Communications for Statistical Applications and Methods
2020, Vol. 27, No. 6, 625–639

https://doi.org/10.29220/CSAM.2020.27.6.625
Print ISSN 2287-7843 / Online ISSN 2383-4757

Fixed-accuracy confidence interval estimation of $P(X > c)$ for a two-parameter gamma population

Yan Zhuang[a], Jun Hu[1,b], Yixuan Zou[c]

[a]Department of Mathematics and Statistics, Connecticut College, USA;
[b]Department of Mathematics and Statistics, Oakland University, USA;
[c]Department of Statistics, University of Kentucky, USA

Abstract

The gamma distribution is a flexible right-skewed distribution widely used in many areas, and it is of great interest to estimate the probability of a random variable exceeding a specified value in survival and reliability analysis. Therefore, the study develops a fixed-accuracy confidence interval for $P(X > c)$ when X follows a gamma distribution, $\Gamma(\alpha,\beta)$, and c is a preassigned positive constant through: 1) a purely sequential procedure with known shape parameter α and unknown rate parameter β; and 2) a nonparametric purely sequential procedure with both shape and rate parameters unknown. Both procedures enjoy appealing asymptotic first-order efficiency and asymptotic consistency properties. Extensive simulations validate the theoretical findings. Three real-life data examples from health studies and steel manufacturing study are discussed to illustrate the practical applicability of both procedures.

Keywords: fixed-accuracy confidence interval, gamma distribution, sequential procedure

1. Introduction

As one of the most important distributions in probability and statistics, the gamma distribution is a flexible right-skewed distribution widely used in many areas such as reliability, environment, insurance and medicine. Burgin (1975) justified the applicability of the gamma distribution to inventory control. Vaz and Fortes (1988) discussed fitting a gamma distribution for grain sizes in a poly-crystal. Husak *et al.* (2007) used the gamma distribution to model rainfall data. Most recently, Johnson and Kliche (2020) compared seven estimation procedures for gamma parameters on raindrop size data. Since Ohlsson and Johansson (2010), the gamma distribution has become more or less the standard option in modeling claim cost in the insurance industry.

Many researchers have worked on the estimation of the parameters in two-parameter gamma distributions. Choi and Wette (1969) examined the numerical technique of the maximum likelihood method to estimate both parameters of a gamma distribution. Chen and Mi (1998) discussed the point estimation for the scale parameter of a gamma distribution based on grouped data, assuming that the shape parameter was known. Iliopoulos (2016) constructed exact confidence intervals for the shape parameter of a gamma distribution. The author compared the exact confidence intervals with bootstrap confidence intervals via simulation studies. Son and Oh (2006) developed a Gibbs sampling Bayesian estimator of the two-parameter gamma distribution under the non-informative prior.

[1] Corresponding author: Department of Mathematics and Statistics, Oakland University, Mathematics and Science Center, Room 367, 146 Library Drive, Rochester, MI 48309-4479, USA. E-mail: junhu@oakland.edu

Published 30 November 2020 / journal homepage: http://csam.or.kr

Most of the literature on the parameter estimation of gamma distributions is based on fixed-sample-size procedures. That is, one aims to find the estimate for parameters of interest based on the obtained data, no matter how large or small it is. In certain situations, especially when data collecting process is time-consuming or costly, it is of great importance to understand what sample size is needed to obtain the value of estimators with prescribed accuracy, which, however, depends on some unknown nuisance parameter and cannot be fixed in advance. Thus, sequential sampling becomes necessary to solve such problems. Takada and Nagata (1995) considered a sequential procedure for building a fixed-width confidence interval for the mean of a gamma distribution. Isogai and Uno (1995) developed a sequential procedure for estimating the mean of a gamma distribution under a loss function of squared error plus linear cost. Liu (2001) worked on approximating the optimal fixed sample size expected reward through a two-stage sampling procedure. Recently, Zacks and Khan (2011) developed two-stage and sequential procedures of conducting fixed-width confidence interval estimation for the scale parameter when the shape parameter is known. Mahmoudi and Roughani (2015) studied a bounded risk two-stage sampling procedure for estimating the scale parameter of a gamma distribution with the shape parameter known. Roughani and Mahmoudi (2015) derived explicit formulas for the expected value and risk of the estimator of the scale parameter in a gamma distribution, where the shape parameter was assumed known. One may achieve a broad-ranging review in the field of sequential analysis by combining selected parts of interest from the following monographs and references therein: Stein (1949), Anscombe (1952, 1953), Chow and Robbins (1965), Woodroofe (1977), Ghosh and Mukhopadhyay (1981), Siegmund (1985), Ghosh *et al.* (1997) and Mukhopadhyay and de Silva (2009).

In this paper, we will focus on constructing a fixed-accuracy confidence interval for $P(X > c)$ where X is a random variable that follows a gamma distribution, and c is a preassigned positive constant. In many applications, it is desired to estimate the probability of a random variable exceeding a specified value. For instance, in industrial hygiene, it is of interest to estimate the probability that the exposure level (level of exposure to a contaminant in a workplace) of a worker exceeds the occupational exposure limit (OEL; usually set by the Occupational Safety and Health Administration). See Krishnamoorthy and Mathew (2009). In lifetime data analysis, people are often interested in crucial events such as failure of equipment, death of a person, and development of symptoms of disease. The time to the occurrence of the event is called lifetime. There is a large volume of literature on modeling lifetime data using the gamma distribution. One is referred to Lawless (1982), Ansell and Phillips (1994), Balakrishnan and Ling (2014), Balakrishnan *et al.* (2019), etc. Moreover, the widely used exponential distribution is considered as a special case of the gamma distribution. One may see Nadarajah and Gupta (2007), Piegorsch (2015) and Dagpunar (2019) for more details. Therefore, it is essential to know the probability of a gamma random variable X exceeding a given value of interest, especially when at least one parameter is unknown. To the best of our knowledge, little literature has been focused on the confidence interval estimation of the rate parameter of a gamma distribution. And no literature exists in building a fixed-accuracy confidence interval for a function of the rate parameter (e.g., $P(X > c)$). Due to the fact that a fixed-width confidence interval is very likely to contain a negative lower bound or an upper bound being greater than 1, and it does not make sense for $P(X > c)$ to be negative or to exceed 1, the problem for estimating $P(X > c)$ will be better addressed using a fixed-accuracy confidence interval when the rate parameter is unknown.

