

# Genuine Exact Two-Stage Methodologies for Producing Assigned Accuracy Estimators for a Gamma Mean

Kevin P. Tolliver<sup>1</sup>

U.S. Census Bureau

Office of Statistical Methods and Research for Economic Programs

Kevin.p.tolliver@census.gov

## Abstract

When sampling from largely right-skewed populations it is better to assume the population is Gamma rather than Normal. The Gamma distribution is often assumed when modeling mean-time-to-failure in the biological field of Survival Analysis and the engineering field of Reliability Analysis. The paper often makes references to estimating times-to-events in clinical trials. The sequential methods in this paper are used to determine what sample size is required to attain an accurate estimator assuming the data comes from a Gamma population. This paper proposes two methods for finding estimators with pre-assigned accuracy: (1) point estimator and (2) an interval estimator. It implements a genuine two-stage sampling procedure. The term genuine refers to the fact that, in contrast to previous methods, the procedures proposed herein are based on the combined samples from both stages, rather than ignoring the data from the first-stage sample. Theoretical results are exact, which means at no point was an asymptotic or large sampling approximation used and all the derivations assumed an underlying distribution of Gamma. These results are accompanied by more practical solutions. Results are found for when shape is both known and unknown.

## I. Introduction

The Gamma distribution is a flexible right skewed density that has a wealth of applications, and because of the flexibility of the family of Gamma distributions it is often assumed random variables such as time-to-event has an underlying Gamma distribution. It is not restrained to just time. As a family of distributions the Gamma distribution can be assumed in any area where values have a positive support and have justifiably large values in the support. For example, it is assumed as the underlying distribution for both precipitation rates and precipitation intensity, Maureil (2007), censor imaging, Chatelian (2007) and (2008), and often with general queues. The Gamma mean is most widely used in statistically modeling time-to-events in Survival and Reliability Analysis seen in biomedical and engineering fields. Perhaps, it is most commonly used in clinical trials. For example, in a drug trial the time estimated until a drug goes into effect can be modeled with an underlying Gamma distribution. Many clinical trials are subject to budgetary restraints and oversampling as described by Wald (1947) can lead to high costs, since each volunteer in the study is typically compensated. To reduce spending sampling procedures like the two laid out in subsequent sections can be implemented and still ensure accuracy. For this paper, references to mean time-to-event will be associated with the Gamma mean even though a Gamma mean is not exclusive to times. In order to correctly describe the process it is pertinent to develop an accurate estimator of the mean with low bias and low variation. This paper assumes the estimator is unbiased and focuses on developing a sampling procedure that will ensure low variation

Suppose there is a sequence of independent observations  $X_1, X_2, \dots, X_n$  having a common Gamma distribution with the density function

$$f(x) = \frac{x^{\theta-1} e^{-x/\lambda}}{\Gamma(\theta)\lambda^\theta}, \text{ for } x > 0 \quad (1.1)$$

---

<sup>1</sup> This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. Any views expressed on statistical, methodological, technical, or operational issues are those of the authors and not necessarily those of the U.S. Census Bureau.

with the mean  $\theta\lambda$  representing the average time needed till the event occurs. Observing  $X_1, X_2, \dots, X_n$  the mean time  $\theta\lambda$  can be estimated by the sample mean  $\bar{X}_n = n^{-1} \sum_{i=1}^n X_i$ . The intent of this paper is to develop a sampling scheme that will produce a reliable estimator for the mean. The variance of the mean of a Gamma population is well-documented to be  $\theta\lambda^2 n^{-1}$ . As is commonly seen in literature, a constant  $A$  is added to penalize the variance more or less as the user sees fit. There are two specific goals in this paper: (1) to develop a sampling scheme estimator with variance below a preassigned value  $w$

$$A \cdot \text{Var}(\bar{X}_n) = \frac{A\theta\lambda^2}{n} \leq w \quad (1.2)$$

and (2) to produce an interval estimator  $C(X)$  with preassigned interval width and coverage probability  $1-\alpha$ .

1.  $P(\theta\lambda \in C(X)) \geq 1 - \alpha$
2.  $C(X) \leq 2d$

When no prior knowledge of the distribution is known then there is no fixed sample size solution. The problem requires drawing an initial sample size before determining the final sample size, making this a sequential problem. For this problem there are only two sets of data: the initial set and subsequent set. This particular type of sequential sampling procedure is referred to as a two-stage design.

