.312	0.195
25	Mean	0.079	0.141	0.197	0.238	0.265	0.278	0.271	0.249	0.173
50	Mean	0.054	0.102	0.137	0.169	0.190	0.202	0.202	0.186	0.142
100	Mean	0.037	0.069	0.099	0.121	0.136	0.145	0.146	0.135	0.104
10	SD	0.058	0.089	0.109	0.105	0.107	0.097	0.104	0.133	0.156
25	SD	0.022	0.036	0.041	0.049	0.050	0.048	0.047	0.051	0.071
50	SD	0.010	0.018	0.022	0.025	0.026	0.027	0.025	0.028	0.032
100	SD	0.005	0.009	0.012	0.013	0.014	0.015	0.014	0.016	0.016

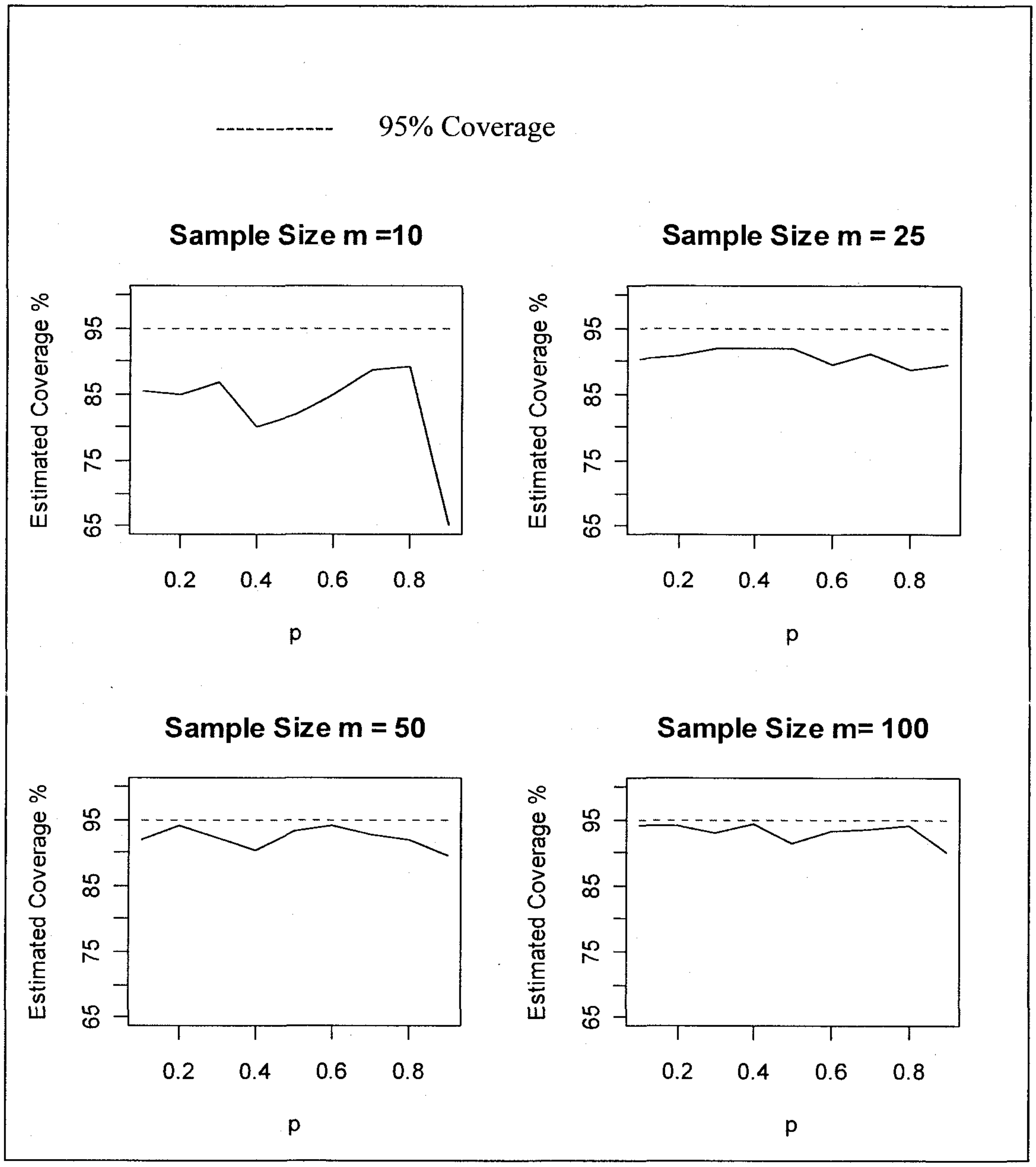

Figure 6 Lattice Plot of Bootstrap Estimated Coverage (%)
Vs. Proportion p for Different Sample Sizes

