Uniformly Minimum-Variance Unbiased Estimator (UMVUE) for the Gamma Cumulative Distribution Function with Known and Integer Scale Parameter

Jessica Kubrusly

Institute of Mathematics and Statistics, Fluminense Federal University (UFF), Niterói, Brazil
Email: jessicakubrusly@id.uff.br

Abstract

Uniformly minimum-variance unbiased estimator (UMVUE) for the gamma cumulative distribution function with known and integer scale parameter. This paper applies Rao-Blackwell and Lehmann-Scheffeé Theorems to deduce the uniformly minimum-variance unbiased estimator (UMVUE) for the gamma cumulative distribution function with known and integer scale parameters. The paper closes with an example comparing the empirical distribution function with the UMVUE estimates.

Keywords

UMVUE, Cumulative Distribution Estimates, Gamma Distribution, Erlang Distribution, Lehmann-Scheffeé Theorem, Rao-Blackwell Theorem

1. Introduction

Statistical inference is an important topic in scientific studies, whether from the theoretical or applied aspects. Studying efficient estimators for the probability density function (PDF) or for the cumulative distribution function (CDF) can be useful for various applications, such as the estimation of Fisher information or the estimation of quantiles. Other applications are mentioned in [1].

Some studies on the estimation of CDF have appeared in recent literature for some continuous distributions, for instance, Pareto-Rayleigh distribution [2], Exponentiated Burr XII distribution [1] or general distributions [3].

The purpose of this paper is to present the Uniformly minimum-variance unbiased estimator (UMVUE) for the gamma cumulative distribution function with

known and integer scale parameter.

The Gamma distribution is a member of the two-parameter family of continuous probability distributions. There are two different parameterizations currently used. In this article we focus on the shape (k) and scale (λ) parameters, so $X \sim Gamma(k, \lambda)$ and the probability density function of X is defined by

$$f_X(x) = \frac{\lambda^k}{\Gamma(k)} x^{k-1} e^{-\lambda x}, \; x > 0.$$

If k is a positive integer, $\Gamma(k) = (k-1)!$, such a distribution is called an Erlang distribution and represents the sum of k independent exponential random variables, each of which has mean equal to $1/\lambda$.

Let $X_1, X_2, \cdots, X_n$ be a random sample of size n from the population X. The most common way to estimate a cumulative distribution function of X, F, is from the empirical distribution function defined by

$$\hat{F}_e(q) = \frac{\text{number of elements in the sample} \leq q}{n} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{X_i \leq q}$$

where $\mathbf{1}_A$ is the indicator of event A. This paper presents another estimator for F when k is known as an integer parameter, which is referred to as the minimum-variance unbiased estimator (UMVUE).

The paper is organized as follows. Section 2 presents two lemmas for the main result. Section 3 presents and demonstrates the main result of this paper. To finish, Section 4 presents a simple example that compares the empirical distribution function and the UMVUE estimator of $F(q)$.

2. Preliminaries

The following lemmas will be used for establishing the main results.

Lemma 1 *Let* $q, t \in \mathbb{R}$, $q < t$, $a \in \mathbb{R}^+$ *and* $b \in \mathbb{N}^+$. *Then,*

$$\int_q^t (t-x)^a x^b dx = a! b! \sum_{j=1}^{b+1} \frac{(t-q)^{a+j} q^{b-j+1}}{(a+j)!(b-j+1)!}.$$

Proof. The proof goes by induction on b. Let's call $\int_q^t (t-x)^a x^b dx = h(q,t,a,b)$.

Base case: check the result for $b = 1$. That is,

$$
\begin{aligned}
h(q,t,a,1) = \int_q^t (t-x)^a x dx &= a! \sum_{j=1}^{2} \frac{(t-q)^{a+j} q^{2-j}}{(a+j)!(2-j)!} \\
&= \frac{(t-q)^{a+1} q}{a+1} + \frac{(t-q)^{a+2}}{(a+2)(a+1)} \\
&= (t-q)^{a+1} \frac{q(a+2) + t - q}{(a+1)(a+2)} \\
&= (t-q)^{a+1} \frac{q(a+1) + t}{(a+1)(a+2)}, \quad \forall q, t, a.
\end{aligned}
$$

For this, an integration by substitution will be performed.