Historically, when estimating times until event of random processes researchers have used sample designs based on incorrect assumptions about the underlying distribution. For many years the underlying distribution was assumed to be Normal, even for estimates with positive support and distributions that are right skewed. A two-stage bounded variance estimator for Normal populations that ensured an accurate mean estimator was given by Stein (1945). As noted, assuming random time processes are Normal may be inaccurate. It might be more appropriate to assume an Exponential underlying distribution. This distribution is right-skewed and a relative of the Gamma distribution. An exact two-stage design described a terminal size that ensured the variance of the estimator falls below a preassigned variance assuming an Exponential population was given by Birnbaum and Healy (1960). The term "exact" refers to the fact that this result does not assume anything asymptotically about the initial observations and it never approximates the underlying distribution. This procedure does, however, use the initial sample size to determine the final sample size and ignores the initial sample size in the final estimate. This was later improved by combining both the initial and secondary samples together to produce the estimate of Kubokawa (1989), making the solution a genuine two-stage sampling procedure. The term "genuine" refers to the fact that, in contrast to previous methods, the procedures proposed are based on the combined samples from both the first and second stages, rather than ignoring the data from the first-stage sample. Accordingly, the terminal sample size and the estimate are no longer independent, which complicates the theory development significantly. With a relatively large number of initial observations this result did as well as the Birnbaum and Healy method. A two-stage sampling procedure that is both genuine and exact assuming Exponential underlying distribution is given by Mukhopadhyay and Pepe (2006). If  $X_1, X_2, \dots, X_m$  are Exponential with mean  $\lambda$ , and

$$B = \frac{2Am(m+1)}{(m-1)(m-2)} \quad (1.4)$$

then the terminal sample size

$$N = \max \left\{ m, \left\langle \frac{B\bar{X}_m^2}{w} \right\rangle + 1 \right\} \quad (1.5)$$

will ensure the variance is bounded below the predetermined variance  $w$ , where  $\langle \bullet \rangle$  is the integer value of  $\bullet$ . This procedure was extended to the Gamma distribution when shape is known by Tolliver and Carpenter (2008), given in section II.

Much like the bounded variance problem, two-stage bounded interval estimators for Normal populations have been studied extensively, Stein (1949). A solution for a general density was found, Chow and Robbins (1965), and a more specific solution for an Exponential density was found, Govindarajulu (1995). However, both of these results

assume Normal approximation at some point in their work. In fact, there is little research for fixed-width confidence intervals for asymmetric distributions. Many papers on fixed-width confidence intervals provide intervals of the form  $C(X) = \{\mu \mid \bar{X}_n - d < \mu < \bar{X}_n + d\}$ , making the width of the interval estimator completely independent of the final sample mean itself. This is a luxury asymmetric distributions such as the Gamma do not have. However, there will be an explicit interval estimator of that form provided in this paper. The difference is that since the Gamma is not symmetric, the upperbound and lowerbound will not be equidistant from the sample mean. The final interval estimator will be completely independent of the final sample mean much like the Normal case. That is, our confidence interval  $C(X)$  will be of the form

$$C(X) = \{\mu \mid \bar{X}_n - kd < \mu < \bar{X}_n + (2-k)d\} \quad (1.6)$$

where  $k$  is an unknown nonnegative real number less than two. This quantity signifies that the two lengths will not be the same.

## II. Bounded Variance Estimator

The first goal of the paper is to develop bounded variance estimators when shape  $\theta$  is known and scale  $\lambda$  is unknown and when both parameters are unknown. This result will lead to the development of a bounded interval estimator, as the two concepts are related. When sampling from Normal populations with known variance, the methods used to bound variance and the methods used to bind the width of the interval estimator are the same. This is seen in many undergraduate textbooks. When sampling from Normal populations with unknown variance the two methodologies are similar, Stein (1945) and Stein (1949). The idea is that if the variance can be bounded then consequently the interval estimator which is a function of the variance can also be bounded.

As stated before, we desire bounding the variance under some predetermined bound, as in (1.2). This means that the number of observations  $n^*$  that will ensure the variance is within the bound is  $A\theta\lambda^2w^{-1}$ . Ignoring the fact that the quantity above may not be an integer, we allow the optimal sample size

$$n^* = \left\langle \frac{A\theta\lambda^2}{w} \right\rangle + 1.$$

Notice  $n^*$  is dependent on the unknown parameters, so a sequential sampling procedure must be implemented so some knowledge can be gained on these parameters. A pilot sample of  $m$  observations,  $X_1, X_2, \dots, X_m$  i.i.d variables will be taken following a Gamma distribution  $(\theta, \lambda)$ , with  $m\theta \geq 3$ . At this point, one might note a theoretical dilemma. Similarly to the optimal sample size being dependent on the unknown parameters, an assumption must be made using an unknown parameter. However, in practice  $m\theta > 3$  is a fairly safe assumption since  $\theta$  is not usually miniscule. From this sample the maximum likelihood estimator of the mean is  $\bar{X}_m$ . That estimate is used to determine the terminal sample size  $N$ . To reemphasize, there is no fixed sample size solution to this problem. In addition to that, this method may also decrease the expected sample size and this will be a key tool in the construction of the interval estimator. After observing the first  $m$  observations, our first stage, a decision is needed to determine if the procedure can continue with the  $m$  observations, or if more need to be added, our second stage. This yields the two-stage procedure:

*Theorem 1. Let  $X_1, X_2, \dots, X_m$  i.i.d. Gamma( $\theta, \lambda$ ) initial observations be drawn, with  $m\theta \geq 3$ . The terminal sample size*

$$N = \max \left\{ m, \left\langle \frac{B\bar{X}_m^2}{\theta w} \right\rangle + 1 \right\}, \quad (2.1)$$

where

$$B = \frac{A(2m^2\theta^2 + 2m\theta)}{(m\theta - 1)(m\theta - 2)} \quad (2.2)$$

if  $\lambda$  unknown and  $\theta$  known and

$$B = \frac{A(8m)}{3m - 9} \quad (2.3)$$

if both parameters are unknown, will ensure that the variance over all  $N$  observations will be less than or equal to the predetermined variance  $w$ ; that is, that  $A \cdot \text{Var}(\bar{X}_N) \leq w$ ,

The proof is given in the appendix (A.1).

The proof ensures the variance is beneath the bound. However after exploring the distribution of  $N$ , the sampling procedure still draws a large number of observations. Recall  $n^* = A\theta\lambda^2/w$ . This is due to the fact that the ratio of  $B$  to  $A$  will always be larger than two, causing  $N$  to be larger than  $n^*$ . Instinctively it would appear necessary that reducing  $B$  50% would reduce sample size and still maintain the variance bound. Because of the skewness in the distribution of the variable  $\bar{X}^2$ , simply dividing by two does not work.

The terminal sample size can be reduced further along the lines of Mukhopadhyay and Zacks (2007). In that paper the authors conclude they can reduce  $B$  by investigating the distribution of the variance under the sampling procedure. This was done by looking at the variance as a function of  $\lambda$  and identifying what value of  $\lambda$  gives the maximal variance, and then empirically decreasing  $B$  so that maximal variance is just within the bound. This yielded a new  $B$  as a ratio of the old  $B$ . The same was done for this study for different values for  $\theta$ . Varying  $\lambda$  from one to forty,  $\theta$  from one to forty, and fixing everything extensive simulations were performed. For each value of  $\theta$  the value of  $\lambda$  that produced the highest value of variance. Afterwards, each of these were repeated empirically decreasing  $B$  so that risk would increase but not exceed its bound. Finally, a regression was implemented with  $B$  and  $\theta$  to produce a new value for  $B$

$$\begin{aligned} B_{new} &= 0.6B & \theta < 1 \\ B_{new} &= [-0.028\log(\theta) + 0.6]B & 1 \leq \theta \leq 20 \\ B_{new} &= 0.505B & \theta > 20 \end{aligned}$$

This new  $B$  will give a smaller expected value of  $N$  and should still give a variance less than  $w$ . We can see there is a significant reduction in this ratio, which will result in a decreased sample size.

To verify this result a simulation study was conducted using R software. In the simulation, differing values for the optimal sample size were chosen: 25, 50, 100, and 500. Each of these values will have a corresponding variance bound. We fixed  $\lambda = 5$  since the result is not dependent upon knowledge of this parameter and vary  $\theta = \{1, 2, 5, 10\}$ .  $A$  is a constant expression that is chosen to be two. One thousand replications were used for each case. The quantity  $\bar{N}$  is an expected value of  $N$  over 1,000 replications and  $\bar{se}^2$  is an estimate of the variance with the original terminal sample size. The simulation is repeated using the bound coefficient when both parameters are unknown. Tables 2.1 and 2.2 show how the expected value for  $N$  compares with the optimal sample size and how the expected value for variance compares with the variance bound.

**Table 2.1. Improved Variance Estimates (Shape Known)**

	$n^*$	$w$	$\bar{N}$ $m=10$	$se^2$	$\bar{N}$ $m=20$	$se^2$	$\bar{N}$ $m=30$	$se^2$
$\theta=1$	25	2.000	51.12	1.576	40.13	1.603	38.48	1.175
	50	1.000	99.63	0.975	77.97	0.930	71.07	0.853
	100	0.500	195.44	0.472	152.96	0.441	140.31	0.431
	200	0.250	403.03	0.195	311.61	0.223	284.67	0.221
$\theta=2$	25	4.000	38.73	4.038	33.80	3.695	34.14	2.378
	50	2.000	74.12	1.732	65.31	1.885	63.03	1.831
	100	1.000	147.23	0.964	132.26	0.856	126.16	0.925
	200	0.500	293.12	0.426	265.22	0.442	248.10	0.453

$\theta=5$	25	10.000	31.26	9.634	29.85	8.589	31.43	7.082
	50	5.000	61.70	4.727	59.07	4.561	57.73	4.893
	100	2.500	123.99	2.145	117.01	2.129	114.25	2.149
	200	1.250	304.12	0.872	288.75	0.904	286.97	0.862
$\theta=10$	25	20.000	28.66	18.523	27.76	19.214	30.42	14.983
	50	10.000	56.89	9.657	55.43	9.142	54.96	9.569
	100	5.000	112.81	4.866	110.75	4.736	108.70	4.474
	200	2.500	227.25	2.275	218.60	2.496	218.37	2.234