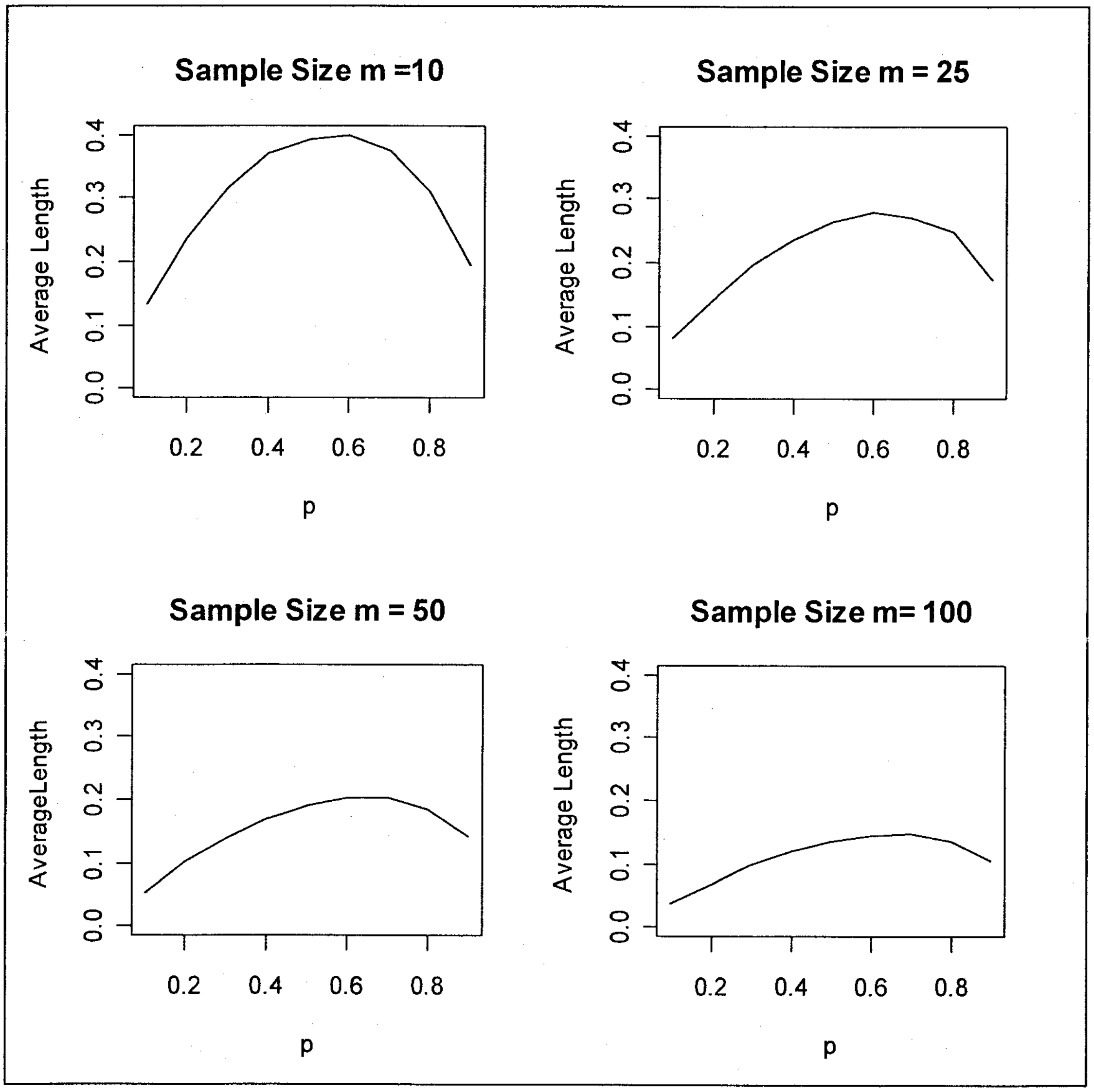

Figure 7 Plot of Bootstrap Average Length vs. Proportion p
For Different Sample Sizes