Let us begin with a gamma random variable ($X \sim \Gamma(\alpha, \beta)$) with the associated density function:

$$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, \quad \text{for } x > 0,\ \alpha, \beta > 0. \tag{1.1}$$

Here, β, the rate parameter, is of interest and remains unknown to us. We shall point out that in some research work, authors alternatively use $\theta = 1/\beta$, the scale parameter, to characterize the gamma distribution. α is called the shape parameter, and can be either known or unknown.

The rest of the paper is organized as the following: In Section 2, we provide a purely sequential procedure to estimate $P(X > c)$ with X coming from a two-parameter gamma population with known shape parameter α and unknown rate parameter β. We discuss some appealing properties of this procedure, and we support our findings through extensive simulations. In Section 3, we develop a nonparametric purely sequential procedure to estimate $P(X > c)$ when both α and β are unknown. We also present some interesting properties of our proposed stopping rule followed by extensive simulations. Section 4 includes illustrations of the purely sequential procedures discussed in Sections 2 and 3 on three real-life data sets. Section 5 provides some concluding thoughts.

2. Sequential fixed-accuracy confidence intervals with known α

In the situation when α is known (for example, the exponential distribution is a special case of $\Gamma(\alpha, \beta)$ with $\alpha = 1$), we can express the desired probability of X exceeding a constant $c(> 0)$ with an incomplete gamma function as follows:

$$p \equiv p(\beta) = P_\beta(X > c) = P_\beta(X\beta > c\beta) = 1 - F(c\beta), \tag{2.1}$$

where $F(\cdot)$ stands for the distribution function of $\Gamma(\alpha, 1)$. Further, we define

$$q \equiv q(\beta) = \frac{p}{1-p}, \tag{2.2}$$

and q has a range of $(0, \infty)$.

Starting with a random sample, $X_1, \ldots, X_n$, from a $\Gamma(\alpha, \beta)$ population, one can easily obtain the maximum likelihood estimator (MLE) of β, say $\hat{\beta}_n$:

$$\hat{\beta}_n = \frac{\alpha}{\bar{X}_n}, \tag{2.3}$$

where $\bar{X}_n = n^{-1} \sum_{i=1}^{n} X_i$ is the sample mean. Further, using the invariant property of MLE, we have the MLEs of p and q given as follows, respectively:

$$\hat{p}_n = 1 - F\left(c\hat{\beta}_n\right), \tag{2.4}$$

and

$$\hat{q}_n = \frac{\hat{p}_n}{1-\hat{p}_n}. \tag{2.5}$$

The central limit theorem (CLT) for MLE tells us that as $n \to \infty$,

$$\sqrt{n}\left(\hat{\beta}_n - \beta\right) \xrightarrow{D} N\left(0, \alpha^{-1}\beta^2\right), \tag{2.6}$$

where $\xrightarrow{D}$ represents convergence in distribution. Applying the delta method, we have that as $n \to \infty$,

$$\sqrt{n}\left(\log \hat{q}_n - \log q\right) \xrightarrow{D} N\left(0, \sigma_\beta^2\right), \tag{2.7}$$

where "log" represents the natural logarithm, and

$$\sigma_\beta^2 = \alpha^{-1}\beta^2\left(\frac{d\log q}{d\beta}\right)^2. \tag{2.8}$$

Note that by definition

$$\log q = \log p - \log(1-p), \tag{2.9}$$

so

$$\begin{aligned}\frac{d\log q}{d\beta} &= \frac{1}{p}\frac{dp}{d\beta} - \frac{1}{1-p}\left(-\frac{dp}{d\beta}\right)\\ &= -\frac{c^\alpha\beta^{\alpha-1}e^{-c\beta}}{\Gamma(\alpha)F(c\beta)(1-F(c\beta))}.\end{aligned} \tag{2.10}$$

And we have the expression of σ_β^2:

$$\sigma_\beta^2 = \frac{c^{2\alpha}\beta^{2\alpha}e^{-2c\beta}}{\alpha\Gamma^2(\alpha)F^2(c\beta)\,(1-F(c\beta))^2}. \tag{2.11}$$

To estimate $p = P_\beta(X > c)$, we first focus on $q = p/(1-p)$, and consider a *fixed-accuracy confidence interval* for q of the form given by

$$J_{n;1} = \left(d^{-1}\hat{q}_n, d\hat{q}_n\right), \tag{2.12}$$

with $d > 1$ fixed. Obviously, $J_{n;1}$ is a subset of $\mathcal{R}^+$, and thus takes into account the positivity of the parameter q. One may refer to Mukhopadhyay and Banerjee (2014), Banerjee and Mukhopadhyay (2016), Mukhopadhyay and Zhuang (2016), Bapat (2018) and other sources for additional background information on fixed-accuracy confidence interval estimation.

With a prescribed significance level $0 < \gamma < 1$, $J_{n;1}$ should also satisfy the condition that the coverage probability should be at least $1-\gamma$ or approximately $1-\gamma$; that is,

$$P_\beta\{q \in J_{n;1}\} \geq 1-\gamma, \tag{2.13}$$

from which it follows that

$$P_\beta\left\{\sqrt{n}\left|\log\hat{q}_n - \log q\right|\sigma_\beta^{-1} \leq \sqrt{n}\sigma_\beta^{-1}\log d\right\} \geq 1-\gamma \Rightarrow n \geq \left(\frac{z}{\log d}\right)^2\sigma_\beta^2,$$

where $z \equiv z_{\gamma/2}$ stands for the upper $100(\gamma/2)$ quantile of a standard normal distribution. Now, we define the *optimal fixed sample size* to be

$$n_1^* \equiv n_1^*(d) = \left(\frac{z}{\log d}\right)^2\sigma_\beta^2. \tag{2.14}$$

The magnitude of n_1^*, unfortunately, remains unknown due to the fact that σ_β^2 depends on the unknown parameter β. Therefore, it is essential to estimate σ_β^2 by updating its estimator stage-wise as needed. Customarily, in the light of (2.3) and (2.11), one may use the MLE of σ_β^2 given by

$$\hat{\sigma}_{n;1}^2 \equiv \sigma_{\hat{\beta}_n}^2 = \frac{c^{2\alpha}\hat{\beta}_n^{2\alpha}e^{-2c\hat{\beta}_n}}{\alpha\Gamma^2(\alpha)F^2\left(c\hat{\beta}_n\right)\left(1-F\left(c\hat{\beta}_n\right)\right)^2}. \tag{2.15}$$

2.1. A purely sequential procedure

In the light of Anscombe (1953) and Chow and Robbins (1965) who established fundamental theory of sequential estimation, we propose a purely sequential procedure to deal with the problem of constructing a fixed-accuracy confidence interval for $P_\beta(X > c)$ under a gamma population when the rate parameter is unknown. The basic idea is that after a pilot sample, we draw observations one by one, and terminate sampling immediately when there are enough observations according to some predefined stopping rule.