**Table 2.2. Improved Variance Estimates (Shape Unknown)**

	$n^*$	$w$	$\bar{N}$ $m=10$	$se^2$	$\bar{N}$ $m=20$	$se^2$	$\bar{N}$ $m=30$	$se^2$
$\theta=1$	25	2.000	63.92	1.395	50.96	1.409	47.17	1.216
	50	1.000	130.05	0.688	100.38	0.672	91.97	0.705
	100	0.500	258.59	0.348	195.71	0.338	184.77	0.334
	200	0.250	506.92	0.161	384.24	0.163	378.55	0.163
$\theta=2$	25	4.000	118.21	1.060	95.35	1.224	90.02	1.311
	50	2.000	232.64	0.583	190.80	0.638	180.93	0.614
	100	1.000	470.88	0.263	387.76	0.280	359.00	0.278
	200	0.500	931.85	0.130	737.12	0.147	723.75	0.159
$\theta=5$	25	10.000	291.32	0.939	238.50	1.133	220.97	1.130
	50	5.000	587.52	0.475	471.27	0.537	443.80	0.538
	100	2.500	1149.31	0.245	945.17	0.270	894.37	0.269
	200	1.250	2920.06	0.091	2354.86	0.114	2215.37	0.126
$\theta=10$	25	20.000	577.71	0.903	471.60	1.080	443.44	1.211
	50	10.000	1152.15	0.439	945.68	0.534	888.98	0.559
	100	5.000	2301.34	0.230	1868.90	0.297	1769.56	0.290
	200	2.500	4563.46	0.110	3747.98	0.131	3556.06	0.152

The very first thing that should be noted is how in each case, the estimated value for the variance is below the variance bound  $w$  for every case except for  $m=10, \theta = 2, w = 4.0$  in Table 2.1. These are only estimates but this is a good indication that mean variance falls below the variance bound. This shows that the sampling procedure does a decent job of producing a variance estimate below the variance predetermined bound. Furthermore, note in the shape known case how close the estimated variance value gets to  $w$ . Because of the inverse relationship between sample size and variance, this means that the sampling procedure does its job of reducing sample size. This is beneficial for any trial that is trying to reduce cost for experiments. With that said the bound coefficient  $B$  is a function of the scale parameter, and decreases as the value for  $\theta$  increases. The sampling procedure continues to over sample when it lacks knowledge of shape. Future research might entail using the initial sample size to find a bound for  $\theta$ .

### III. Genuine Two-Stage Interval Estimator

Many users feel that it is more useful to report interval estimators than other measures of variation because interval estimators give more interpretable results, Ramsey and Schafer (2002). For example in manufacturing applications an interval estimator might be useful for warranty purposes.

This is why interval estimators are beneficial. However, while there is a plethora of research on fixed-width confidence intervals, there is little research on fixed-width confidence intervals for asymmetric distributions using exact methods. Unlike the Normal distribution and other symmetric distributions the width of the Gamma interval estimator is dependent on the final sample mean, whereas the Normal populated sample will be completely

independent of the location of the final sample mean. One can see below an example of how a Gamma estimator differs from a Normal estimator.

The interval estimator for a Normal distributed population when the variance  $\sigma^2$  is known is

$$C(X) = \left\{ \mu \mid \bar{X}_n - z_{\alpha/2} \frac{\sigma}{n} < \mu < \bar{X}_n + z_{1-\alpha/2} \frac{\sigma}{n} \right\}$$

and a well-documented example (Casella and Berger pp. 429) of an interval estimator for a Gamma distributed population when the shape  $\theta$  is known is

$$C(X) = \left\{ \lambda \mid \frac{2\bar{X}_n}{c_{1-\alpha/2}} < \lambda < \frac{2\bar{X}_n}{c_{\alpha/2}} \right\} \quad (3.1)$$

where  $c_q$  is the  $q$ th quantile of a Chi-Square random variable with  $2n\theta$  degrees of freedom.

Arriving at this term is done simply with inverting the statistic  $\bar{X}$  and using the Chi-Square transformation. For more information on this see Appendix A.2. Unlike the Normal case where the width of the interval estimate is only dependent on sample variation, the width of this interval estimator is completely dependent on the final sample mean. If one were to create a two-stage sampling design with this type of interval estimator, then one can substitute  $\bar{X}_m$  for the final sample mean  $\bar{X}_N$ . This produces a decent estimate of the sample size needed to fulfill properties given in (1.3) but mathematically lacks the rigor to ensure them. In fact, there is no way of bounding the interval estimator without finding a bound on the sample mean of all  $N$  observations when using an interval estimator of this form. The proposed sampling methodology is unlike that of any interval estimator for skewed populations and will not resemble (3.1) above at all. The interval estimator we propose will be similar to (1.6) making it independent of the final sample size. Since the Gamma population is asymmetric the sample mean will not be in the center of the interval estimate. The optimal sample size must be redefined to be the smallest  $n$  for which both criteria (1.3) are achieved,

$$n^* = \{n \mid P(\lambda \in C(X)) \geq 1 - \alpha, C(X) \leq 2d\} \quad (3.2)$$

According to Ghosh (1991), terminal sample sizes for fixed-width confidence intervals should have the following properties:

1.  $N$  is non-decreasing in  $2d > 0$ .
2.  $N$  is finite with probability 1 for every  $2d > 0$ .
3.  $N/n^* \rightarrow 1$  as  $2d \rightarrow 0$  in probability or almost surely
4.  $E(N)/n^* \rightarrow 1$  as  $2d \rightarrow 0$ .
5.  $\lim_{d \rightarrow 0} P(\mu \in C(X)) = 1 - \alpha$

This section only provides a solution if shape is known. There are a number of reasons why the shape known case is studied: (1) there are particular instances where the shape parameter is either known or can be assumed as known similar to how variance can be assumed known with the Normal case, (2) studying the shape known case allows us to see how robust the Exponential assumption is, and (3) it gives ground work for methodologies where the shape parameter is unknown.

*Theorem 2. Let  $X_1, X_2, \dots, X_m$  i.i.d. Gamma  $(\theta, \lambda)$  initial observations be drawn, with  $m\theta \geq 3$ . For significance level  $\alpha$  and predetermined width  $d$ , if  $N$  is defined in (2.1) and  $g_q$  is the  $q$ th quantile of the Gamma( $n\theta, 1/n$ ) distribution. the terminal sample size*

$$M = \min \left\{ n \geq N \mid \sqrt{\frac{Nw}{A\theta}} [g_{1-\alpha/2} - g_\alpha] = 2d \right\} \quad (3.3)$$

*will yield an interval estimator*

$$C(X) = \left\{ \theta\lambda \mid \bar{X}_M - \sqrt{\frac{Nw}{A\theta}} [g_{1-\alpha/2} - \theta] < \theta\lambda < \bar{X}_M - \sqrt{\frac{Nw}{A\theta}} [g_{\alpha/2} - \theta] \right\} \quad (3.4)$$

that has the properties given in (1.3)

The proof is given in Appendix A.3.

The proof validates the properties will hold, but the sampling procedure itself can be evaluated with simulations conducted in R software. These simulations will show whether some of the properties given by Ghosh hold and affirm our belief that the coverage probability is larger than the predetermined coverage probability. It is not necessary to test to see if the terminal sample size  $M$  will fall within the width bound; the terminal size by nature will always fall within the width bound since it utilizes a numeric procedure.

The first simulation illustrates how the terminal sample size increases as the predetermined bound  $2d$  decreases and also how initial sample size affects the terminal sample size. The shape parameter was varied  $\theta = \{1, 2, 5, 10\}$  and  $w = \{5, 10, 25, 50\}$  so that the ratio between the two is the same throughout. The variance bound was varied  $\{10, 20, 30\}$ , and  $2d = \{4, 3, 2, 1\}$ . The risk bound is fixed to be three and the coverage probability is fixed to be 0.90. This simulation used 1,000 replications from a Gamma population with scale equal to five.

**Table 3.1. Relationship between  $2d$ ,  $p$ , and Terminal Sample Size  $M$**

		m=10		m=20		m=30	
		$\bar{M}$	P	$\bar{M}$	P	$\bar{M}$	P
$\theta=1,$ w=5	4.0	35.8	0.921	35.7	0.988	50.5	0.998
	3.0	63.4	0.941	64.8	0.985	90.0	0.994
	2.0	140.1	0.951	143.5	0.988	202.7	0.994
	1.0	573.8	0.960	575.1	0.977	812.6	0.969
$\theta=2,$ w=10	4.0	102.3	0.915	92.3	0.919	105.1	0.971
	3.0	170.3	0.893	164.5	0.937	187.2	0.967
	2.0	408.6	0.930	370.8	0.935	420.2	0.958
	1.0	1649.5	0.933	1481.2	0.937	1682.1	0.958
$\theta=5$ w=25	4.0	527.2	0.905	490.3	0.910	490.6	0.894
	3.0	910.6	0.936	879.6	0.921	867.6	0.921
	2.0	2098.2	0.918	1974.2	0.914	1932.1	0.904
	1.0	8421.1	0.920	7947.6	0.919	7797.5	0.931
$\theta=10$ w=50	4.0	1899.0	0.913	1856.3	0.898	1838.1	0.896
	3.0	3382.0	0.900	3314.1	0.930	3279.2	0.915
	2.0	7569.6	0.924	7419.7	0.915	7425.3	0.905
	1.0	31893.1	0.920	29895.1	0.906	29560.7	0.898

Table 3.1 illustrates that as  $2d$  decreases the terminal sample size increases  $M$  at a very large rate. This is a good indication that the terminal sample size is non-decreasing. Since the sample size is calculated numerically the sample size would have to be finite. Unlike the variance problem where there is a closed form of the optimal sample size, this sampling procedure does not have that so there are no comparisons made with  $M$  and  $n^*$ . It was expected that as  $2d$  decreased that the estimated coverage probability would decrease to coverage probability bound. The simulations do not indicate this. This is probably due to the fact that there are simply too many factors to consider when implementing this process. Finally, with more information given in the initial sample size a smaller expected number of total observations will be drawn provided the initial sample size does not eclipse expected value of  $N$ .