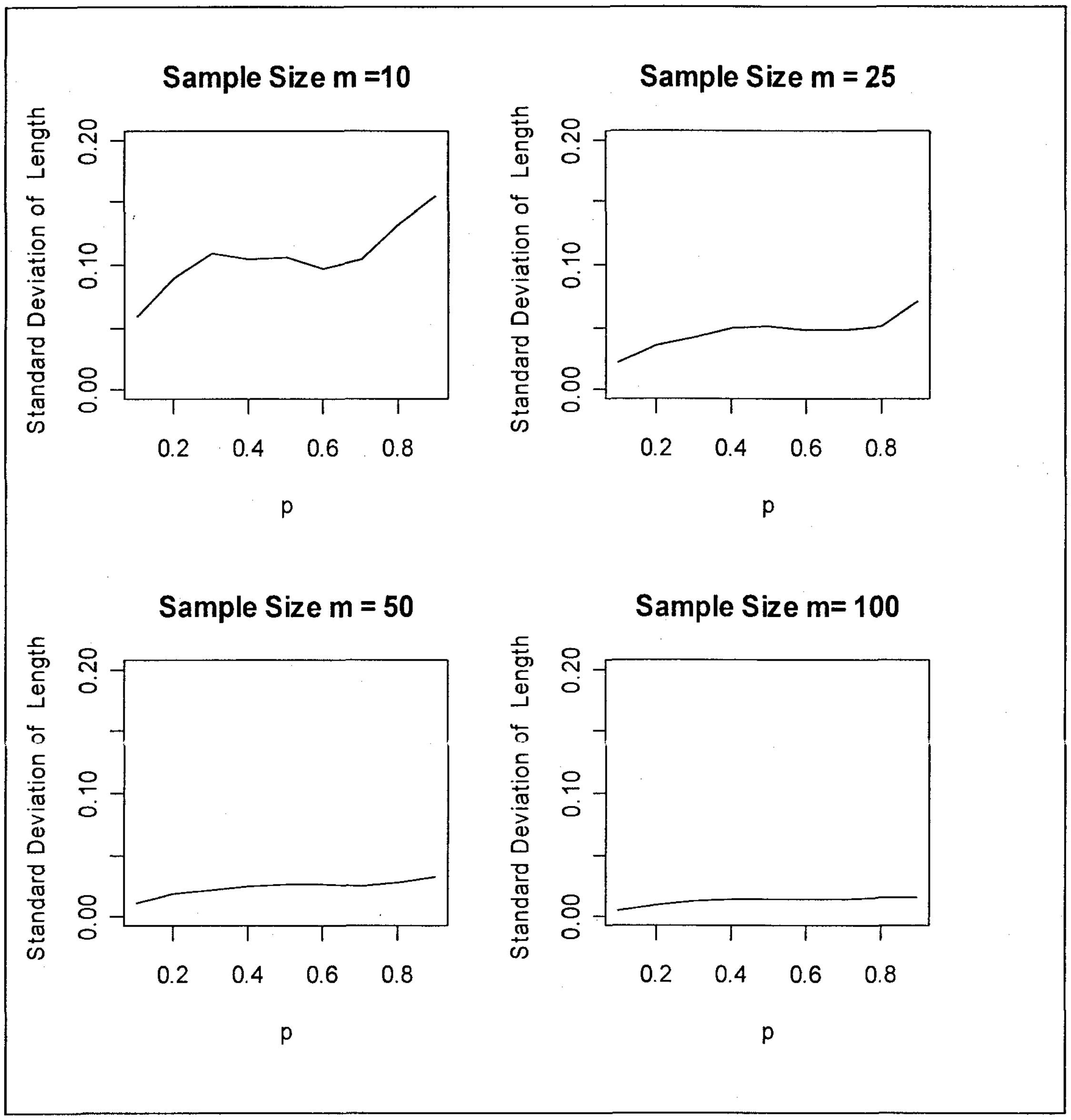

Figure 8 Plot of Bootstrap Standard Deviation of Length vs. Proportion p For Different Sample Sizes

4.2 Experiment 2: Bayesian and Bootstrap

In this section, samples are generated from a Geometric distribution with parameter p0, where p0 has a Beta (α,β) distribution, and the Bayesian method combined with bootstrap as explained in detail in Chapter 3 is used. In this experiment, we used Beta (1, 1) prior and Beta (5, 5) prior. As we mentioned in Chapter 3, these two priors are reasonable choices for p. For each type of prior, a table of estimated coverage and average length for the two interval procedures will be presented.

The results of this experiment are summarized in Tables 7 and 8. It can be seen from the tables that the Bayesian approaches produce coverage greater or equal than the specified confidence and greater than the bootstrap approach. However, the bootstrap intervals have shorter length than the Bayesian intervals. Notice again that the Bayesian intervals with Beta (5, 5) give slightly higher coverage but longer length than Beta (1, 1). This happened because the prior (Beta (5, 5) is informative.

Table 7 Summary of Bayesian and Bootstrap intervals as function of Estimated Coverage (%) and Average length with Beta (1, 1)

Sample Size	Estimated Coverage %		Average Length	
	Bayes	Boot	Bayes	Boot
10	94.93	79.22	0.3157	0.2751
25	95.31	87.79	0.2054	0.1921
50	94.98	91.05	0.1454	0.1402
100	95.12	92.29	0.1038	0.1016

Table 8 Summary of Bayesian and Bootstrap intervals as function of Estimated Coverage (%) and Average length with Beta (5, 5)

