Lower Confidence Limit for Reliability Based on Grouped Data with a Quantile Filling Algorithm

Mimi Zhang[1], Qingpei Hu[2], Min Xie[1], Dan Yu[2]

[1]Department of Systems Engineering and Engineering Management, City University of Hong Kong, Hong Kong, China

[2]Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, China

Abstract

The purpose of this article is to derive a lower confidence limit for reliability given a grouped data set. This is done by using a quantile filling algorithm which generates pseudo failure data from grouped data. A general framework of this approach is first introduced. The cases for the exponential distribution, Weibull distribution, and lognormal distribution are used to illustrate this approach. Simulation studies are carried out and the results show that it is a useful method handling grouped data. Two field lifetime data sets are also analyzed to demonstrate its feasibility. Some further improvements are discussed as well.

Key words: Quantile-Filling algorithm; Lower confidence limit for reliability; Grouped data; Exponential Model; Weibull Model; Lognormal Model.

1. Introduction

Grouped data is one type of incomplete data for which data are categorized into intervals (Murthy et al., 2003). Generally, these intervals are non-overlapped. Grouped data is common in many disciplines such as engineering, finance, and biostatistics. In view of its prevalence, feasible and powerful methods should be explored as grouped data contains quite small information in survival time. Various approaches to analyze grouped data have been developed, via frequentist, Bayesian view or fuzzy logic. Kalbfleisch and Prentice (1973) developed a Cox proportional hazards model for grouped data. Davison and Tsai (1992) extended generalized linear model to grouped data. Yang and Yu (2005) proposed a method to estimate the parameters in fuzzy class models using a fuzzy clustering algorithm. Recently, many techniques have been studied concerning grouped data, e.g., Bassetti et al. (2007), Meister (2007), Rivero and Valdes (2008), Lambert (2011) and Ryan et al. (2011).

Data completion is a general statistical method for the analysis of incomplete data sets. It is natural as the most challenging is the lack of information when dealing with incomplete data. Data completion (or imputation) technique has been extensively developed to handle missing data. Rubin (1976) first came up with the multiple imputations to estimate incomplete data regression model. Some articles further studied the relative accuracy of multiple imputations e.g., Reiter and Raghunathan (2007), Reiter (2007) and Holan et al. (2010). While for censored/truncated data, different approaches have been studied on data completion, including regression models for parameters like the survival function in a single point, the restricted mean survival time, and transition or state occupation probabilities in multi-state models, e.g., the competing risks cumulative incidence function. Buckley and James (1979) adopted a data completion technique (replacing the censored observations by their estimated conditional expectations) to estimate the parameters of linear regression model. Lai and Ying (1991b) proved that a modified Buckley-James estimator is consistent and asymptotically normal under certain conditions. Wang et al. (2008) related high dimensional genomic data to survival outcomes using a semi-parametric accelerated failure time model, with a doubly penalized Buckley-James method used for estimation. Liu et al. (2011) proposed a robust multiple imputation approach directly imputes restricted lifetimes over the study period on a model of the mean restricted life as a linear function of covariates. Andersen (2010) presented a review of recent works on the application of pseudo observations in survival and event history analysis.

Yu and Dai (1996) and Yu and Guo (2001) developed an algorithm, called quantile filling algorithm, to generate complete data from censored data on the basis of maximum likelihood estimates. One of the drawbacks of the original procedure is the cumbersome computation caused by maximum likelihood estimation (MLE). Besides, the convergence and consistency for these algorithms have not been studied. Jiang (2008) proposed an alternative quantile filling algorithm using moment invariance criterion and proved the convergence of the algorithm as well as the consistency of the estimators. Considering this, the combination of moment estimation and conditional quantile filling is not only feasible, efficient but also convenient.

An important purpose of this paper is to compute a lower confidence limit for reliability, based on pseudo complete data. With respect to the performance, we usually require an assurance regarding the minimum value of some indices. Hence, lower confidence limits are of interest. Heard and Pensky (2006) considered the construction of confidence intervals for reliability when the sample size is relatively small. Some works on lower confidence limit can be found in Weerahandi (1993), Liu and Lindsay (2009) and McKane et al. (2005). The lower confidence limit for reliability in these papers is a standard one-sided interval estimation based on pivot statistics. A precondition for this approach is the complete lifetime data. When dealing with incomplete data, we utilize the quantile filling algorithm and replace the incomplete data by its corresponding pseudo complete data.

The rest of this article is organized as follows. The general framework of grouped data, quantile filling algorithm, and a lower confidence limit for reliability are introduced in Section 2. From Sections 3 to 5, we derive the corresponding quantile filling algorithm and a lower confidence limit for reliability based on exponential distribution, Weibull distribution, and lognormal distribution, respectively. Section 6 provides three illustrative examples: a simulation, an application to characterize the lifetime distribution of aluminum coupon and an application to model the distribution of ball bearing. Some further discussions and concluding remarks are given in Section 7.