The second simulation produces estimates for the coverage probability by calculating the percentage of times the mean fell within the confidence interval. Much like variance problem, ideally the percentages  $p$  will be just above

the predetermined coverage probability  $1-\alpha$ . The estimate of the coverage probability is the percentage of times the scale parameter is within our interval. The shape parameter was varied  $\theta=\{1,2,5,10\}$ ,  $w=\{10,5,2,1,0.5,0.1\}$ ,  $m$  fixed to be 30 and  $\lambda$  fixed to be five. The percentage of times the parameter lies within the confidence interval is an estimate of the coverage probability of the scale parameter. This is observed for predetermined 0.80, 0.85, 0.90, and 0.95 coverage probabilities.

**Table 3.2. Coverage Percentages of Mean by Interval Estimate**

		$\theta=1$	$\theta=2$	$\theta=5$	$\theta=10$
$1-\alpha$		p	p	p	p
0.80	w=10	0.999	0.906	0.819	0.807
	w=5.0	0.977	0.837	0.822	0.796
	w=2.0	0.905	0.843	0.822	0.826
	w=1.0	0.831	0.846	0.821	0.827
	w=0.5	0.825	0.836	0.816	0.837
	w=0.1	0.839	0.818	0.829	0.821
0.85	w=10	1.000	0.946	0.879	0.846
	w=5.0	0.990	0.868	0.860	0.877
	w=2.0	0.936	0.885	0.861	0.861
	w=1.0	0.886	0.874	0.877	0.868
	w=0.5	0.855	0.889	0.870	0.862
	w=0.1	0.898	0.875	0.857	0.869
0.90	w=10	1.000	0.974	0.937	0.916
	w=5.0	0.997	0.895	0.917	0.918
	w=2.0	0.957	0.907	0.926	0.914
	w=1.0	0.919	0.936	0.908	0.917
	w=0.5	0.936	0.928	0.914	0.904
	w=0.1	0.926	0.925	0.916	0.916
0.95	w=10	1.000	1.000	0.994	0.959
	w=5.0	0.999	0.999	0.938	0.955
	w=2.0	0.978	0.989	0.960	0.950
	w=1.0	0.945	0.939	0.960	0.949
	w=0.5	0.955	0.957	0.964	0.957
	w=0.1	0.966	0.951	0.966	0.962

Table 3.2 shows with varying levels of shape, the estimated coverage probability consistently surpasses the  $1-\alpha$  mark. The percentages are most notably affected by the value chosen for  $w$ . As  $w$  is chosen to be smaller the percentages near the goal coverage probability. However, one still runs the risk of producing a large  $N$  value if  $w$  is chosen to be too small. To reiterate, the goal is to reduce the number of observations while simultaneously keeping the properties given in (1.3). It was expected to see nearly 100% coverage as  $w=0.5$  and  $w=0.1$  was selected, but simulations showed otherwise. Both of these tables show that  $w$  is probably the largest contributor affecting the coverage probability. If  $w$  selected to be relatively small compared with  $\theta$  the interval estimate will always near 100% coverage. The most important piece of information to take from these two simulations is that the percentage of times the mean fell within the confidence interval is consistently larger than the  $1-\alpha$  coverage probability.

## V. Discussion

Two two-stage sampling procedures were proposed that ensured the sample mean will be accurate for a Gamma population. Throughout the paper, many examples where Gamma is used to model data were given, specifically its use of estimating times in clinical trials. There are two benefits to implementing these procedures: these methods

ensure accuracy without use of asymptotics or any other approximations as well as these methods reduce the sample size. Though the Gamma is often assumed when estimating times, it is not exclusive to that and can be assumed for anything that is right-skewed and have a positive support.

One of the major limitations with the interval estimator result is that it always assumed shape known. This is a luxury that is not always present. Section 3 gives three strong reasons why the shape known is relevant. One way of getting around this is to select an initial sample large enough to estimate the shape parameter and insert it into the result (2.1), before finding the interval estimate. This lacks the mathematical rigor given in the previous sections. At this point, it is unclear of how one is to go about developing an interval without knowing the shape parameter.

Finally, there are a number of situations that happen in practice that alter the distribution; all of which lead to possible future research. In many voluntary clinical trials the subjects drop from the trial midway through. The knowledge of the subjects before dropping out of the trial can still be useful. Modifying this procedure so that it includes censored values can be a worthwhile improvement. In addition to that, if there is a minimum amount of time a subject has to be in a trial, the distribution laid out in (1.1) may not be appropriate. Mukhopadhyay and Zacks (2007), develop a two-stage bounded variance procedure for the Exponential distribution where the parameter of interest was a linear combination of location and scale. Similarly, this can be done for the Gamma mean. It is well documented that the ratio of two Gamma distributed variables are Beta. Each of these results could be potentially extended to the Beta distribution.