Let

$$\mathcal{P}_1 : N_1 \equiv N_1(d) = \inf\left\{n \geq m : n \geq \left(\frac{z}{\log d}\right)^2 \hat{\sigma}^2_{n;1}\right\}, \tag{2.16}$$

where $m \geq 1$ indicates a pilot sample size, and $\hat{\sigma}^2_{n;1}$ is defined in (2.15). That is, we first take a sample of size m, and check whether $m \geq (z/\log d)^2\hat{\sigma}^2_{m;1}$ is true or not. If it is true, we do not take any extra observations, and the final sample will be the pilot sample. Otherwise, we draw one observation at-a-time, check the stopping rule (2.16) successively with renewed $\hat{\sigma}^2_{n;1}$, and stop sampling at the first time when $n \geq (z/\log d)^2\hat{\sigma}^2_{n;1}$ is observed and the terminated sample size is $N_1 = n$. We state Theorem 1 to show that the procedure $\mathcal{P}_1$ terminates w.p.1.

Theorem 1. *For the purely sequential fixed-accuracy confidence interval estimation procedure* $\mathcal{P}_1$ *given in* (2.16)*, with all fixed* α*,* β*,* γ*,* d*,* c*, and* m*, we have:*

$$P_\beta(N_1 < \infty) = 1. \tag{2.17}$$

Proof: Let $\xrightarrow{P}$ denote convergence in probability. It is well-known that $\hat{\beta}_n$ is a consistent estimator of β, and hence $\hat{\beta}_n \xrightarrow{P} \beta$ as $n \to \infty$. Note that $\hat{\sigma}^2_{n;1}$ is a continuous function of $\hat{\beta}_n$, so by Slutsky's theorem, $\hat{\sigma}^2_{n;1} \xrightarrow{P} \sigma^2_\beta$ as $n \to \infty$. In view of (2.16), it is easily seen that $P_\beta(N_1 < \infty) = 1$. □

Upon termination, with the acquired data

$$\{N_1, X_1, \ldots, X_m, \ldots, X_{N_1}\},$$

the $100(1-\gamma)\%$ fixed-accuracy confidence interval for q will be:

$$J_{N_1;1} = \left(d^{-1}\hat{q}_{N_1}, d\hat{q}_{N_1}\right) = \left(\frac{\hat{p}_{N_1}}{d(1-\hat{p}_{N_1})}, \frac{d\hat{p}_{N_1}}{1-\hat{p}_{N_1}}\right),$$

and accordingly, the confidence interval for $p = P_\beta(X > c)$ will be:

$$J^*_{N_1;1} = \left(\frac{\hat{p}_{N_1}}{d-(d-1)\hat{p}_{N_1}}, \frac{d\hat{p}_{N_1}}{1+(d-1)\hat{p}_{N_1}}\right),$$

where $\hat{p}_{N_1}$ comes from (2.4) with the fully acquired data.

Now, we are in a position to discuss the appealing asymptotic efficiency and consistency properties for this newly proposed purely sequential procedure $\mathcal{P}_1$.

Theorem 2. *For the purely sequential fixed-accuracy confidence interval estimation procedure $\mathcal{P}_1$ given in* (2.16)*, with all fixed α, β, c, γ, and m, we have:*

$$\lim_{d\to 1} E_\beta\left[\frac{N_1}{n_1^*}\right] = 1 \text{ [Asymptotic First-Order Efficiency]}, \tag{2.18}$$

$$\lim_{d\to 1} P_\beta\{q \in J_{N_1;1}\} = 1-\gamma \text{ [Asymptotic Consistency]}, \tag{2.19}$$

where n_1^* is defined in (2.14)*.*

Proof: We prove the asymptotic first-order efficiency and consistency by applying a general framework of purely sequential fixed-width confidence intervals based on MLE proposed in Yu (1989) with the stopping rule given by

$$N = \inf\left\{n \geq m : I_n\left(\hat{\theta}_n\right) \geq \lambda\right\},$$

where m is a pilot sample size, $\hat{\theta}_n$ is the MLE of a generic parameter θ, $I_n(\hat{\theta}_n)$ is the observed Fisher information, and λ is a multiplier such that $\lambda\theta$ indicates an optimal sample size.

Observe that a corresponding confidence interval can be constructed for $\log q$ in the light of $J_{n;1}$ from (2.12) for q as follows:

$$(\log \hat{q}_n - \log d, \log \hat{q}_n + \log d), \tag{2.20}$$

which is symmetric about $\log \hat{q}_n$ with a fixed width of $2\log d$. The reciprocal for the asymptotic variance of $\sqrt{n}(\log \hat{q}_n - \log q)$, given by $1/\sigma_\beta^2$, can be interpreted as the Fisher information in terms of $\log q$, and accordingly $n/\hat{\sigma}_{n;1}^2$ is the observed Fisher information obtained from n observations.

For this fixed-width confidence interval estimation problem, the optimal fixed sample size is still n_1^* defined in (2.14). Therefore, the associated stopping rule remains to be (2.16), and it would match Yu's (1989) stopping rule by noting that $I_n(\hat{\theta}_n) = n/\hat{\sigma}_{n;1}^2$ and $\lambda = (z/\log d)^2$. Along the lines of Yu (1989, Theorems 2 and 3), both (2.18) and (2.19) stand. □

2.2. Simulated performances

To investigate the performance of the purely sequential fixed-accuracy confidence interval estimation procedures $\mathcal{P}_1$ based on the stopping rule given by (2.16), we include a simulation study under the exponential case; that is, the shape parameter α is fixed and known to be 1. In order to draw eligible pseudo-random observations, we fixed $\beta = 2$, but pretended it had been unknown to us. Then, we set $c = \alpha/\beta = 0.5$ so that $P_\beta(X > c)$ to be estimated represented the probability that an observation would turn out larger than the mean value. For brevity alone, we present summaries when the pilot sample size was $m = 20$, and the significance level $\gamma = 0.05$, while a wide range of d values were chosen, including $d = 1.10, 1.15, 1.20, 1.25, 1.30, 1.35, 1.40, 1.45, 1.50$.

We summarized the following quantities in Table 1 by running 10,000 independent trials: the mean and standard deviation of the terminated sample sizes, $\bar{N}_1$ and $s(N_1)$; the ratio of $\bar{N}_1$ to the optimal fixed sample size n_1^* as well as the difference $\bar{N}_1 - n_1^*$; the average coverage probability $\overline{\text{cp}}_1$, which is calculated by checking the percentage of the obtained confidence intervals containing the true value of p out of 10,000 intervals, and the corresponding standard error $s(\overline{\text{cp}}_1)$; and the mean of estimated values of $p = P_\beta(X > c)$ with its standard error, $\bar{\hat{p}}_{N_1}$ and $s(\bar{\hat{p}}_{N_1})$.

In Table 1, $\bar{N}_1$ increases as the preassigned accuracy d decreases. $\bar{N}_1/n_1^*$ is close to 1 and it gets closer to 1 as d goes smaller. These are also shown in Figures 1 and 2. The average coverage

Table 1: Simulation results by implementing $\mathcal{P}_1$ as (2.16) using $\Gamma(1, 2)$ with $\gamma = 0.05$, $c = 0.5$, and $m = 20$ under 10,000 runs, where $p = P_\beta(X > c) = 0.3679$

d	n_1^*	$\bar{N}_1$	$s(N_1)$	$\bar{N}_1/n_1^*$	$\bar{N}_1 - n_1^*$	$\overline{\text{cp}}_1$	$s(\overline{\text{cp}}_1)$	$\hat{\bar{p}}_{N_1}$	$s(\hat{\bar{p}}_{N_1})$
1.50	58.48	59.76	6.39	1.0219	1.28	0.9525	0.0021	0.3695	0.0005
1.45	69.64	70.94	7.00	1.0187	1.30	0.9569	0.0020	0.3692	0.0004
1.40	84.92	86.21	7.74	1.0152	1.29	0.9524	0.0021	0.3691	0.0004
1.35	106.75	107.95	8.60	1.0113	1.20	0.9539	0.0021	0.3691	0.0003
1.30	139.66	140.82	9.78	1.0082	1.16	0.9558	0.0021	0.3690	0.0003
1.25	193.08	194.36	11.59	1.0066	1.28	0.9517	0.0021	0.3684	0.0003
1.20	289.21	290.28	14.11	1.0037	1.07	0.9536	0.0021	0.3685	0.0002
1.15	492.17	493.32	18.46	1.0023	1.15	0.9535	0.0021	0.3682	0.0002
1.10	1058.32	1059.17	26.94	1.0008	0.85	0.9511	0.0022	0.3682	0.0001

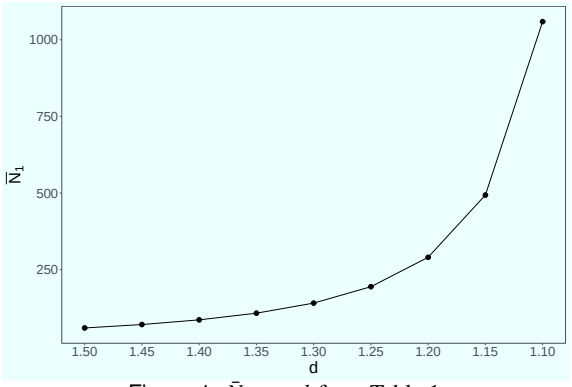

Figure 1: $\bar{N}_1$ *v.s.* d *from Table 1.*

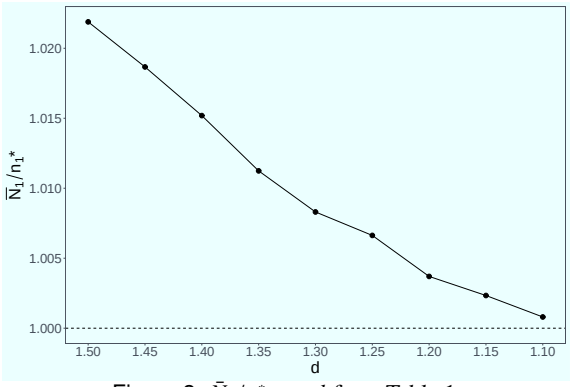

Figure 2: $\bar{N}_1/n_1^*$ *v.s.* d *from Table 1.*

probability $\overline{\text{cp}}_1$ is close to $1 - \gamma = 0.95$, and its standard error $s(\overline{\text{cp}}_1)$ is small across the board. Moreover, the differences between $\bar{N}_1$ and n_1^* are small and consistent for different values of d, which is around 1 across the board. This empirically validates that our newly proposed sampling procedure $\mathcal{P}_1$ gives a very consistent and efficient determination of the required sample size for estimation. From the last two columns, we can tell $\hat{\bar{p}}_{N_1}$ values are close to p with tiny standard errors, which suggests that our sampling procedure can provide perfect estimates for this probability of interest.

3. Sequential fixed-accuracy confidence intervals with unknown α

For a two-parameter gamma distribution $\Gamma(\alpha, \beta)$ with both α and β unknown, we can no longer implement $\mathcal{P}_1$ as per (2.16) to obtain a fixed-accuracy confidence interval for p that is defined in (2.1). Also, in this case, the MLEs of α and β have no closed-form expressions, and can only be solved numerically. In this section, we therefore develop a sequential fixed-accuracy confidence interval estimation methodology from a nonparametric perspective.

Let us define a new random variable Y based on the gamma random variable X: for a random sample of size n, $X_1, \ldots, X_n$, let $Y_i = I_{\{X_i > c\}}$ with $i = 1, \ldots, n$, where I_A is the indicator function of an event A. Clearly, $Y_1, \ldots, Y_n$ are independent and identically distributed Bernoulli random variables with probability of success being p as is defined in (2.1).

Then,

$$\tilde{p}_n = \bar{Y}_n = n^{-1} \sum_{i=1}^{n} Y_i \tag{3.1}$$

serves as an unbiased and consistent estimator of p. Furthermore, it is also the MLE of p under circumstances where both α and β are unknown in a gamma distribution. By CLT, it further holds that as $n \to \infty$,

$$\sqrt{n}\,(\tilde{p}_n - p) \xrightarrow{D} N(0, p(1-p)). \tag{3.2}$$

Similarly, we define

$$\tilde{q}_n = \frac{\tilde{p}_n}{1 - \tilde{p}_n}, \tag{3.3}$$

so following the delta method, as $n \to \infty$, we have

$$\sqrt{n}\,(\log \tilde{q}_n - \log q) \xrightarrow{D} N\left(0, \sigma_p^2\right), \tag{3.4}$$

where $\sigma_p^2 = 1/\{p(1-p)\}$. In the light of (2.12) and (2.13), with preassigned d and γ, we consider the $100(1-\gamma)\%$ fixed-accuracy confidence interval for q:

$$J_{n;2} = \left(d^{-1}\tilde{q}_n, d\tilde{q}_n\right), \tag{3.5}$$

which additionally satisfies the condition that

$$P_{\alpha,\beta}\{q \in J_{n;2}\} \geq 1 - \gamma \tag{3.6}$$

in order to obtain a confidence interval with $100(1-\gamma)\%$ confidence level.

Likewise, we derive an alternative optimal fixed sample size given by

$$n_2^* \equiv n_2^*(d) = \left(\frac{z}{\log d}\right)^2 \sigma_p^2, \tag{3.7}$$

which remains unknown, again. Therefore, it is essential to estimate σ_p^2 through sequential procedures. Here, we adopt the estimator

$$\hat{\sigma}_{n;2}^2 \equiv \sigma_{\tilde{p}_n}^2 = \frac{1}{\tilde{p}_n(1-\tilde{p}_n)}. \tag{3.8}$$

3.1. A purely sequential procedure

One can immediately identify a similarity between (2.14)–(2.15) and (3.7)–(3.8). For situations when both α and β are unknown, we can follow along the lines of Section 2, and develop a purely sequential fixed-accuracy confidence interval estimation procedure associated with the following stopping rule:

$$\mathcal{P}_2 : N_2 \equiv N_2(d) = \inf\left\{n \geq m : n \geq \left(\frac{z}{\log d}\right)^2 \left(\hat{\sigma}^2_{n;2} + n^{-1}\right)\right\}, \tag{3.9}$$

where $m \geq 1$ continues to indicate a pilot sample size, and $\hat{\sigma}^2_{n;2}$ is defined in (3.8). Different from the stopping rule (2.16), the term n^{-1} is incorporated here to prevent unexpected early termination of sampling due to the fact that $\tilde{p}_n$ is discrete and hence has a positive probability to be 0 or 1.

With the pilot sample data, if $m \geq (z/\log d)^2(\hat{\sigma}^2_{m;2} + m^{-1})$, we do not take any extra observations, and the final sample is the pilot sample. Otherwise, we draw one observation at-a-time and check the stopping rule (3.9) successively with updated $\hat{\sigma}^2_{n;2}$ until for the first time $n \geq (z/\log d)^2(\hat{\sigma}^2_{n;2} + n^{-1})$ is observed. And the final sample size will be $N_2 = n(> m)$. The following theorem states that the procedure $\mathcal{P}_2$ also terminates w.p.1.

Theorem 3. *For the purely sequential fixed-accuracy confidence interval estimation procedure $\mathcal{P}_2$ given in* (3.9)*, with all fixed α, β, γ, d, c, and m, we have:*

$$P_{\alpha,\beta}(N_2 < \infty) = 1. \tag{3.10}$$

Proof: Clearly, $\bar{Y}_n \xrightarrow{P} p$ as $n \to \infty$. Applying Slutsky's theorem, we have that $\hat{\sigma}^2_{n;2} \to \sigma^2_p$ as $n \to \infty$. Therefore, $P_{\alpha,\beta}(N_2 < \infty) = 1$ holds. □

Upon termination, with the acquired data

$$\{N_2, X_1, \ldots, X_m, \ldots, X_{N_2}\},$$

we propose the $100(1-\gamma)\%$ confidence interval for q is:

$$J_{N_2;2} = \left(d^{-1}\tilde{q}_{N_2}, d\tilde{q}_{N_2}\right) = \left(\frac{\tilde{p}_{N_2}}{d(1-\tilde{p}_{N_2})}, \frac{d\tilde{p}_{N_2}}{1-\tilde{p}_{N_2}}\right),$$

and accordingly, the confidence interval for $p = P_{\alpha,\beta}(X > c)$ is:

$$J^*_{N_2;2} = \left(\frac{\tilde{p}_{N_2}}{d-(d-1)\tilde{p}_{N_2}}, \frac{d\tilde{p}_{N_2}}{1+(d-1)\tilde{p}_{N_2}}\right) \quad \text{for } p,$$

where $\tilde{p}_{N_2}$ comes from (3.1). The purely sequential procedure $\mathcal{P}_2$ also enjoys asymptotic first-order efficiency and consistency.

Theorem 4. *For the purely sequential fixed-accuracy confidence interval estimation procedure $\mathcal{P}_2$ given in* (3.9)*, with all fixed α, β, γ, c, and m, we have:*

$$\lim_{d\to 1} E_\beta\left[\frac{N_2}{n_2^*}\right] = 1 \text{ [}\textit{Asymptotic First-Order Efficiency}\text{]}, \tag{3.11}$$

$$\lim_{d\to 1} P_\beta\{q \in J_{N_2;2}\} = 1-\gamma \text{ [}\textit{Asymptotic Consistency}\text{]}, \tag{3.12}$$

Table 2: Simulation results by implementing $\mathcal{P}_2$ as (3.9) using $\Gamma(2,2)$ with $\gamma = 0.05$, $c = 1$, and $m = 20$ under 10,000 runs, where $p = P_{\alpha,\beta}(X > c) = 0.4060$

d	n_2^*	$\bar{N}_2$	$s(N_2)$	$\bar{N}_2/n_2^*$	$\bar{N}_2 - n_2^*$	$\overline{\text{cp}}_2$	$s(\overline{\text{cp}}_2)$	$\bar{\tilde{p}}_{N_2}$	$s(\bar{\tilde{p}}_{N_2})$
1.70	56.57	58.27	3.35	1.0301	1.70	0.9599	0.0020	0.4086	0.0006
1.65	63.52	65.24	3.40	1.0271	1.72	0.9557	0.0021	0.4083	0.0006
1.60	72.11	73.87	3.63	1.0245	1.76	0.9615	0.0019	0.4079	0.0006
1.55	82.93	84.69	3.82	1.0212	1.76	0.9551	0.0021	0.4075	0.0005
1.50	96.89	98.68	4.10	1.0184	1.79	0.9562	0.0020	0.4075	0.0005
1.45	115.38	117.19	4.38	1.0157	1.81	0.9491	0.0022	0.4075	0.0004
1.40	140.70	142.53	4.72	1.0131	1.83	0.9503	0.0022	0.4072	0.0004
1.35	176.86	178.66	5.35	1.0102	1.80	0.9540	0.0021	0.4070	0.0004
1.30	231.40	233.15	6.00	1.0075	1.75	0.9479	0.0022	0.4069	0.0003
1.25	319.90	321.69	6.96	1.0056	1.79	0.9526	0.0021	0.4068	0.0003
1.20	479.19	480.97	8.63	1.0037	1.78	0.9469	0.0022	0.4064	0.0002
1.15	815.46	817.23	11.03	1.0022	1.77	0.9521	0.0021	0.4063	0.0002
1.10	1753.49	1755.38	16.14	1.0011	1.89	0.9527	0.0021	0.4061	0.0001

where n_2^* is defined in (3.7)*.*

Theorem 4 can be proved in the same fashion as we proved Theorem 2, as long as one notes that $n/\hat{\sigma}_{n;2}$ is the observed Fisher information in terms of $\log q$ and refers to Yu (1989). We leave out many details here for brevity.

3.2. Simulated performances

Next, we include a simulation study by implementing the purely sequential fixed-accuracy confidence interval estimation procedures $\mathcal{P}_2$ based on the stopping rule given by (3.9) in an analogous way as we did in Sections 2. We drew pseudo-random observations from a $\Gamma(2,2)$ population; that is, we had fixed $\alpha = 2$ and $\beta = 2$, but pretended that they were unknown. Then, we set $c = \alpha/\beta = 1$ so that $P_{\alpha,\beta}(X > c)$ represented the probability that an observation would turn out larger than the mean value. For brevity alone, we present in Table 2 the selected summaries when the pilot sample size $m = 20$, the significance level $\gamma = 0.05$, and $d = 1.10, 1.15, 1.20, 1.25, 1.30, 1.35, 1.40, 1.45, 1.50, 1.55, 1.60, 1.65, 1.70$.

Running 10,000 independent trials, we reported the mean and standard deviation of the terminated sample sizes $\bar{N}_2$ and $s(N_2)$, the ratio of $\bar{N}_2$ to the optimal fixed sample size n_2^* which should be close to 1, the average coverage probability $\overline{\text{cp}}_2$ which should be comparable with $1 - \gamma (= 0.95)$ along with its standard error $s(\overline{\text{cp}}_2)$, and the mean of estimated values of $p = P_{\alpha,\beta}(X > c)$ with its standard error, $\bar{\tilde{p}}_{N_2}$ and $s(\bar{\tilde{p}}_{N_2})$. The simulation results in Table 2 obviously double validate all the findings of the nonparametric procedure that we proposed for (3.9). Moreover, it is worth mentioning that the differences between $\bar{N}_2$ and n_2^* were empirically around 1.7–1.8 for different values of d, which further shows that our proposed procedure $\mathcal{P}_2$ gives a very consistent and efficient determination of the required sample size for estimation. And that $\bar{\tilde{p}}_{N_2}$ values are close to p with tiny standard errors suggests the sampling procedure can provide perfect estimates for this probability of interest.

4. Real data illustrations

To demonstrate the practical applicability of our newly proposed fixed-accuracy confidence interval estimation procedures, we include illustrations using three real-life data sets: (i) the urine albumin-to-creatinine ratios (UACR, mg/g) of 5255 adolescent survey participants from NHANES 1999–2004, referred to as the "UACR data"; (ii) excess cycle times data in steel manufacturing; (iii) survival times

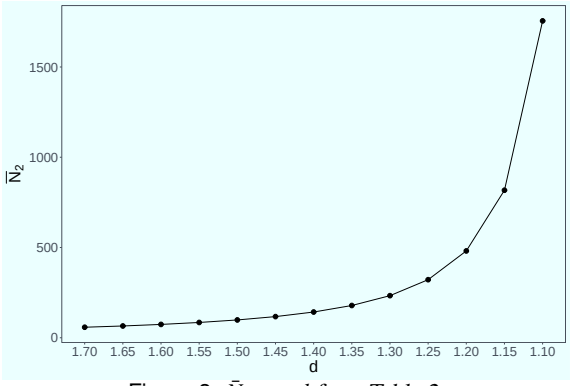

Figure 3: $\bar{N}_2$ *v.s. d from Table 2.*

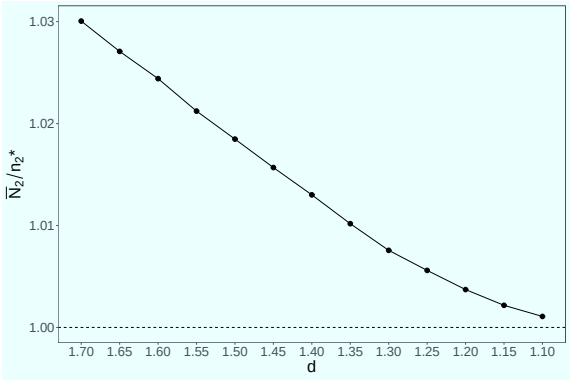

Figure 4: $\bar{N}_2/n_2^*$ *v.s. d from Table 2.*

data from a group of 97 female dementia patients diagnosed at age 70–74.

4.1. Illustration I: using UACR data

The reference population for our analysis is created using survey participants from NHANES 1999–2014 who met the following criteria: between 12 and 17 years old, not pregnant, blood pressure $< 120/80$ mmHg, without diabetes, no prescription medications used within the previous 30 days, and a Z-score for weight-to-height ratio ≤ 2. This yields a reference sample of size $n = 5255$. The UACR data were analyzed in Zou and Young (2020), where the authors had established a one-sided upper tolerance limit of 0.7358 mg/g and suggested to classify adolescents with UACR values falling below it as "strictly normal." Interested readers may refer to the paper, and more background information is omitted here.

For illustrative purposes, we also treated the UACR data of size $n = 5255$ as our population, to which the exponential distribution gave a good fit (Zou and Young, 2020); that is, the shape parameter $\alpha = 1$, but the rate parameter β was assumed unknown. We set $c = 0.7358$ and proceeded to estimate $p = P_\beta(X > c)$, which can be interpreted as the proportion of healthy adolescents with mildly increased UACR. Implementing the procedures $\mathcal{P}_1$ introduced earlier to draw observations from the UACR data set under simple random sampling without replacement, we constructed fixed-accuracy confidence intervals for q and p, respectively. The results are summarized in Table 3.

Table 3: Fixed-accuracy confidence intervals using the UACR data with $\gamma = 0.05$, $c = 0.7358$ and $m = 20$ implementing $\mathcal{P}_1$ from (2.16)

d	N_1	$J_{N_1;1}$ for q	$J^*_{N_1;1}$ for p
1.20	745	(0.0993, 0.1429)	(0.0903, 0.1251)
1.18	1045	(0.0758, 0.1055)	(0.0705, 0.0955)
1.16	1388	(0.0707, 0.0951)	(0.0660, 0.0868)
1.14	1586	(0.0890, 0.1156)	(0.0817, 0.1037)
1.12	2448	(0.0685, 0.0859)	(0.0641, 0.0791)
1.10	2939	(0.0911, 0.1103)	(0.0835, 0.0993)

Table 4: Final sample data of excess cycle times using $\mathcal{P}_1$ from (2.16)

5	32	3	21	7	3	1	7	4	4	9	7	2	11	4	11
5	1	5	7	13	3	10	8	10	11	32	11	3	11	3	7
3	5	12	3	3	2	2	8	10	21	13	3	8	2	8	3
14	8	1	2	3	15	1	3	1	2	5	10	5	1	10	3
2	2	5	4	12	5	8	7	5	10	6	12	3	8	1	1
7	5	2	2	21											

4.2. Illustration II: using excess cycle times data

The excess cycle times data in steel manufacturing was first given in Example 6.1 from Barnett and Lewis (1994) and it is assumed to be a sample from an exponential population. Kimber (1982) and Lin and Balakrishnan (2009) declared that the observed value 92 is an outlier. For this illustration, we use the sample data without the data point 92, and set the shape parameter $\alpha = 1$ with an unknown rate parameter. We are interested in knowing the excess cycle time being greater than 35 as 35 is the largest normal value that is observed. We further assigned $m = 5$ and $\gamma = 0.05$. In the first step, we took the first five data points as our pilot sample data. Then, we checked with the stopping rule as per (2.16) and found it was not satisfied. Thus, we continued our sampling one-at-a-time according to (2.16). In the end, the sampling was terminated with 85 observations, which was recorded in Table 4 as the observations came in. The final 95% confidence interval estimation for $P_\beta(X > 35)$ is $(0.00214, 0.01897)$.

4.3. Illustration III: using survival time data from dementia patients

To illustrate the nonparametric procedure as per (3.9), we consider a data set on survival times from a group of 97 female dementia patients diagnosed at age 70–74. This data set was originally from Elandt-Johnson and Johnson (1999) and was recently discussed by Ozonur and Paul (2020) where they showed that the two-parameter gamma distribution adequately fits the dementia data. One may find the data set from Table 8 of Ozonur and Paul (2020).

According to the analysis by Xie *et al.* (2008), the estimated median survival time from onset of dementia to death was 4.6 years for women. Thus, it will be very helpful to know the probability of living exceeding the median time. We first randomly ordered dementia data and assuming that the data came in with that order. Then, we set $\gamma = 0.05$, $d = 1.6$, $c = 4.6$, $m = 5$. That is, we took the first five data points for initial analysis, and found that the stopping rule as per (3.9) was not satisfied. Thus, we continued our sampling one-at-a-time according to (3.9). The sampling was terminated after collecting the data from 71 patients, and the final data we collected are listed in Table 5.

In the end, using the final observations from Table 5, we were able to get a 95% confidence interval estimation for $p = P_{\alpha,\beta}(X > 4.6)$ to be $(0.3263, 0.5535)$, where X denotes the survival time for a female dementia patient. Thus, using the data we obtained, we would conclude with 95% confidence

Table 5: Final sample data of survival time (in years) for female dementia patients using $\mathcal{P}_2$ from (3.9)

6.75	1.59	0.50	0.50	4.17	3.58	8.16	2.33	1.67	10.17	3.75	1.83
21.00	1.00	1.83	0.50	1.66	1.42	1.67	9.33	4.08	18.08	1.00	7.84
2.67	1.58	1.67	7.83	9.17	1.42	2.00	12.50	4.92	21.83	0.58	1.67
1.25	11.25	4.67	5.25	2.17	3.92	13.83	5.83	0.83	4.17	1.25	3.42
8.50	11.50	10.33	5.25	1.08	3.42	11.25	5.75	2.00	5.58	0.83	1.08
6.92	4.58	7.00	4.92	7.83	2.92	7.84	2.25	3.08	9.92	6.58	

that there is a 32.63% to 55.36% chance that women aged 70–74 can live more than 4.6 years from the onset of dementia.

5. Concluding thoughts

Survival and reliability analysis are two of the most important scientific fields where the gamma distribution is often used to model data. And in these two fields, it is crucial to understand when the measurement, the random variable that is modeled using a gamma distribution, goes beyond a "dangerous" value. Therefore, in the paper, we focus on estimating $P(X > c)$ from a two-parameter gamma population with an unknown rate parameter, and the shape parameter is either known or unknown. In cases of the known shape parameter, such as a sample from an exponential population, we provide an estimation strategy along with a purely sequential procedure for determining the necessary sample size of required accuracy. When the shape parameter is unknown, we come up with a nonparametric purely sequential procedure for achieving the prefixed accuracy. Both procedures perform well in terms of asymptotic efficiency and asymptotic consistency.

Finally, it is also worth mentioning that the nonparametric sequential fixed-accuracy confidence interval estimation procedure developed in Section 3 can be further extended to estimate $P(X > c)$ in a general distribution-free case, since the idea does not depend on any specific population distribution. The gamma population with unknown shape parameter α and unknown rate parameter β can be viewed as an illustration.

Acknowledgement

We are grateful to the referees, whose comments helped improve this article.

References

Ansell JI and Phillips MJ (1994). *Practical Methods for Reliability and Data Analysis*, Clarendon Press, Oxford.

Anscombe FJ (1952). Large-sample theory of sequential estimation, *Mathematical Proceedings of Cambridge Philosophical Society*, **48**, 600–607.

Anscombe FJ (1953). Sequential estimation, *Journal of Royal Statistical Society, Series B*, **15**, 1–29.

Balakrishnan N, Castilla E, Martín N, and Pardo L (2019). Robust estimators for one-shot device testing data under gamma lifetime model with an application to a tumor toxicological data, *Metrika*, **82**, 991–1019.

Balakrishnan N and Ling MH (2014). Gamma lifetimes and one-shot device testing analysis, *Reliability Engineering & System Safety*, **126**, 54–64.

Banerjee S and Mukhopadhyay N (2016). A general sequential fixed-accuracy confidence interval estimation methodology for a positive parameter: Illustrations using health and safety data, *Annals of Institute of Statistical Mathematics*, **68**, 541–570.

Bapat SR (2018). Purely sequential fixed accuracy confidence intervals for $P(X < Y)$ under bivariate exponential models, *American Journal of Mathematical and Management Sciences*, **37**, 386–400.

Barnett V and Lewis T (1994). *Outliers in Statistical Data* (3rd edition), John Wiley & Sons, Chichester, England.

Burgin TA (1975). The gamma distribution and inventory control, *Journal of the Operational Research Society*, **26**, 507–525.

Chen Z and Mi J (1998). Statistical estimation for the scale parameter of the gamma distribution based on grouped data, *Communications in Statistics-Theory and Methods*, **27**, 3035–3045.

Choi SC and Wette R (1969). Maximum likelihood estimation of the parameters of the gamma distribution and their bias, *Technometrics*, **11**, 683–690.

Chow YS and Robbins H (1965). On the asymptotic theory of fixed-width sequential confidence intervals for the mean, *Annals of Mathematical Statistics*, **36**, 457–462.

Dagpunar J (2019). The gamma distribution, *Significance*, **16**, 10–11.

Elandt-Johnson RC and Johnson NL (1999). *Survival Models and Data Analysis*, John Wiley and Sons, New York.

Ghosh M and Mukhopadhyay N (1981). Consistency and asymptotic efficiency of two-stage and sequential procedures, *Sankhyā, Series A*, **43**, 220–227.

Ghosh M, Mukhopadhyay N, and Sen PK (1997). *Sequential Estimation*, Wiley, New York.

Husak GJ, Michaelsen J, and Funk C (2007). Use of the gamma distribution to represent monthly rainfall in Africa for drought monitoring applications, *International Journal of Climatology*, **27**, 935–944.

Iliopoulos G (2016). Exact confidence intervals for the shape parameter of the gamma distribution, *Journal of Statistical Computation and Simulation*, **86**, 1635–1642.

Isogai E and Uno C (1995). On the sequential point estimation of the mean of a gamma distribution, *Statistics & Probability Letters*, **22**, 287–293.

Johnson RW and Kliche DV (2020). Large sample comparison of parameter estimates in gamma raindrop distributions, *Atmosphere*, **11**, 333.

Kimber AC (1982). Tests for many outliers in an exponential sample, *Applied Statistics*, **31**, 263–271.

Krishnamoorthy K and Mathew T (2009). *Statistical Tolerance Regions: Theory, Applications, and Computation*, John Wiley & Sons, New York.

Lawless JF (1982). *Statistical Models and Methods for Lifetime Data*, John Wiley & Sons, New York.

Lin C and Balakrishnan N (2009). Exact computation of the null distribution of a test for multiple outliers in an exponential sample, *Computational Statistics & Data Analysis*, **53**, 3281–3290.

Liu JF (2001). Two-stage approximation of expected reward for gamma random variables, *Communications in Statistics-Theory and Methods*, **30**, 1471–1480.

Mahmoudi E and Roughani G (2015). Bounded risk estimation of the scale parameter of a gamma distribution in a two-stage sampling procedure, *Sequential Analysis*, **34**, 25–38.

Mukhopadhyay N and Banerjee S (2014). Purely sequential and two stage fixed-accuracy confidence interval estimation methods for count data for negative binomial distributions in statistical ecology: One-sample and two-sample problems, *Sequential Analysis*, **33**, 251–285.

Mukhopadhyay N and de Silva BM (2009). *Sequential Methods and Their Applications*, CRC, Boca Ratton.

Mukhopadhyay N and Zhuang Y (2016). On fixed-accuracy and bounded accuracy confidence interval estimation problems in Fisher's "Nile" example, *Sequential Analysis*, **35**, 516–535.

Nadarajah S and Gupta AK (2007). The exponentiated gamma distribution with application to drought

data, *Calcutta Statistical Association Bulletin*, **59**, 29–54.
Ohlsson E and Johansson B (2010). *Non-Life Insurance Pricing with Generalized Linear Models*, Springer, Berlin.
Ozonur D and Paul S (2020). Goodness of fit tests of the two-parameter gamma distribution against the three-parameter generalized gamma distribution, *Communications in Statistics-Simulation and Computation*, in press.
Piegorsch WW (2015). *Statistical Data Analytics: Foundations for Data Mining, Informatics, and Knowledge Discovery*, John Wiley & Sons, University of Arizona.
Roughani G and Mahmoudi E (2015). Exact risk evaluation of the two-stage estimation of the gamma scale parameter under bounded risk constraint, *Sequential Analysis*, **34**, 387–405.
Siegmund D (1985). *Sequential Analysis: Tests and Confidence Intervals*, Springer, New York.
Son YS and Oh M (2006). Bayesian estimation of the two-parameter gamma distribution, *Communications in Statistics-Simulation and Computation*, **35**, 285–293.
Stein C (1949). Some problems in sequential estimation, *Econometrica*, **17**, 77–78.
Takada Y and Nagata Y (1995). Fixed-width sequential confidence interval for the mean of a gamma distribution, *Journal of Statistical Planning and Inference*, **44**, 277–289.
Vaz MF and Fortes MA (1988). Grain size distribution: the lognormal and the gamma distribution functions, *Scripta Metallurgica*, **22**, 35–40.
Woodroofe M (1977). Second order approximations for sequential point and interval estimation, *Annals of Statistics*, **5**, 984–995.
Xie J, Brayne C, and Matthews FE (2008). Survival times in people with dementia: analysis from population based cohort study with 14 year follow-up, *BMJ: British Medical Journal*, **336**, 258–262.
Yu KF (1989). On fixed-width confidence intervals associated with maximum likelihood estimation, *Journal of Theoretical Probability*, **2**, 193–199.
Zacks S and Khan RA (2011). Two-stage and sequential estimation of the scale parameter of a gamma distribution with fixed-width intervals, *Sequential Analysis*, **30**, 297–307.
Zou Y and Young DS (2020). Improving coverage probabilities for parametric tolerance intervals via bootstrap calibration, *Statistics in Medicine*, **39**, 2152–2166.

Received July 17, 2020; Revised August 23, 2020; Accepted September 19, 2020