Sample Size	Estimated Coverage %		Average Length	
	Bayes	Boot	Bayes	Boot
10	94.91	85.50	0.3359	0.3690
25	95.02	91.37	0.2378	0.2491
50	95.26	93.02	0.1755	0.1789
100	95.30	93.71	0.1271	0.1282

CHAPTER 5

CONCLUSIONS

There are a number of methods for computing and analyzing confidence intervals of the parameter p for a geometric distribution. The two methods compared in this thesis are bootstrap confidence intervals and Bayesian credible sets.

After simulating a geometric distribution with sample sizes 10, 25, 50, and 100 and known parameter p, we are able to make the following conclusions regarding the bootstrap approach for computing an interval estimate for the proportion of a geometric distribution. Moreover, after simulating p, draws from a $\text{Beta}(\alpha,\beta)$, we are able to compare the two interval procedures.

1. The 95% Bootstrap CIs only cover the true proportion p approximately 93% of the time even in larger sample sizes. The Bootstrap coverage increases with sample size and with p, but starts to decrease when p gets close to 1. In general, the bootstrap intervals undercover the stated confidence level.
2. The Bootstrap average length decreases with sample sizes, but increases with p. However, it starts to decrease when p reaches high values (when p gets close to1).
3. The Bayesian method can be used to get a credible interval for the proportion of the geometric distribution if we believe that the proportion has some type of prior distribution. We used a Beta (α,β) distribution as the prior for p.