$$h(t,q,a,1) = \int_q^t (t-x)^a \, x \, dx = -\int_{t-q}^0 y^a \, (t-y) \, dy = \int_0^{t-q} y^a \, (t-y) \, dy$$

$$= \int_0^{t-q} ty^a - y^{a+1} \, dy = \frac{ty^{a+1}}{a+1} - \frac{y^{a+2}}{a+2} \Big|_0^{t-q} = \frac{t(t-q)^{a+1}}{a+1} - \frac{(t-q)^{a+2}}{a+2}$$

$$= (t-q)^{a+1} \left(\frac{t}{a+1} - \frac{t-q}{a+2} \right) = (t-q)^{a+1} \left(\frac{t(a+2) - (t-q)(a+1)}{(a+1)(a+2)} \right)$$

$$= (t-q)^{a+1} \frac{t + q(a+1)}{(a+1)(a+2)}.$$

Inductive step: show that the result is true for $b+1$ if it is true for b. That is, assuming that

$$h(t,q,a,b) = \int_q^t (t-x)^a \, x^b \, dx$$

$$= a! \, b! \sum_{j=1}^{b+1} \frac{(t-q)^{a+j} \, q^{b-j+1}}{(a+j)!(b-j+1)!} \quad \text{(induction hypothesis)}$$

we will conclude that

$$h(t,q,a,b+1) = \int_q^t (t-x)^a \, x^{b+1} \, dx = a!(b+1)! \sum_{j=1}^{b+2} \frac{(t-q)^{a+j} \, q^{b-j+2}}{(a+j)!(b-j+2)!}.$$

To solve $\int_q^t (t-x)^a \, x^{b+1} \, dx$ an integration by parts will be done. Consider

$$u = x^{b+1} \quad \text{and} \quad dv = (t-x)^a \, dx, \text{ then } \quad du = (b+1)x^b \, dx \quad \text{and} \quad v = -\frac{(t-x)^{a+1}}{a+1}.$$

$$h(t,q,a,b+1) = \int_q^t (t-x)^a \, x^{b+1} \, dx$$

$$= -x^{b+1} \frac{(t-x)^{a+1}}{a+1} \Big|_q^t + \int_q^t \frac{(t-x)^{a+1}}{a+1}(b+1)x^b \, dx$$

$$= q^{b+1} \frac{(t-q)^{a+1}}{a+1} + \frac{b+1}{a+1} \int_q^t (t-x)^{a+1} \, x^b \, dx$$

$$= q^{b+1} \frac{(t-q)^{a+1}}{a+1} + \frac{b+1}{a+1}(a+1)! \, b!$$

$$\times \sum_{j=1}^{b+1} \frac{(t-q)^{a+1+j} \, q^{b-j+1}}{(a+1+j)!(b-j+1)!} \quad \text{(by induction hypothesis)}$$

Replace $j+1$ with l in the above summation:

$$= q^{b+1} \frac{(t-q)^{a+1}}{a+1} + a!(b+1)! \sum_{l=2}^{b+2} \frac{(t-q)^{a+l} \, q^{b-l+2}}{(a+l)!(b-l+2)!}$$

$$= q^{b+1} \frac{(t-q)^{a+1} \, a!(b+1)!}{(a+1) \, a!(b+1)!} + a!(b+1)! \sum_{l=2}^{b+2} \frac{(t-q)^{a+l} \, q^{b-l+2}}{(a+l)!(b-l+2)!}$$

$$= a!(b+1)! \left(\frac{(t-q)^{a+1} \, q^{b+1}}{(a+1)!(b+1)!} + \sum_{l=2}^{b+2} \frac{(t-q)^{a+l} \, q^{b-l+2}}{(a+l)!(b-l+2)!} \right)$$

$$= a!(b+1)! \sum_{l=1}^{b+2} \frac{(t-q)^{a+l} \, q^{b-l+2}}{(a+l)!(b-l+2)!}$$

which closes the proof by induction.

Lemma 2 *Let* $\mathbf{X} = \{X_1, \cdots, X_n\}$ *be a random sample with size n from the population* $X \sim Gamma(k, \lambda)$. *Let* $T = \sum_{i=1}^{n} X_i$. *Then,*

$$f_{T|X_1}(t \mid x_1) = \frac{\lambda^{(n-1)k}}{\Gamma((n-1)k)}(t - x_1)^{(n-1)k-1}\, e^{\lambda(t-x_1)}, \quad t > x_1.$$

Proof. First, notice that

$$f_{T|X_1}(t \mid x_1) = f_{\sum_{i=1}^{n} X_i | X_1}(t \mid x_1) = f_{\sum_{i=1}^{n} X_i}(t - x_1).$$