2. Quantile Filling Algorithm and Lower Confidence Limit for Reliability

Before introducing the data completion algorithm and the lower confidence limit for reliability, a list of notation is given here.

LCLR	lower confidence limit for reliability
QF	quantile filling
ME	moment estimation
k	sampling number
n_i	sample size of the *i*th sampling
f_i	failure number at the *i*th sampling
$t_{i,j}^{(m)}$, $x_{i,j}^{(m)}$	pseudo complete data got at the *m*th step of QF algorithm for the *j*th item of the *i*th sampling
$T_i^{(m)}$, $X_i^{(m)}$	pseudo complete data vector got at the *m*th step of QF algorithm for the *i*th sampling
$T^{(m)}$, $X^{(m)}$	pseudo complete data vector got at the *m*th step of QF algorithm

Suppose that at the ith inspection time t_i, n_i units are drawn at random from the population and the number of expired units f_i is recorded ($i = 1,2, \dots, k$). We hence obtain a grouped data $(f_1, f_2, \dots, f_k)$ with sample sizes $(n_1, n_2, \dots, n_k)$ and sampling times $(t_1, t_2, \dots, t_k)$, where k is a pre-assigned constant. We can simply denote this grouped data set as $\{(t_i, n_i, f_i),\ i = 1,2, \dots, k\}$.

Assume that the survival times are independent and identically distributed random variables from a continuous distribution with distribution function $G(t,\ \theta)$, where θ is the unknown parameter vector to be estimated. Let $T_i^{(m)} = (t_{i,1}^{(m)}, \ldots, t_{i,f_i}^{(m)}, t_{i,f_i+1}^{(m)}, \ldots, t_{i,n_i}^{(m)})$ denote the result (pseudo complete data) obtained after m cycles of the QF algorithm. Set $\theta^{(m)}$ to be the moment estimation vector of the unknown parameter vector θ , calculated from $T^{(m)} = (T_1^{(m)}, T_2^{(m)}, \ldots, T_k^{(m)})$ $(m = 1,2, \ldots)$.

Based on moment invariance criterion, the QF algorithm starts with an initial guess at the parameter vector $\theta^{(0)}$. $\theta^{(0)}$ can be derived by regular parameter estimation techniques. QF algorithm seeks to replace the grouped data and to estimate the unknown parameters by iteratively applying the following two steps:

QF step: Replace $T_i^{(m-1)} = (t_{i,1}^{(m-1)}, \ldots, t_{i,f_i}^{(m-1)}, t_{i,f_i+1}^{(m-1)}, \ldots, t_{i,n_i}^{(m-1)})$ by $T_i^{(m)} = (t_{i,1}^{(m)}, \ldots, t_{i,f_i}^{(m)}, t_{i,f_i+1}^{(m)}, \ldots, t_{i,n_i}^{(m)})$ where $t_{i,j}^{(m)}$ $(i = 1,2, \ldots, k, m = 1,2, \ldots)$ is given by

$$t_{i,j}^{(m)} = \begin{cases} G_{\theta^{(m-1)}}^{-1}(\dfrac{j}{f_i + 1} | T \leq t_i) & 1 \leq \mathrm{j} \leq f_i \\ G_{\theta^{(m-1)}}^{-1}(\dfrac{j - f_i}{n_i - f_i + 1} | T > t_i) & f_i < j \leq n_i \end{cases}$$

and where $T_i^{(0)} = \{\underbrace{t_i, \ldots, t_i}_{n_i}\}$.

ME step: Calculate the *m*th moment estimate vector $\theta^{(m)}$ based on $T^{(m)} = (T_1^{(m)}, T_2^{(m)}, \ldots, T_k^{(m)})$, using standard complete-data methods.

Stop when $||\theta^{(m)} - \theta^{(m-1)}|| < \varepsilon$, some pre-assigned tolerance limit. $\theta^{(m)}$ is the estimate of the unknown parameter vector and $T^{(m)}$ is the pseudo complete survival times. The tolerance limit ε should be small enough to guarantee the accuracy.