4. Bayesian approach gives coverage greater or equal than the specified confidence.
5. Beta (1, 1) and Beta (5, 5) are the two prior distributions that are appropriate for constructing Bayesian credible intervals for the proportion of a geometric distribution.
6. Compared to the classical parametric bootstrap method, the Bayesian approach worked better for computing credible intervals for the true proportion p of the geometric distribution.

It should be noted that the Bayesian method for this set of experiments (fixed p0 values) also yields specified coverage. It was observed (in another series of experiments not reported in this thesis) that if the chosen prior pdf is also too far from the true prior pdf, then the coverage may drop all the way to 0%. When p0 was generated from the Beta prior, the Bayes method gave specified coverage.

APPENDIX I

R SOURCE CODE

A copy of the R code used to obtain the results in Chapter 4 are presented below.

R code for experiment 1

```
bootsim <- function(x, p, N = 1000)
  {
  # Bootstrap method applied to a single geometric sample.

  # Generate an Nxm matrix of bootstrap resamples
   bootmat <- matrix(sample(x, N*length(x), replace = TRUE), nrow = N, byrow =
                         TRUE)
   bootmean <- rowMeans(bootmat)
   phat <- 1/(1 + bootmean)
   q <- quantile(sort(phat), c(0.025, 0.975))
   length.boot <- q[2] - q[1]
   cover <- p > q[1] && p < q[2]
   list(length.boot = length.boot, cover = cover)
  }

geomsim <- function(m, p0, N = 1000)
  {
  # Calling program that:
  #  (i) inputs the vector of success probabilities p0
  #  (ii) generates the geometric random samples
  #  (iii) produces the %coverage of the bootstrap intervals

 # Step 1: Generate the vector of probabilities and the geometric samples
  n<-length(p0)
  x <- matrix(rep(0, m*n), nrow = n)  # initialize data matrix
  for (i in seq(along = p0)) x[i, ] <- rgeom(m, prob = p0[i])
  meanx <- rowMeans(x)   # sample averages of each geometric sample
  # Step 2: Generate the bootstrap intervals, return the coverage indicator and length
  indic <- length.boot <- rep(NA, n)
  for(i in 1:n) {
     store <- bootsim(x[i, ], p0[i])
     indic[i] <- store$cover
     length.boot[i] <- store$length.boot
             }

  list(bootcov =as.numeric(indic),
      lboot = length.boot)
  }


geomrep <- function(m, p0, M = 500)
  {
   # Wrapper function to perform replicate simulations with geomsim

    n <- length(p0)
    cover.boot <- bootlgth <- matrix(0, nrow = M, ncol = n)
    for(k in 1:M)  {
      temp <- geomsim (m, p0)
      cover.boot[k, ] <- temp$bootcov
      bootlgth[k, ] <- temp$lboot
                }
     bootcov <- colMeans(cover.boot) * 100
     list(m = m, p0 = p0, bootcov = bootcov,
         lboot = bootlgth)
  }

# test runs

p0 <- seq(0.1, 0.9, by = 0.1)
run10 <- geomrep(10, p0)
run25 <- geomrep(25, p0)
run50 <- geomrep(50, p0)
run100 <- geomrep(100, p0)

###########
## Summaries for different sample sizes of coverage
###########

coverage10 <- (run10$bootcov)
coverage25 <- (run25$bootcov)
coverage50 <- (run50$bootcov)
coverage100 <- (run100$bootcov)
```

```
###########
## Summaries for different sample sizes of average length
##########

avglength10 <-(colMeans(run10$lboot))
avglength25 <- (colMeans(run25$lboot))
avglength50 <- (colMeans(run50$lboot))
avglength100 <-(colMeans(run100$lboot))

###########
## Summaries for different sample sizes of standard deviation of length
##########
sdlength10 <- (apply(run10$lboot, 2, sd))
sdlength25 <-(apply(run25$lboot, 2, sd))
sdlength50 <-(apply(run50$lboot, 2, sd))
sdlength100 <- (apply(run100$lboot, 2, sd))


# true coverage
truecov <- c(95, 95,  95, 95, 95, 95, 95, 95, 95)

# combine the true coverage and the estimated coverage
mat1 <- cbind(coverage10, truecov)
mat2 <- cbind(coverage25, truecov)
mat3 <-cbind(coverage50, truecov)
mat4 <-cbind(coverage100, truecov)

# Create a 2 x 2 plot for each true coverage/ estimated coveage  set
par(mfrow = c(2, 2))
matplot(p0, mat1, type = 'l', lty = c(1, 2), xlab = "p",
   ylab = "Estimated Coverage %", main = "Sample Size m =10")
matplot(p0, mat2, type = 'l', lty = c(1, 2), xlab = "p",
   ylab = "Estimated Coverage %", main = "Sample Size m = 25")
matplot(p0, mat3, type = 'l', lty = c(1, 2), xlab = "p",
   ylab = "Estimated Coverage %", main = "Sample Size m = 50")
matplot(p0, mat4, type = 'l', lty = c(1, 2), xlab = "p",
   ylab = "Estimated Coverage %", main = "Sample Size m= 100")
invisible()

# Create a 2 x 2 plot for average length
par(mfrow = c(2, 2))
plot(p0, avglength10, type = 'l', lty = 1, xlab = "p",
   ylab = "Average Length ", main = "Sample Size m =10")
plot(p0, avglength25, type = 'l', lty = 1, xlab = "p",
   ylab = "Average Length ", main = "Sample Size m = 25")
plot(p0, avglength50, type = 'l', lty = 1, xlab = "p",
   ylab = "AverageLength ", main = "Sample Size m = 50")
plot(p0, avglength100, type = 'l', lty = 1, xlab = "p",
   ylab = "Average Length ",  main = "Sample Size m= 100")
invisible()


# Create a 2 x 2 plot for standard deviation of  length
par(mfrow = c(2, 2))
plot(p0, sdlength10, type = 'l', lty = 1, xlab = "p",
   ylab = "Standard Deviation of  Length ", main = "Sample Size m =10")
plot(p0, sdlength25, type = 'l', lty = 1, xlab = "p",
   ylab = "Standard Deviation of  Length ", main = "Sample Size m = 25")
plot(p0, sdlength50, type = 'l', lty = 1, xlab = "p",
   ylab = "Standard Deviation of  Length ", main = "Sample Size m = 50")
plot(p0, sdlength100, type = 'l', lty = 1, xlab = "p",
   ylab = "Standard Deviation of  Length ",  main = "Sample Size m= 100")
invisible()
```