Set $W = \sum_{i=2}^{n} X_i$, $W \sim Gamma((n-1)k, \lambda)$ and

$$f_W(w) = \frac{\lambda^{(n-1)k}}{\Gamma((n-1)k)} w^{(n-1)k-1} e^{-\lambda w}, \quad w > 0.$$

Then,

$$f_{T|X_1}(t \mid x_1) = f_{\sum_{i=2}^{n} X_i}(t - x_1) = \frac{\lambda^{(n-1)k}}{\Gamma((n-1)k)}(t - x_1)^{(n-1)k-1}\, e^{-\lambda(t-x_1)}, \, t - x_1 > 0.$$

3. Main Result

Theorem 1 *Let* $\mathbf{X} = \{X_1, X_2, \cdots, X_n\}$ *be a random sample with size n from the population* $X \sim Gamma(k, \lambda)$. *Consider* $k \in \mathbb{N}$ *a known parameter. For any* $q \in \mathbb{R}^+$ *let* $F_X(q) = P(X \le q) = \theta$ *be an unknown parameter. The Uniform Minimum Variance Unbiased Estimator* (*UMVUE*) *for* θ *is given by*

$$\hat{\theta} = 1 - \left(\frac{\sum_{i=1}^{n} X_i - q}{\sum_{i=1}^{n} X_i}\right)^{nk-1} \sum_{j=1}^{k} \binom{nk-1}{k-j} \left(\frac{q}{\sum_{i=1}^{n} X_i - q}\right)^{k-j}, \quad if \ q < \sum_{i=1}^{n} X_i.$$

If $q \ge \sum_{i=1}^{n} X_i$, $\hat{\theta} = 1$.

Proof. Rao-Blackwell Theorem [4] [5] states that if $\tilde{\theta}$ is an unbiased estimator for θ and $T(\mathbf{X})$ is a sufficient statistic for θ, then $\hat{\theta} = E(\tilde{\theta} | T(\mathbf{X}))$ is an unbiased estimator for θ based on $T(\mathbf{X})$ and $Var(\hat{\theta}) \le Var(\tilde{\theta})$. Lehmann-Scheffé Theorem [6] [7] states that if $\tilde{\theta}$ is an unbiased estimator for θ and $T(\mathbf{X})$ is a complete sufficient statistic for θ, then $\hat{\theta} = E(\tilde{\theta} | T(\mathbf{X}))$ is the (unique) uniformly minimum-variance unbiased estimator (UMVUE) for θ. Let us find some unbiased estimator and a complete sufficient statistic for θ to deduct its UMVUE.

First note that $\tilde{\theta}$ defined below is an unbiased estimator for θ.

$$\tilde{\theta} = \begin{cases} 1, & X_1 \le q \\ 0, & X_1 > q \end{cases}$$

Also note that $X \sim Gamma(k, \lambda)$, where k is a known parameter, belong to an uniparametric exponential family. It is known that for the exponential family it is possible to find directly a sufficient and complete statistic [8]. In the case of X, $T(\mathbf{X}) = \sum_{i=1}^{n} X_i$ is a complete sufficient statistic. Then,

$$\hat{\theta} = E\left(\tilde{\theta} \,\middle|\, \sum_{i=1}^{n} X_i\right)$$

is the UMVUE for θ. Let's show that $\hat{\theta}$ has the expression given in the statement.

171

$$\hat{\theta} = E\left(\tilde{\theta} \,\middle|\, \sum_{i=1}^{n} X_i \right) = P\left(X_1 \le q \,\middle|\, \sum_{i=1}^{n} X_i \right) = 1 - P\left(X_1 > q \,\middle|\, \sum_{i=1}^{n} X_i \right)$$

$$= 1 - P\left(X_1 > q \,\middle|\, \sum_{i=1}^{n} X_i = t \right)\Bigg|_{t=\sum_{i=1}^{n} X_i}$$

$$= \begin{cases} 1 & , t \le q \\ 1 - \int_q^{\infty} f_{X_1|\sum_{i=1}^n X_i}(x_1 \mid t)\,\mathrm{d}x_1 & , q < t \end{cases}\Bigg|_{t=\sum_{i=1}^{n} X_i}$$

Before proceeding with the computations, we develop the expression of $f_{X_1|\sum_{i=1}^n X_i}$. For this, we used Lemma 2.