## Appendix.

### A.1.

We can re-express the variance on all  $N$  observations as  $A \cdot \text{Var}(\bar{X}_N) = AE(\bar{X}_N - \theta\lambda)^2$  as

$$AE\left[\frac{m^2}{N^2}(\bar{X}_m - \theta\lambda)^2 + \frac{\theta\lambda^2}{N}\left(\frac{N-m}{N}\right)\right]$$

Recall  $m \leq N$  and  $N \geq B\bar{X}_m^2 w^{-1}$ .

Now,

$$AE\left[\frac{m^2}{N^2}(\bar{X}_m - \theta\lambda)^2\right] \leq AE\left[\frac{m}{N}(\bar{X}_m - \theta\lambda)^2\right] \leq AE\left[\frac{mw\theta}{B\bar{X}_m^2}(\bar{X}_m^2 - 2\theta\lambda\bar{X}_m + \theta^2\lambda^2)\right]$$

Also,

$$AE\left[\theta\lambda^2\left(\frac{N-m}{N^2}\right)\right] \leq AE\left[\theta\lambda^2\left(\frac{1}{N}\right)\right] \leq AE\left[\theta^2\lambda^2\left(\frac{w\theta}{B\bar{X}_m^2}\right)\right]$$

Thus, using the two inequalities above with the re-expression fact we have

$$A \cdot \text{Var}(\bar{X}_N) \leq AE\left[\frac{w\theta}{B}\left(m - \frac{2m\theta\lambda}{\bar{X}_m} + \frac{m\theta^2\lambda^2}{\bar{X}_m^2} + \frac{\theta\lambda^2}{\bar{X}_m^2}\right)\right]$$

Using the fact that  $\bar{X}_m \sim \text{Gamma}(m\theta, \lambda/m)$ , we can calculate the expectation easily, and using the fact that  $\Gamma(m\theta) = (m\theta - 1)\Gamma(m\theta - 1)$  the upperbound of the variance can be re-expressed as

$$A \cdot \text{Var}(\bar{X}_N) \leq \frac{Aw}{B}\left(\frac{2m^2\theta^2 + 2m\theta}{(m\theta - 1)(m\theta - 2)}\right)$$

If  $\theta$  is known we can ensure the expected loss is less than our variance bound  $w$  by solving the righthand side of the inequality to equal  $w$  then solving for  $B$  accordingly to obtain equation (2.2).

If  $\theta$  is unknown we again use the assumption  $m\theta \geq 3$ . There is a very obvious flaw here:  $\theta$  is assumed unknown yet we are assuming that  $m\theta \geq 3$ . However, in practice even without knowledge of  $\theta$  it is a pretty safe assumption unless one is working with a combination of extremely small sample sizes and extremely small skewed distributions. If there is a belief that the distribution might be skewed and  $\theta$  will be much less than one then that can be corrected by taking a larger initial sample size. Continuing on, since  $1/\theta \leq m/3$

$$\begin{aligned} A \cdot \text{Var}(\bar{X}_N) &\leq \frac{Aw}{B}\left(\frac{2m^2\theta^2 + 2m\theta}{(m\theta - 1)(m\theta - 2)}\right) \leq \frac{Aw}{B}\left(\frac{2m^2}{m^2 - 3m}\right) + \frac{Aw}{B}\left(\frac{2m}{m^2 - 3m}\right)\left(\frac{1}{\theta}\right) \\ &\leq \frac{Aw}{B}\left(\frac{2m^2}{m^2 - 3m}\right) + \frac{Aw}{B}\left(\frac{2m}{m^2 - 3m}\right)\left(\frac{m}{3}\right) \leq \frac{Aw8m^2}{B(3m^2 - 9m)} = \frac{Aw8m}{B(3m - 9)} \end{aligned}$$

Setting this to  $w$  and solving for  $B$  accordingly we obtain equation (2.3).

**A.2.**

Suppose that  $X_1, X_2, \dots, X_n$  are iid Exponential( $\lambda$ ). Then  $T = \sum X_i$  is a sufficient statistic for  $\lambda$ . In the Gamma pdf  $t$  and  $\lambda$  appear together as  $t/\lambda$  and, in fact the Gamma( $n, \lambda$ ) pdf  $(\Gamma(n)\lambda^n)^{-1} t^{n-1} e^{-t/\lambda}$  is a scale family. Thus if  $Q(T, \lambda) = 2T/\lambda$ , then

$$Q(T, \lambda) \sim \text{Gamma}(n, 2),$$

which does not depend on  $\lambda$ . The quantity  $Q(T, \lambda)$  is a pivot with a Chi-Square distribution with  $2n$  degrees of freedom, Casella and Berger (pp. 428).

Similarly, if  $X_1, X_2, \dots, X_n$  are iid Gamma( $\theta, \lambda$ ), then  $Q(T, \lambda) = 2X/\lambda \sim \text{Chi-Square}(2n\theta)$ .