R code for experiment 2

```
asmeth <- function(alpha, beta, m, N = 10000)
  {
  # Perform N iterations of the following process:
  #    * randomly sample p0 from a Beta(alpha, beta) prior
  #    * take a random sample of size m from a Geometric(p0) distribution
  #    * find Bayesian credible interval from resulting posterior
  #    * independently, use bootstrap method to get a bootstrap interval
  #    * check whether or not each of the Bayesian and bootstrap intervals
  #contain p0
  #
  # After iterations are complete, find coverage probability and average length
  # of the set of simulated intervals.

  p <- rbeta(N, alpha, beta)
  boot.cover <- boot.length <- bayes.length <- bayes.cover <- numeric(N)
   for(i in 1:N) {
    x <- rgeom(m, p[i])
    s <- sum(x)
    bci <- qbeta(c(0.025, 0.975), alpha + m, beta + s)
    bayes.cover[i] <- bci[1] < p[i] && p[i] < bci[2]
    bayes.length[i] <- diff(bci)
    xx <- bootsim(x, p[i])
    boot.cover[i] <- xx$cover
    boot.length[i] <- xx$length.boot
            }
   bayes.cp <- mean(as.numeric(bayes.cover))
   boot.cp <- mean(as.numeric(boot.cover))
   avgl.bayes <- mean(bayes.length)
   avgl.boot <- mean(boot.length)
   cat(paste("Prior: Beta(", alpha, ", ", beta, ")", sep = ""), "\n\n")
   cat("          Coverage probability    Average length\n")
   cat(paste("Bayesian:          ", round(bayes.cp, 4), "                ",
round(avgl.bayes, 4), sep = ""), "\n")
   cat(paste("Bootstrap:         ", round(boot.cp, 4), "                ",
round(avgl.boot, 4), sep = ""), "\n")
   invisible()
  }

# Test runs:

asmeth(1, 1, 10)
asmeth(1, 1, 25)
asmeth(1, 1, 50)
asmeth(1, 1, 100)
asmeth(0, 0.5, 10)
asmeth(0, 0.5, 25)
asmeth(0, 0.5, 50)
asmeth(0, 0.5, 100)
asmeth(5, 5, 10)
asmeth(5, 5, 25)
asmeth(5, 5, 50)
asmeth(5, 5, 100)
```