$$f_{X_1|\sum_{i=1}^n X_i}(x_1 \mid t) = \frac{f_{\sum_{i=1}^n X_i|X_1}(t \mid x_1) f_{X_1}(x_1)}{f_{\sum_{i=1}^n X_i}(t)}$$

$$= \frac{\lambda^{(n-1)k}(t-x_1)^{(n-1)k-1}\mathrm{e}^{-\lambda(t-x_1)}}{\Gamma((n-1)k)} \frac{\lambda^k x_1^{k-1}\mathrm{e}^{-\lambda x_1}}{\Gamma(k)} \frac{\Gamma(nk)}{\lambda^{nk} t^{nk-1}\mathrm{e}^{-\lambda t}}, t-x_1 > 0, x_1 > 0, t > 0.$$

$$= \frac{\Gamma(nk)}{\Gamma(nk-k)\Gamma(k)} \frac{(t-x_1)^{nk-k-1} x_1^{k-1}}{t^{nk-1}}, x_1 < t, x_1 > 0, t > 0.$$

Assuming $q < t$, and using Lemma 1,

$$\hat{\theta} = 1 - \int_q^{\infty} f_{X_1|\sum_{i=1}^n X_i}(x_1 \mid t)\,\mathrm{d}x_1 \Bigg|_{t=\sum_{i=1}^{n} X_i}$$

$$= 1 - \int_q^{t} \frac{\Gamma(nk)}{\Gamma(nk-k)\Gamma(k)} \frac{(t-x_1)^{nk-k-1} x_1^{k-1}}{t^{nk-1}}\,\mathrm{d}x_1 \Bigg|_{t=\sum_{i=1}^{n} X_i}$$

$$= 1 - \frac{\Gamma(nk)}{\Gamma(nk-k)\Gamma(k) t^{nk-1}} \int_q^{t} (t-x_1)^{nk-k-1} x_1^{k-1}\,\mathrm{d}x_1 \Bigg|_{t=\sum_{i=1}^{n} X_i}$$

$$= 1 - \frac{\Gamma(nk)}{\Gamma(nk-k)\Gamma(k) t^{nk-1}} (nk-k-1)!(k-1)!$$

$$\times \sum_{j=1}^{(k-1)+1} \frac{(t-q)^{(nk-k-1)+j} q^{(k-1)-j+1}}{((nk-k-1)+j)!((k-1)-j+1)!}\Bigg|_{t=\sum_{i=1}^{n} X_i}$$

$$= 1 - \frac{\Gamma(nk)}{t^{nk-1}} \sum_{j=1}^{k} \frac{(t-q)^{nk-k-1+j} q^{k-j}}{\Gamma(nk-k+j)\Gamma(k-j+1)}\Bigg|_{t=\sum_{i=1}^{n} X_i}$$

$$= 1 - \left(\frac{t-q}{t}\right)^{nk-1} \sum_{j=1}^{k} \frac{\Gamma(nk)}{\Gamma(nk-k+j)\Gamma(k-j+1)}\left(\frac{q}{t-q}\right)^{k-j}\Bigg|_{t=\sum_{i=1}^{n} X_i}$$

$$= 1 - \left(\frac{\sum_{i=1}^{n} X_i - q}{\sum_{i=1}^{n} X_i}\right)^{nk-1} \sum_{j=1}^{k} \frac{(nk-1)!}{(nk-k+j-1)!(k-j)!}\left(\frac{q}{\sum_{i=1}^{n} X_i - q}\right)^{k-j}$$

$$= 1 - \left(\frac{\sum_{i=1}^{n} X_i - q}{\sum_{i=1}^{n} X_i}\right)^{nk-1} \sum_{j=1}^{k} \binom{nk-1}{k-j}\left(\frac{q}{\sum_{i=1}^{n} X_i - q}\right)^{k-j}.$$

4. A Very Simple Example

Consider the population $X \sim Gamma(3, \lambda)$ and the random sample of size 5:

$$2.049282 \qquad 2.429458 \qquad 1.288630 \qquad 3.967066 \qquad 3.527220$$

which was generated in the R Program [9] by the command

$$\text{set.seed(1);rgamma(5,3,1).}$$

For this sample, $\sum_{i=1}^{5} x_i = 13.26166$ and the uniformly minimum-variance unbiased estimator for the gamma cumulative distribution function with scale parameter $k = 3$ is

$$\hat{\theta} = \hat{F}_X(q) = 1 - \left(\frac{\sum_{i=1}^{5} X_i - q}{\sum_{i=1}^{5} X_i} \right)^{14} \sum_{j=1}^{3} \binom{5 \times 3 - 1}{3 - j} \left(\frac{q}{\sum_{i=1}^{5} X_i - q} \right)^{3-j}$$

$$= 1 - \left(\frac{13.26166 - q}{13.26166} \right)^{14} \left(91 \left(\frac{q}{13.26166 - q} \right)^{2} + 14 \left(\frac{q}{13.26166 - q} \right) + 1 \right)$$