Remark. $G_{\theta^{(m-1)}}^{-1}(t_{i,j}^{(m)}|T \leq t_i)$ and $G_{\theta^{(m-1)}}^{-1}(t_{i,j}^{(m)}|T > t_i)$ are the inverse functions of $G_{\theta^{(m-1)}}(t_{i,j}^{(m)}|T \leq t_i)$ and $G_{\theta^{(m-1)}}(t_{i,j}^{(m)}|T > t_i)$ respectively:

$$G_{\theta^{(m-1)}}\left(t_{i,j}^{(m)}\middle|T \leq t_i\right) = \frac{P(T \leq t_{i,j}^{(m)})}{P(T \leq t_i)} = \frac{G(t_{i,j}^{(m)}, \theta^{(m-1)})}{G(t_i, \theta^{(m-1)})} = \frac{j}{f_i + 1} \qquad 1 \leq \mathrm{j} \leq f_i$$

$$G_{\theta^{(m-1)}}\left(t_{i,j}^{(m)}\middle|T > t_i\right) = \frac{G\left(t_{i,j}^{(m)}, \theta^{(m-1)}\right) - G\left(t_i, \theta^{(m-1)}\right)}{1 - G(t_i, \theta^{(m-1)})} = \frac{j - f_i}{n_i - f_i + 1} \qquad f_i < j \leq n_i$$

where $T \sim G(t, \theta^{(m-1)})$ is the survival time random variable.

Below we derive a lower confidence limit for reliability based on this pseudo survival time data $T^{(m)}$. We do not give a derivation of the algorithm used to obtain the LCLR, but the following example should clarify its application.

Suppose that the survival time T of a component is a random variable following the cumulative distribution $G(t, \theta)$ given before. The reliability function, say $R(t, \theta)$, is then reduced to

$$R(t, \theta) = P(T > t) = 1 - G(t, \theta).$$

The α lower confidence limit for reliability at time *t*, to be denoted by $R_L(\alpha)$, is given by

$$P\{R(t,\theta) \geq R_L(\alpha)\} = \alpha \qquad (2.1)$$

It should be noted that the LCLR is a bivariate function of *t* and α.

For convenience, we denote $T^{(m)} = \left(T_1^{(m)}, T_2^{(m)}, \dots, T_k^{(m)}\right)$ as $T^{(m)} = (T_1, T_2, \dots, T_n)$, where $n = \sum_{i=1}^{k} n_i$. As can be evidenced, $T_1, T_2, \dots, T_n$ is a random sample from $G(t, \theta)$. It is assumed that $X = g(T_1, T_2, \dots, T_n, \theta)$ is a pivotal quantity for θ and that X follows a well-known sampling distribution, say chi-square distribution with *n*-1 degrees of freedom. By denoting θ as $\theta = g^{-1}(T_1, T_2, \dots, T_n, X)$, (2.1) can be rewritten as

$$P\left(M \geq R_L(\alpha)\right) = \alpha$$

where

$$M = R(t, g^{-1}(T_1, T_2, ..., T_n, X)) \qquad (2.2)$$

Sample from chi-square distribution with *n*-1 degrees of freedom so we have data $X_1, X_2, \dots, X_m$. By substituting X_i into (2.2) we have data $M_1, M_2, \dots, M_m$, where $M_i = R\left(t, g^{-1}(T_1, T_2, \dots, T_n, X_i)\right)$. The empirical distribution of random variable M is then obtained based on data $M_1, M_2, \dots, M_m$. The α lower quantile of this empirical distribution is the estimate for $R_L(\alpha)$, i.e. the lower confidence limit for reliability at time t under confidence level α. We can augment the sample size m to improve the estimation accuracy.

3. The Case of Exponential Distribution

Consider the simplest case when T follows the exponential distribution with the cumulative probability distribution given by

$$F(t) = 1 - \exp(-\frac{t}{\theta}) \tag{3.1}$$

where $t > 0$ and $\theta > 0$ is the scale parameter. The moment estimator of the unknown parameter θ is

$$\hat{\theta} = \overline{T}$$

where $\overline{T}$ is the sample mean.

Given a grouped data $\{(t_i, n_i, f_i),\ i = 1,2, \dots, k\}$ described in Section 2, from the exponential distribution (3.1), the QF algorithm can be specified as follows:

QF step: Replace $T_i^{(m-1)} = (t_{i,1}^{(m-1)}, \dots, t_{i,f_i}^{(m-1)}, t_{i,f_i+1}^{(m-1)}, \dots, t_{i,n_i}^{(m-1)})$ by $T_i^{(m)} = (t_{i,1}^{(m)}, \dots, t_{i,f_i}^{(m)}, t_{i,f_i+1}^{(m)}, \dots, t_{i,n_i}^{(m)})$ where $t_{i,j}^{(m)}$ $(i = 1,2, \dots, k, m = 1,2, \dots)$ is given by

$$t_{i,j}^{(m)} = \begin{cases} -\theta^{(m-1)} \ln(1 - \frac{j}{f_i+1}(1 - \exp\{-\frac{t_i}{\theta^{(m-1)}}\})) & 1 \le j \le f_i \\ t_i - \theta^{(m-1)} \ln(1 - \frac{j-f_i}{n_i-f_i+1}) & f_i < j \le n_i \end{cases}$$

and where $T_i^{(0)} = \{\underbrace{t_i, \dots, t_i}_{n_i}\}$.