BIBLIOGRAPHY

Bain L.J., Engelhardt M. Introduction To Probability and Mathematical Statistics. 2nd Edition. PWS-Kent Publishing Company, Boston, Massachusetts, USA. 1992.

Bernardo J.M., Smith A.F.M. Bayesian Theory. Wiley. 1994.

Bhat U.N. Elements of Applied Stochastic Process. 2nd Edition. New York:Wiley. 1984.

Casella G, Berger R.L. Statistical Inference. Duxbury Press, 2001.

Chen R. A surveillance system for congenital malformations. Journal of the American Statistical Association. pp (323-327). 1978.

Daniels H.E. Mixtures of geometric distributions. Journal of the Royal Statistical Society. Series B. pp (409-413). 1961

Davison A.C, Hinkley D.V. Bootstrap Methods and their Application. Cambridge, University Press.1997.

Efron B. Tibshirani. R.J. An Introduction to the Bootstrap. Chapman & Hall. 1993.

Gabriel K.R, Neumann J. On a distribution of weather cycles by length. Quarterly Journal of the Royal Meteorologica; Society. pp (375-380). 1962.

Jagers P. How many people pay their tram fares? Journal of American Statistical Association. pp (801-804). 1973.

Johnson N.L., Kotz S., Kemp A.W.. Univariate Discrete Distributions. New York: John Wiley and Sons. 1992.

Merrill J.C. VISSIM Model Development of a Mid-Pedestrian Crossing between Signalized Intersections. Thesis.pp (41-43-, 80-84). 2005.

Pfeiffer P.E. Probability for Applications. Springer Verlag, New York, Berlin, Heidelberg. 1990.

Pielou E.C Runs of one species with respect to another in transect through plant populations. Biometrics. Pp (579-593). 1963.

Seber G.A.F.. The estimation of Animal Abundance.2nd Edition. London, Griffin. 1982.

Taylor H.M., Karlin S. An Introduction to Stochastic Modeling. Orlando, Fl: Academic Press. 1984.

Verzani J. Using R for Introductory Statistics. Chapman and Hall/CRC Press. 2005.

VITA

Graduate College
University of Nevada, Las Vegas

Majgan Beria

Home Address:
4484 Sweet Stone Pl
Las Vegas, NV 89147

Degrees:
Bachelor of Science, Mathematical Sciences, 2004
University of Nevada, Las Vegas

Thesis Title:
Confidence interval estimation for a geometric distribution

Thesis Examination Committee:
Chair, Dr.Ashok K. Singh, Ph.D.
Committee Member, Dr.Dennis Murphy, Ph.D.
Committee Member, Dr.Rohan Dalpatadu, Ph.D.
Graduate Faculty Representative, Dr.Anthony Lucas, Ph.D.