$$= 1 - 91 \left(\frac{q}{13.26166 - q} \right)^{16} + 14 \left(\frac{q}{13.26166 - q} \right)^{15} + \left(\frac{q}{13.26166 - q} \right)^{14}$$
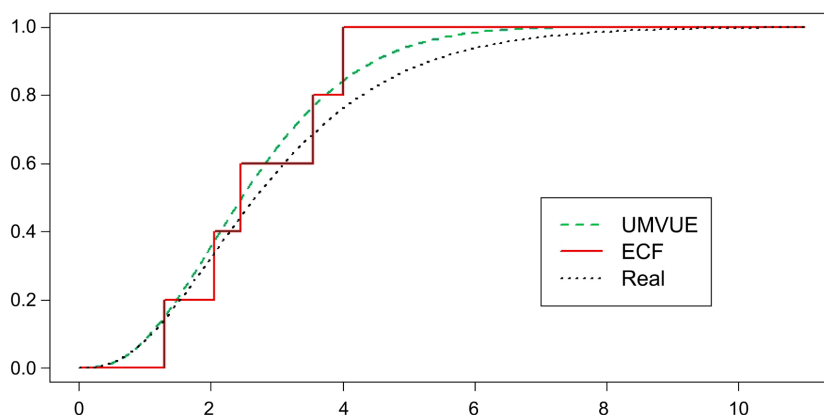

Figure 1. Gamma cumulative distribution function—approximate and exact curves.

In **Figure 1** can be compared the curves of the uniformly minimum-variance unbiased estimator for the gamma cumulative distribution function $\hat{F}$ (UMVUE) and the empirical cumulative function (ECF), both created from the random samples presented above. The black dotted line represents the exact curve (Real), considering $\lambda = 1$, parameter used to generate the sample.

5. Conclusions

In the previous sections, we argue about the construction of the UMVUE for the gamma cumulative distribution function with known and integer scale parameter. This is a pontual estimator for the $P(X \leq q)$, for $q \in \mathbb{R}$, where $X \sim Gama(k, \lambda)$ and $k \in \mathbb{N}$ is a well-known parameter.

The advantage of using the UMVUE estimator is that besides being the uniformly minimum-variance unbiased estimator it is a continuous estimator with respect to q. Moreover, with small samples, the results are already satisfactory, which can be seen in Section 4. The disadvantage of the present approach is that

for large values of k the estimator will have a complex expression and if $n \times k$ is too large, it may not be simple to compute $\begin{pmatrix} nk-1 \\ k-j \end{pmatrix}$.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

[1] Hassan, A.S., Assar, S.M., Ali, K.A. and Nagy, H.F., *et al.* (2021) Estimation of the Density and Cumulative Distribution Functions of the Exponentiated Burr XII Distribution. *Statistics*, **22**, 171-189. https://doi.org/10.21307/stattrans-2021-044

[2] Jebeli, M. and Deiri, E. (2020) Estimation Methods for the Probability Density Function and the Cumulative Distribution Function of the Pareto-Rayleigh Distribution. *Statistics*, **54**, 135-151. https://doi.org/10.1080/02331888.2019.1689979

[3] Zamanzade, E., Mahdizadeh, M. and Samawi, H.M. (2020) Efficient Estimation of Cumulative Distribution Function Using Moving Extreme Ranked Set Sampling with Application to Reliability. *AStA Advances in Statistical Analysis*, **104**, 485-502. https://doi.org/10.1007/s10182-020-00368-3

[4] Rao, C. (1945) Information and Accuracy Attainable in Estimation of Statistical Parameters. *Bulletin of the Calcutta Mathematical Society*, **37**, 81-91.

[5] Blackwell, D. (1947) Conditional Expectation and Unbiased Sequential Estimation. *The Annals of Mathematical Statistics*, **18**, 105-110. https://doi.org/10.1214/aoms/1177730497

[6] Lehmann, E.L. and Scheffe, H. (1950) Completeness, Similar Regions, and Unbiased Estimation: Part I. *Sankhyā: The Indian Journal of Statistics* (1933-1960), **10**, 305-340.

[7] Lehmann, E.L. and Scheffe, H. (1955) Completeness, Similar Regions, and Unbiased Estimation: Part II. *Sankhyā: The Indian Journal of Statistics* (1933-1960), **15**, 219-236.

[8] Casella, G. and Berger, R.L. (2021) Statistical Inference. Cengage Learning.

[9] R Core Team (2020) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna.