Vidyodaya J. of Sci. (2010) Vol. 15 No.1 & 2, pp 37-43

Confidence intervals for the median of a gamma distribution

B.M.S.G. Banneheka
Dept. of Statistics and Computer Science,
University of Sri Jayewardenepura,
Nugegoda, Sri Lanka
email: banneheka@dscs.sjp.ac.lk

Received on : 08 - 05 - 2010
Accepted no : 10 - 06 - 2010

Abstract

The gamma distribution is often used as a model for positively skewed distributions. The median is better than the mean as the representative of the 'average' in such situations. Literature is available for inference concerning the mean of a gamma distribution, but the literature concerning the median of a gamma distribution is rare.

In this paper we present a method for constructing confidence intervals for the median of a gamma distribution. The method involves inverting the likelihood ratio test to obtain 'large sample' confidence intervals. A difficulty arises as it is not possible to write the likelihood function in terms of the median. In this paper we propose a method to avoid this difficulty. The method works well even for moderately large sample sizes. The methodology is illustrated using an example.

Keywords and phrases: positively skewed, likelihood ratio test, large sample theory

Introduction

Consider the problem of estimating the 'average' of a positively skewed distribution. The gamma distribution is often used as a model in such situations and the median is considered as better than the mean as the representative of the 'average.' Then the problem becomes estimating the median ν of a gamma distribution. Ekanayake (2005) has studied the problem of making inference concerning the median of a gamma distribution, for her M.Sc. dissertation.

The sample median and maximum-likelihood estimator are two possible point estimators. Banneheka and Ekanayake (2009) compare these two estimators with a new estimator

$$\hat{\upsilon} = \frac{(3\hat{\alpha} - 0.8)\bar{x}}{(3\hat{\alpha} + 0.2)},$$

where $\bar{x}$ is the sample mean and $\hat{\alpha} = \bar{x}^2/(\frac{\sum_{i=1}^{n} x_i^2}{n} - \bar{x}^2)$. Considering the easiness of calculation and the size of mean square error, they recommend using the new estimator.

One method used to construct confidence intervals is called the 'pivotal quantity method' (Mood A.M., et.al., 2001, pg. 379). In order to apply this method to construct a confidence interval for ν, one must find a pivotal quantity with only ν in it as a parameter. In addition, the distribution of the pivotal quantity must be known. Such a pivotal quantity exists when the shape parameter $\alpha = 1$. However, it is not possible to find a pivotal quantity when $\alpha \neq 1$.

Another method used to construct a confidence interval for a parameter θ is to invert the likelihood ratio test for $H_0: \theta = \theta_0$ (a known value) vs. $H_1: \theta \neq \theta_0$. According to the standard theory, twice the log-likelihood ratio has a χ^2 distribution under H_0, for large samples. The degree of freedom of the χ^2 distribution is equal to the difference between the dimension of the unrestricted parameter space and the dimension of the parameter space restricted by H_0. The set of θ_0 values for which H_0 is not rejected at a given significance level γ gives a $(1 - \gamma)100\%$ confidence interval for θ.

The density function of a gamma distribution with shape parameter α and scale parameter β is written as

$$f_X(x; \alpha, \beta) = \frac{e^{-x/\beta} x^{\alpha-1}}{\Gamma(\alpha)\beta^{\alpha}}; \; x > 0, \alpha > 0, \beta > 0 \quad (1)$$

Then, the log-likelihood function for a random sample $x_1, x_2, \ldots, x_n$ from the above distribution is

$$l(\alpha, \beta) = \frac{\sum_{i=1}^{n} x_i}{\beta} + (\alpha - 1)\sum_{i=1}^{n} \ln x_i - n \ln \Gamma(\alpha) - n\alpha \ln \beta. \quad (2)$$

According to the large sample theory, when H_0 is true and n is large,

$$T = 2\left[l\left(\hat{\alpha}_{mle}, \hat{\beta}_{mle}\right) - l\left(\hat{\alpha}_{mle}(v_0), \hat{\beta}_{mle}(v_0)\right)\right] \tag{3}$$

has a χ_1^2 distribution. Here, $\hat{\alpha}_{mle}$ and $\hat{\beta}_{mle}$ are the global maximum likelihood estimates of α and β respectively, while $\hat{\alpha}_{mle}(v_0)$ and $\hat{\beta}_{mle}(v_0)$ are respectively the maximum likelihood estimates of α and β under H_0. Theoretically, The set $\{v_0 : T < \chi^2_{\gamma,1}\}$ should be a $(1-\gamma)100\%$ confidence interval for v.

In order to calculate $l\left(\hat{\alpha}_{mle}(v_0), \hat{\beta}_{mle}(v_0)\right)$, we should be able to write α, β in terms of v. Unfortunately, this is not possible. Therefore, T cannot be calculated and hence the above procedure cannot be implemented. We use an approximation given in Ekanayake (2005) to overcome this difficulty. That is to use the approximate relationship

$$v \approx \mu \frac{(3\alpha - 0.8)}{(3\alpha + 0.2)} \tag{4}$$

to obtain an approximate likelihood function in terms of v and v and use it in place of the exact likelihood function. In order to calculate the maximum of this function under H_0 and H_1, Ekanayake (2005) has used profile likelihood method. This method involves a considerable amount of computing. In this paper, we use a different parameterization and provide an alternative method which is computationally simpler.

Method

Let μ be the mean of the distribution. Then,

$$\mu = \propto \beta \tag{5}$$

From (4) and (5),

$$\alpha \approx \frac{(0.2v + 0.8\mu)}{3(\mu - v)} \tag{6}$$

and

$$\beta \approx \frac{3(\mu - v)\mu}{(0.2v + 0.8\mu)} \tag{7}$$

Let $l^*(\mu, v)$ be the function that we get when α and β in (2) are replaced by the approximate values given by (6) and (7) respectively. That is, let

$$l^*(\mu, v) = -\frac{(0.2v + 0.8\mu)}{3(\mu - v)\mu} \sum_{i=1}^{n} x_i + \left[\frac{(0.2v + 0.8\mu)}{3(\mu - v)} - 1\right] \sum_{i=1}^{n} ln\, x_i$$

$$-n\ln\left\{\Gamma\left[\frac{(0.2v+0.8\mu)}{3(\mu-v)}\right]\right\}-n\frac{(0.2v+0.8\mu)}{3(\mu-v)}ln\left[\frac{3(\mu-v)\mu}{(0.2v+0.8\mu)}\right]. \tag{8}$$

We propose to use the 'adjusted likelihood ratio test statistic'

$$T^* = 2[l(\hat{\alpha}_{mle}, \hat{\beta}_{mle}) - l^*(\hat{\mu}(v_0), v_0)\ ] \tag{9}$$

in place of the likelihood ratio test statistic T given by (3). Here, $l(\hat{\alpha}_{mle}, \hat{\beta}_{mle})$ is the global maximum of $l(\alpha,\beta)$ while $l^*(\hat{\mu}(v_0), v_0)$ is the maximum of $l^*(\mu, v)$ under H_0.

Steps of the procedure

1. The global maximum likelihood estimates $\hat{\alpha}_{mle}, \hat{\beta}_{mle}$ are found using the iterative numerical method mentioned in Anita S. et.al (2002). Some statistical packages (e.g. Minitab) also provide these estimates. Therefore, it is not difficult to calculate $\hat{\alpha}_{mle}, \hat{\beta}_{mle}$.
2. The global maximum log-likelihood is calculated by substituting $\hat{\alpha}_{mle}, \hat{\beta}_{mle}$ values in $l(\alpha,\beta)$.
3. For a fixed value of v_0, $l^*(\hat{\mu}(v_0), v_0)$ is found by numerically maximizing $l^*(\mu, v_0)$.
4. T^* is calculated using (9). If $T^* < \chi^2_{\gamma,1}$ then v_0 is taken as a value in the confidence interval.
5. Steps 3 and 4 are repeated for different values of v_0 and the set of v_0 values such that $T^* < \chi^2_{\gamma,1}$ is determined.
6. The resulting set is taken as the *(1* $-\ \gamma$ *)100%* confidence interval for v

Adequacy

As a measure of the adequacy of the above procedure, we compare the nominal and estimated confidence coefficients for several combinations of α and β. In each case, the actual confidence coefficient is estimated using 1000 Monte Carlo simulations. Results are shown in Table 1.

Table 1. Observed confidence coefficients, each based on 1000 Monte Carlo simulations

		Nominal confidence coefficient		
		0.90	0.95	0.99
∝	n	Observed confidence coefficient		
1	20	0.807	0.940	0.979
	25	0.874	0.944	0.982
	30	0.891	0.951	0.988
2	20	0.865	0.937	0.985
	25	0.888	0.955	0.993
	30	0.891	0.945	0.985
4	20	0.876	0.941	0.978
	25	0.894	0.945	0.984
	30	0.904	0.949	0.995
8	20	0.884	0.935	0.984
	25	0.885	0.949	0.984
	30	0.896	0.951	0.988

These results were generated with G(8,1) data. However, the results do not depend on the value of β

Illustration

Table 2 contains the Minitab data set named 'Tiles.MTW'. (This is a data set that Minitab uses in Process Capability examples in on-line Help). These data are the warping in 10 floor tiles each working day for 10 days. Suppose one wants to construct a confidence interval for the median warping of tiles.

Table 2. Warpings of Tiles. Source: Minitab

1.60103	2.31426	0.52829	2.89530	0.44426	5.31230	1.22095	4.24464	1.12465	0.28186
0.84326	2.55635	1.01497	2.86853	2.48648	1.92282	6.32858	3.21267	0.78193	0.57069
3.00679	4.72347	1.12573	2.18607	3.91413	1.22586	3.80076	3.48115	4.14333	0.70532
1.29923	1.75362	2.56891	1.05339	2.28159	0.76149	4.22622	6.66919	5.30071	2.84843
2.24237	1.62502	4.23217	1.2556	0.96705	2.39930	4.33233	2.44223	3.79701	6.25825
2.63579	5.63857	1.34943	1.97268	4.98517	4.96089	0.42845	3.51246	3.24770	3.37523
0.34093	4.64351	2.84684	0.84401	5.79428	1.96775	1.20410	8.03245	5.04867	3.23538
6.96534	3.95409	0.76492	3.32894	2.52868	1.35006	3.44007	1.13819	3.06800	6.08121
3.46645	4.38904	2.78092	4.15431	3.08283	4.79076	2.51274	4.27913	2.45252	1.66735
1.41079	3.24065	0.63771	2.57873	3.82585	2.20538	8.09064	2.05914	4.69474	2.12262

The histogram of these data, shown in Figure 1, is positively skewed. The gamma distribution is a suitable model for the warping of tiles (p-value of the

Anderson-Darling goodness of fit test is 0.238). Therefore we use the above method to construct a 95% confidence interval for the median.

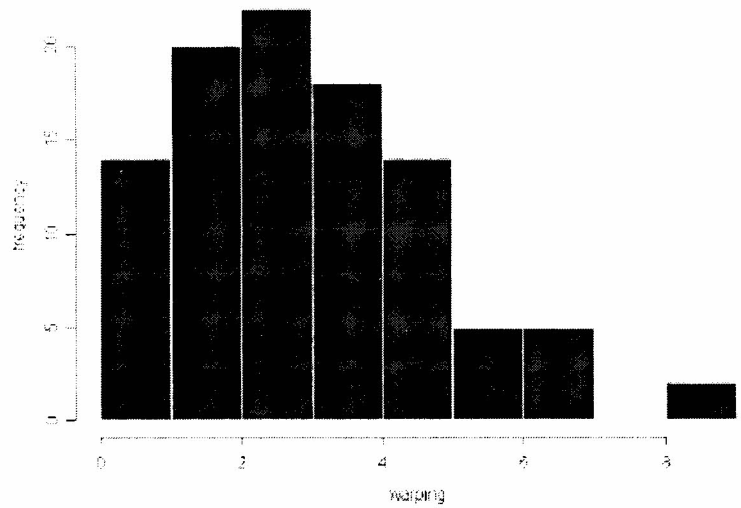

Figure 1. Histogram of warping data

Figure 2 shows the graph of T^* against ν_0. T^* is less than $\chi^2_{0.05,1}$=3.841 when $\nu_0 \in (2.19, 2.88)$. Therefore, (2.19, 2.88) is a 95% confidence interval for the median warping of tiles.

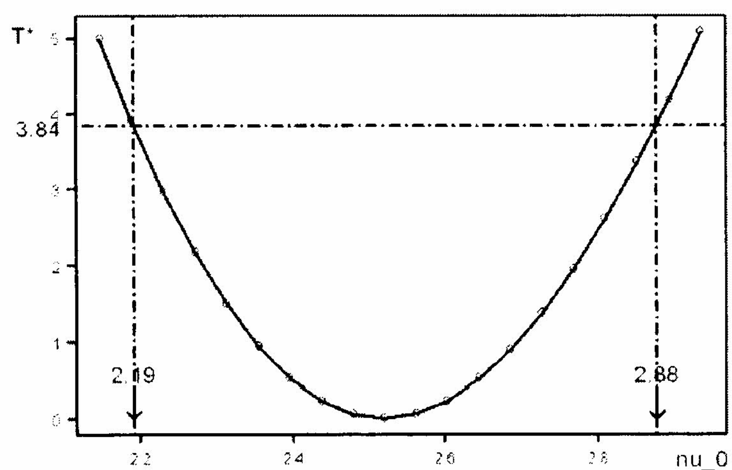

Figure 2. Calculation of 95% confidence interval for the median warping of tiles

Conclusions

Our method of inverting the adjusted likelihood ratio test to construct confidence intervals for the median ν of a gamma distribution works reasonably well even with moderately large samples.

References

Anita Singh, Ashok K. Singh, and Ross J. Iaci. Estimation of the Exposure Point Concentration Term Using a Gamma Distribution, 2002. EPA Technology Support Center Issue, United States Environmental Protection Agency. Available at:
http://www.hanford.gov/dqo/training/289cmb02.pdf

Banneheka, B.M.S.G. and Ekanayake, G.E.M.U.P.D. A new point estimator for the median of gamma distribution. Vidyodaya J. of Sci.. (2009) Vol. 14, pp 95-103.

Ekanayake, G.E.M.U.P.D. Inference concerning the median of a Gamma distribution (2005). M.Sc. dissertation submitted to the university of Sri Jayewardenepura as a partial fulfillment of M.Sc. degree in Applied Statistics.

Mood, A.M., Graybill, F., Boes, D.C. Introduction to the theory of Statistics (2001). Tata McGraw Hill Publishing Company Limited, New Delhi.