**A.3.**

The proof in A.1. ensures that  $AE[\bar{X}_N^2 - \theta\lambda]^2 \leq w$ . This means that

$$\lambda \leq \sqrt{\frac{Nw}{A\theta}}.$$

Denote  $g_q^*$  be the  $q$ th quantile of the Gamma( $n\theta, \lambda/n$ ). The sampling distribution of  $\bar{X}_n \sim \text{Gamma}(n\theta, \lambda/n)$ ,

$$P(g_{\alpha/2}^* < \bar{X} < g_{1-\alpha/2}^*) = 1 - \alpha$$

Let  $g_q$  be the  $q$ th quantile of the Gamma( $n\theta, 1/n$ ). distribution. Due to the scale property of a Gamma population;

$$P(\lambda g_{\alpha/2} < \bar{X} < \lambda g_{1-\alpha/2}) = 1 - \alpha$$

$$P(\lambda[g_{\alpha/2} - \theta] < \bar{X} - \theta\lambda < \lambda[g_{1-\alpha/2} - \theta]) = 1 - \alpha$$

$$P(-\bar{X} + \lambda[g_{\alpha/2} - \theta] < -\theta\lambda < -\bar{X} + \lambda[g_{1-\alpha/2} - \theta]) = 1 - \alpha$$

$$P(\bar{X} - \lambda[g_{1-\alpha/2} - \theta] < \theta\lambda < \bar{X} - \lambda[g_{\alpha/2} - \theta]) = 1 - \alpha$$

$$P(\bar{X} - \sqrt{Nw(\theta A)^{-1}}[g_{1-\alpha/2} - \theta] < \theta\lambda < \bar{X} - \sqrt{Nw(\theta A)^{-1}}[g_{\alpha/2} - \theta]) \geq 1 - \alpha$$

The width of this confidence interval is set to width  $2d$ , and  $M$  is found accordingly. No close form solution exists, but the numeric solution yields the interval estimator given in (3.1).

## References

- Birnbaum, A. and Healy, W. C., Jr. (1960). *Estimates with Prescribed Variance Based on Two-Stage Sampling*, Annals of Mathematical Statistics, Vol 31, pp. 662-676.
- Chatelian, F et al. (2007). *Bivariate Gamma Distributions for Image Registration and Change Detection*. IEEE Transactions on Image Processing. Vol 16, No 7, pp.1796-1806.
- Chatelian, F et al. (2008). *Change Detection in Multisensor SAR Images Using Bivariate Gamma Distributions*. IEEE Transactions on Image Processing. Vol 17, No 3, pp. 249-258.
- Ghosh, B.K. and Sen, P.K. (1991). *Handbook on Sequential Analysis*. Marcel Dekker: New York, NY.
- Govindarajulu, Z. (1995). *Sequential Point and Interval Estimation of Scale Parameter of an Exponential Distribution*. International J. Math. and Sci. Vol 18, No 2, pp. 383-390.
- Gutowksi, W.J et al. (2007). *A Possible Constraint on Regional Precipitation Intensity Under Global Warming, Journal of Hydrometeorology*. Vol 8, pp.1382-1392.
- Kubokawa, T. (1989). *Improving on Two-Stage Estimators for Scale Families*, Metrika, Vol 36, pp. 7-13.
- Maureil, A. et al. (2007). *Impacts of Climate Change on the Frequency and Severity of Floods in the Chateauguay River Basin, Canada*. Can J Civ Eng. Vol 34, pp. 1048-1059.
- Mukhopadhyay, N., Silva, B.M., and Waikar, V. (2006). *On a New Two-Stage Confidence Interval Procedure and Comparisons with Its Competitors for Estimating the Difference of Normal Means*.
- Mukhopadhyay, N. and Pepe, W. (2006). *Exact Bounded Risk Estimation When the Terminal Sample Size and Estimator Are Dependent: The Exponential Case*, Sequential Analysis, Vol 25, No 1, pp. 85 - 101.
- Mukhopadhyay, N. and Duggan, W. T. (1999). *On a Two-Stage Procedure Having Second-Order Properties with Applications*. Annals Inst. Statistical Mathematics. Vol 51, No 4, pp. 621-636.
- Stein, C (1945). *Two-sample Test of a Linear Hypothesis Whose Power is Independent of the Variance*. Ann Math Statistics. Vol 16, pp. 243-258.
- Stein, C. (1949). *Some Problems in Sequential Estimation*, Econometrica. Vol 17, pp. 77-78.
- Wald, A (1947) *.Sequential Analysis*. Wiley: New York.
- Zacks, S. and Mukhopadhyay, N. (2006a). *Exact Risks of Sequential Point Estimators of the Exponential Parameter*, Sequential Analysis, Vol 25, No 2, pp. 203- 226.
- Zacks, S. and Mukhopadhyay, N. (2006b). *Bounded Risk Estimation of the Exponential Parameter in a Two-Stage Sampling*, Sequential Analysis. Vol 25, No 4, pp. 437 - 452.
- Zacks, S. and Mukhopadhyay, N. (2007). *Bounded Risk Estimation of Linear Combinations of the Location and Scale Parameters in Exponential Distributions under Two-Stage Sampling*. Journal of Statistical Planning and Inference. Vol 137, pp. 3672 – 3686.