ME step: Calculate the *m*th moment estimate $\theta^{(m)}$ based on $T^{(m)} = (T_1^{(m)}, T_2^{(m)}, \dots, T_k^{(m)})$:

$$\theta^{(m)} = \overline{T^{(m)}}$$

where $\overline{T^{(m)}}$ is the sample mean.

Stop when $|\theta^{(m)} - \theta^{(m-1)}| < \varepsilon$, some pre-assigned tolerance limit. $\theta^{(m)}$ is the estimate of the unknown parameter vector and $T^{(m)}$ is the pseudo complete survival times.

It is assumed that random variables $T_1, T_2, \dots, T_n$ are independent and identically distributed from distribution (3.1) and let $T = \sum_{k=1}^{n} T_k$. The pivotal quantity $\mathrm{X} = 2\mathrm{T}/\theta$ has the chi-square distribution with $2n$ degrees of freedom, namely $\mathrm{X} \sim \chi_{2n}^2$. We express the unknown parameter as a function of pivotal quantity, that is $\theta = 2\mathrm{T}/\mathrm{X}$. The reliability function can hence be rewritten as

$$R(t) = e^{-\frac{t}{\theta}} = e^{-\frac{\mathrm{X}t}{2T}}$$

For a given confidence level α of the LCLR at time *t*, we have

$$\begin{aligned} \alpha &= P(R(t) \geq R_L(\alpha)) \\ &= P(\exp(-\frac{Xt}{2T}) \geq R_L(\alpha)) \\ &= P(X \leq \frac{-2T\ln(R_L(\alpha))}{t}) \end{aligned}$$

As can be evidenced that $-2T\ln(R_L(\alpha))\,/t$ is an upper α quantile of the chi-square distribution with $2n$ degrees of freedom. In other words, given the upper α quantile of the chi-square distribution with $2n$ degrees of freedom denoted by $\chi_{2n}^2(\alpha)$, the LCLR at time *t* is

$$R_L(\alpha) = e^{-\frac{t \times \chi_{2n}^2(\alpha)}{2T}}$$

Mix all the ultimate pseudo complete data obtained from the above QF algorithm as $T^{(m)} = (t_1^{(m)}, \ldots, t_n^{(m)})$, where $n = \sum_{i=1}^k n_i$. The algorithm for the LCLR is given as follow:

Step 1：Calculate $T = \sum_{k=1}^n t_k^{(m)}$ and $\chi_{2n}^2(\alpha)$ where $\chi_{2n}^2(\alpha)$ is the upper α quantile of the chi-square distribution with $2n$ degrees of freedom;

Step 2：Calculate $R_L(\alpha) = e^{-\frac{t \times \chi_{2n}^2(\alpha)}{2T}}$;

4. The Case of Weibull Distribution

Weibull distribution is probably the most commonly used distribution in reliability modeling when the hazard rate is not constant. Assume that random variable T is distributed according to the Weibull model with the cumulative distribution function given by

$$F(t;\alpha,\beta) = 1 - e^{-(\frac{t}{\alpha})^\beta}, \qquad (4.1)$$

where $t > 0, \alpha > 0$ is the scale parameter and $\beta > 0$ is the shape parameter. Let $X = \ln T$, then random variable X follows the extreme value distribution with the cumulative distribution function parameterized in terms of $\mu = \ln \alpha$ and $\sigma = 1/\beta$

$$F(x;\mu,\sigma) = 1 - \exp[-\exp(\frac{x-\mu}{\sigma})], \qquad -\infty < x < +\infty \tag{4.2}$$

Note that $-\infty < \mu < +\infty$ and $\sigma > 0$. The moment estimators of parameters μ and σ reduce to

$$\hat{\sigma} = \frac{\sqrt{6}}{\pi}\sqrt{var[X]} \qquad \hat{\mu} = \overline{X} + \gamma\hat{\sigma}$$

where $\overline{X}$ is sample mean, $var[X]$ is sample variance and $\gamma = 0.5772156$ ... is the Euler constant.

Models (4.1) and (4.2) are equivalent models in the sense the procedure developed under one model can be easily used for the other model. For simplicity, we only consider the extreme value distribution (4.2) here.

Generate a grouped data $\{(x_i, n_i, f_i),\ i = 1,2,\dots,k\}$ from the extreme value distribution (4.2) as follows:

At the ith round, draw randomly n_i items from the population. Record the number of units as f_i, whose values are no more than x_i ($i = 1,2,\dots,k$). In practice, x_i is the logarithm transformation of inspection time t_i.

Let $X_i^{(m)} = (x_{i,1}^{(m)}, \dots, x_{i,f_i}^{(m)}, x_{i,f_i+1}^{(m)}, \dots, x_{i,n_i}^{(m)})$ denote the result (pseudo sample values) got after m cycles of the QF step ($i = 1,\dots,k$ and $m = 1,2,\dots$). $\theta^{(m)} = (\mu^{(m)}, \sigma^{(m)})$ is the moment estimate vector of the unknown parameter vector calculated based on $X^{(m)} = (X_1^{(m)}, X_2^{(m)}, \dots, X_k^{(m)})$.

The QF algorithm in which the population has the extreme value distribution (4.2) is constructed as follows:

QF step: Replace $X_i^{(m-1)} = (x_{i,1}^{(m-1)}, \dots, x_{i,f_i}^{(m-1)}, x_{i,f_i+1}^{(m-1)}, \dots, x_{i,n_i}^{(m-1)})$ by $X_i^{(m)} = (x_{i,1}^{(m)}, \dots, x_{i,f_i}^{(m)}, x_{i,f_i+1}^{(m)}, \dots, x_{i,n_i}^{(m)})$ where $x_{i,j}^{(m)}$ ($i = 1,\dots,k, m = 1,2,\dots$) is given by

$$x_{i,j}^{(m)} = \begin{cases} \mu^{(m-1)} + \sigma^{(m-1)} \ln\{-\ln\{1 - \frac{j}{f_i+1}[1 - \exp(-e^{\frac{x_i-\mu^{(m-1)}}{\sigma^{(m-1)}}})]\}\} & 1 \le j \le f_i \\ \mu^{(m-1)} + \sigma^{(m-1)} \ln\{e^{\frac{x_i-\mu^{(m-1)}}{\sigma^{(m-1)}}} - \ln(1 - \frac{j-f_i}{n_i-f_i+1})\} & f_i < j \le n_i \end{cases}$$

and where $X_i^{(0)} = \{\underbrace{x_i, \dots, x_i}_{n_i}\}$.

ME step: Calculate the mth moment estimate vector $\theta^{(m)}$ based on $X^{(m)} = (X_1^{(m)}, X_2^{(m)}, \dots, X_k^{(m)})$:

$$\sigma^{(m)} = \frac{\sqrt{6}}{\pi}\sqrt{var[X^{(m)}]} \qquad \mu^{(m)} = \overline{X^{(m)}} + \gamma\sigma^{(m)}$$

where $\overline{X^{(m)}}$ is the sample mean and $var[X^{(m)}]$ is the sample variance.

Stop when $||\theta^{(m)} - \theta^{(m-1)}|| < \varepsilon$, some pre-assigned tolerance limit. $\theta^{(m)}$ is the desired estimate vector and $X^{(m)}$ is the ultimate pseudo complete data.

Independent and identically distributed random variables $X_1, \dots, X_n$, denoting survival times, have extreme value distribution (4.2). Consider two dependent pivotal quantities: $f_1 = (\overline{W} - \mu)/\sigma$ and $f_2 = V^2/\sigma^2$ where

$$\overline{W} = \frac{1}{n}\sum_{k=1}^{n} X_k \qquad V^2 = \frac{1}{n-1}\sum_{k=1}^{n}(X_k - \overline{W})^2$$

The reliability function can be rewritten in terms of f_1 and f_2:

$$R(t) = e^{-e^{\frac{\ln t-\mu}{\sigma}}} = e^{-e^{\frac{\ln t-\overline{W}}{V}\sqrt{f_2}+\sqrt{f_1}}} = e^{-e^{M}}$$

where

$$M = \frac{\ln t - \overline{W}}{V}\sqrt{f_2} + \sqrt{f_1}.$$

For a given confidence level α of the LCLR at time *t*, we have

$$\begin{aligned} \alpha &= P(R(t) \geq R_L(\alpha)) \\ &= P(\exp(-\exp(M)) \geq R_L(\alpha)) \\ &= P(M \leq \ln(-\ln R_L(\alpha))) \end{aligned}$$

We denote the upper α quantile of random variable M as M_α. As can be evidenced, the LCLR can be expressed as a function of M_α, which is given by:

$$R_L(\alpha) = e^{-e^{M_\alpha}}$$

Combine all the pseudo complete data obtained from the above QF algorithm as $X^{(m)} = \left(X_1^{(m)}, X_2^{(m)}, \dots, X_k^{(m)}\right) = (x_1, \dots, x_n)$ where $n = \sum_{i=1}^{k} n_i$. The LCLR algorithm for $R_L(\alpha)$ is given as follows:

Step 1：Calculate $\overline{W} = \frac{1}{n}\sum_{k=1}^{n} x_k, V^2 = \frac{1}{n-1}\sum_{k=1}^{n}(x_k - \overline{W})^2$;

Step 2：Draw independent and identically distributed sample from standard extreme value population, denoted as $(\omega_1, \dots, \omega_n)$;

Step 3：Compute

$$f_{11} = \frac{1}{n}\sum_{k=1}^{n} w_k \qquad f_{21} = \frac{1}{n-1}\sum_{k=1}^{n}(w_k - f_{11})^2\,;$$

Step 4: Repeating **Step 2-Step 3** m times so we have $\overrightarrow{f_1} = (f_{11}, f_{12}, \dots, f_{1m})$ and $\overrightarrow{f_2} = (f_{21}, f_{22}, \dots, f_{2m})$;

Step 5: The empirical distribution F_M of random variable M is then obtained by computing $\overrightarrow{M} = \frac{\ln t - \overline{W}}{V}\sqrt{\overrightarrow{f_2}} + \overrightarrow{f_1}$;

Step 6：Calculate $R_L(\alpha) = e^{-e^{M_\alpha}}$ where M_α is the upper α quantile of M;

5. The Case of Lognormal Distribution

Lognormal distribution is another commonly used distribution in reliability and lifetime data analysis. We assume that the random variable T follows a lognormal distribution with parameters μ and σ. The probability density function for the lognormal distribution is given by

$$f(t;\mu,\sigma) = \frac{1}{t\sigma\sqrt{2\pi}} e^{-\frac{(lnt-\mu)^2}{2\sigma^2}} \tag{5.1}$$

where $t > 0, \mu \in R$ and $\sigma^2 > 0$. The moment estimators of parameters μ and σ^2 are

$$\begin{aligned} \widehat{\mu} &= ln(\overline{T}) - \frac{1}{2}ln(1 + \frac{var(T)}{(\overline{T})^2}) \\ \widehat{\sigma^2} &= ln(1 + \frac{var(T)}{(\overline{T})^2}) \end{aligned}$$

where $\overline{T}$ is the sample mean and $var(T)$ is the sample variance. Let $X = \ln T$, then random variable X is normally distributed with the probability density function reduced to

$$f(x;\mu,\sigma)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{5.2}$$

The moment estimators of parameters μ and σ^2 are thus given by

$$\hat{\mu}=\overline{X} \qquad \widehat{\sigma^2}=var(X).$$

$\overline{X}$ is the sample mean and $var(X)$ is the sample variance calculated on logarithm transformed data.

Models (5.1) and (5.2) are equivalent models in the sense the procedure developed under one model can be easily used for the other model. For simplicity, we below just consider the normal distribution (5.2).

Generate a grouped data $\{(x_i,n_i,f_i),\ i=1,2,\ldots,k\}$ from the normal distribution (5.2) as follows:

At the *i*th round, draw randomly n_i items from the population. Record the number of units as f_i, whose values are no more than x_i ($i=1,\ldots,k$). In practice, x_i is the logarithm transformation of inspection time t_i.

Let $X_i^{(m)}=(x_{i,1}^{(m)},\ldots,x_{i,f_i}^{(m)},x_{i,f_i+1}^{(m)},\ldots,x_{i,n_i}^{(m)})$ denote the result (pseudo sample values) got after m cycles of the QF step ($i=1,\ldots,k$ and $m=1,2,\ldots$). $\theta^{(m)}=(\mu^{(m)},\sigma^{(m)})$ is the moment estimate vector of the unknown parameter vector calculated based on $X^{(m)}=(X_1^{(m)},X_2^{(m)},\ldots,X_k^{(m)})$.

The QF algorithm in which the population has the normal distribution (5.2) is constructed as follows:

QF step: Replace $X_i^{(m-1)}=(x_{i,1}^{(m-1)},\ldots,x_{i,f_i}^{(m-1)},x_{i,f_i+1}^{(m-1)},\ldots,x_{i,n_i}^{(m-1)})$ by $X_i^{(m)}=(x_{i,1}^{(m)},\ldots,x_{i,f_i}^{(m)},x_{i,f_i+1}^{(m)},\ldots,x_{i,n_i}^{(m)})$ where $x_{i,j}^{(m)}$ ($i=1,\ldots,k$ and $m=1,2,\ldots$) is given by

$$x_{i,j}^{(m)}=\begin{cases}\mu^{(m-1)}+\sigma^{(m-1)}\Phi^{-1}\left(\frac{j}{f_i+1}\Phi\left(\frac{x_i-\mu^{(m-1)}}{\sigma^{(m-1)}}\right)\right) & 1\le j\le f_i\\ \mu^{(m-1)}+\sigma^{(m-1)}\Phi^{-1}\left(\frac{j-f_i}{n_i-f_i+1}+\left(1-\frac{j-f_i}{n_i-f_i+1}\right)\Phi\left(\frac{x_i-\mu^{(m-1)}}{\sigma^{(m-1)}}\right)\right) & f_i<j\le n_i\end{cases}$$

and where $X_i^{(0)}=\{\underbrace{x_i,\ldots,x_i}_{n_i}\}$.

ME step: Calculate the *m*th moment estimate vector $\theta^{(m)}$ based on $X^{(m)}=(X_1^{(m)},X_2^{(m)},\ldots,X_k^{(m)})$:

$$\sigma^{(m)} = \sqrt{var[X^{(m)}]} \qquad \mu^{(m)} = \overline{X^{(m)}}$$

where $\overline{X^{(m)}}$ is the sample mean and $var[X^{(m)}]$ is the sample variance.

Stop when $||\theta^{(m)} - \theta^{(m-1)}|| < \varepsilon$, some pre-assigned tolerance limit. $\theta^{(m)}$ is the desired estimate vector and $X^{(m)}$ is the ultimate pseudo complete data.

Independent and identically distributed random variables $X_1, \dots, X_n$, denoting survival time, have normal distribution (5.2). Consider two independent pivotal quantities:

$$f_1 = \frac{\sqrt{n}(\overline{X} - \mu)}{\sigma} \sim N(0,1) \qquad f_2 = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

where

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

The reliability function can be rewritten in terms of f_1 and f_2:

$$R(t) = 1 - \Phi\left(\frac{t-\mu}{\sigma}\right) = 1 - \Phi\left(\frac{t-\overline{X}}{\sqrt{n-1}S}\sqrt{f_2} + \frac{f_1}{\sqrt{n}}\right) = 1 - \Phi(M)$$

where

$$M = \frac{t-\overline{X}}{\sqrt{n-1}S}\sqrt{f_2} + \frac{f_1}{\sqrt{n}}.$$

For a given confidence level α of the LCLR at time *t*, we have

$$\begin{aligned} \alpha &= P\{R(t) \geq R_L(\alpha)\} \\ &= P\{1 - \Phi(M) \geq R_L(\alpha)\} \\ &= P\{M \leq \Phi^{-1}(1 - R_L(\alpha))\}. \end{aligned}$$

We denote the upper α quantile of random variable M as M_α. As can be evidenced, the LCLR can be expressed as a function of M_α, which is given by:

$$R_L(\alpha) = 1 - \Phi(M_\alpha).$$

Combine all the pseudo complete data obtained from the above QF algorithm as $X^{(m)} = \left(X_1^{(m)}, X_2^{(m)}, \dots, X_k^{(m)}\right) = (x_1, \dots, x_n)$ where $n = \sum_{i=1}^{k} n_i$. The LCLR algorithm for $R_L(\alpha)$ is given as follows:

Step 1： Calculate $\overline{X} = \frac{1}{n}\sum_{k=1}^{n} x_k$, $S^2 = \frac{1}{n-1}\sum_{k=1}^{n}(x_k - \overline{X})^2$.

Step 2： Draw m independent and identically distributed samples from standard normal population, denoted as $\overrightarrow{f_1} = (f_{11}, f_{12}, \dots, f_{1m})$.

Step 3： Draw m independent and identically distributed samples from chi-square population ($n-1$ degrees of freedom), denoted as $\overrightarrow{f_2} = (f_{21}, f_{22}, \dots, f_{2m})$.

Step 4: The empirical distribution F_M of random variable M is then obtained by computing $\overrightarrow{M} = \frac{t-\overline{X}}{\sqrt{n-1}S}\sqrt{\overrightarrow{f_2}} + \frac{\overrightarrow{f_1}}{\sqrt{n}}$.

Step 5： Calculate $R_L(\alpha) = 1 - \Phi(M_\alpha)$, where M_α is the upper α quantile of M.

6. Numerical Examples

Three illustrative examples are used here to show the applications of the proposed algorithms. Without loss of generality, we let the sample sizes equal to each other, namely $n_1 =, \dots, = n_k = n$.

6.1 Simulation on Exponential Distribution

The simulation is run 10000 times for each combination of (n, k). Various sampling schemes are considered, namely for fixed sample size n to increase sampling times k and vice versa. Four statistics are calculated: the average of the 10000 estimates of the unknown parameter, the coverage probability, the upper α quantile of the lower confidence limits for reliability at confidence level α, and the mean squared error of these lower confidence limits. Coverage probability is a widely used tool measuring prediction accuracy of confidence interval, see Wang (2008) and Kabaila and Leeb (2006). By the definition of lower confidence limit for reliability at inspection time *t* with confidence level α, we have

$$P(R(t) \geq R_L(\alpha)) = \alpha. \tag{6.1}$$

(6.1) can be rewritten as $P(R_L(\alpha) \leq R(t)) = \alpha$ which means that $R(t)$ is the upper α quantile of $R_L(\alpha)$. Therefore the upper α quantile of the lower confidence limits, at confidence level α, is an estimator of the reliability $R(t)$.

Suppose that the true value of θ in (3.1) is 60. We consider the reliability when the survival time is $t = 3$ ($R(t) = 0.9512$). The confidence level α is set to be 0.95. All the indices except the confidence level are selected arbitrarily. Results are given in Table 1.

Because all the results in the fifth and sixth columns of Table 1 are satisfactory, the performance will be measured by two statistics: average estimate and coverage probability. It is observed that sampling frequency k has effects on MSE, smaller MSE along with lager k. The precision however mainly depends on the sample size n, higher precision along with larger n. In decision making, we pay more attention to sample size. Note that the result of sampling scheme $(n.k) = (500, 11)$ shows high degree of veracity. The coverage probability is almost the same as the confidence level. The upper α quantile of the lower confidence limits is also identical to the true reliability. The average estimate well coincides with the true value, with biasness less than 1.5%. The small MSE also demonstrates the accuracy of prediction.

6.2 Lifetime of Aluminum Coupon

The data set is reported by Birnbaum and Saunders (1958) and it represents the survival times (in circles) of aluminum coupon. Lee and Wang (2003) illustrated an application of the gamma distribution to these survival times. Upadhyay and Mukherjee (2010) used this data set to make a comparison between accelerated Weibull and accelerated Birnbaum-Saunders distributions. The 101 observations are listed in Table 2. The sample size and sampling times are set to be $(n, k) = (20, 5)$. The inspection time vector (in circles) is stochastically set to be $T = (t_1, t_2, \ldots, t_5) = (400, 800, 1200, \ldots, 2000)$. We conduct the sampling procedure described in Section 2 and obtain a grouped data set: $[(400, 20, 1), (800, 20, 2), (1200, 20, 5), (1600, 20, 13), (2000, 20, 18)]$.

We apply Weibull distribution to model this data set. Via QF algorithm, the estimates of scale and shape parameters are obtained as $(\hat{\alpha}_1, \hat{\beta}_1) = (155.44, 4.2921)$. The moment estimates, based on the original complete data, are $(\hat{\alpha}_2, \hat{\beta}_2) = (154.41, 4.1795)$. For comparison, the maximum likelihood estimates based on the original complete data are also calculated: $(\hat{\alpha}_3, \hat{\beta}_3) = (154.92, 3.9536)$. We can see that the scale parameter estimates are very close to each other, so are the shape parameter estimates. A graphical comparison of the original observed data and the fitted cumulative distribution functions is presented in Fig. 1, which shows very good agreement. This is corroborated by a chi-square test of goodness of fit which yields a probability level of 0.4387.

The reliability at time $t = 791$ is 0.95, based on the empirical distribution. Via the fitted Weibull distribution, the lower confidence limit for reliability (at time $t = 791$) is 0.9115 with confidence level $\alpha = 0.95$. The lower confidence limit is close to 0.95, showing the advantage of this method.

6.3 Lifetime of Deep-Groove Ball Bearing

The following 23 observations listed in Table 3 are used to illustrate the QF algorithm on lognormal distribution. This data set was originally given by Lieblein and Zelen (1956) for the lifetime (in millions of revolutions) of ball bearing. These data were later studied by Dumonceaux and Antle (1973), Pavur et al. (1992), and Upadhyay and Mukherjee (2008) etc.. The sample sizes are set to be $(n_1,\ n_2) = (11, 12)$, with sampling time $k = 2$. The inspection time vector (in millions of revolutions) is stochastically set to be $(t_1,\ t_2) = (50, 100)$. We conduct the sampling procedure described in Section 2 and obtain a grouped data set: $\{(50, 11, 6), (100, 12, 8)\}$.

We use lognormal distribution to model this data set. Via the QF algorithm, the estimates of location and scale parameters are obtained as $(\hat{\mu}_1, \hat{\sigma}_1) = (4.1243,\ 0.4809)$. The moment estimates, based on the original complete data, are $(\hat{\mu}_2, \hat{\sigma}_2) = (4.1505,\ 0.5334)$. We can see that the two location parameter estimates are almost equal. The relative bias of $\hat{\sigma}_1$ to $\hat{\sigma}_2$, namely $(\hat{\sigma}_2 - \hat{\sigma}_1)/\hat{\sigma}_2$, is 0.0984. Considering the characteristic of grouped data and the sample size, the QF algorithm performs quite well. A graphical comparison of the original observed data and the fitted cumulative distribution functions is presented in Fig. 2, which also shows very good agreement.

The reliability at time $t = 25.0560$ is 0.95, based on the empirical distribution. Via the fitted lognormal distribution, the lower confidence limit for reliability (at time $t = 25.0560$) is 0.9025 with confidence level $\alpha = 0.95$. This lower confidence limit is close to 0.95, again showing the advantage of this method.

7. Concluding Remarks

In this paper, we proposed an approach dealing with grouped data. This technique provides us with pseudo complete data based on a quantile filling (QF) approach in Yu and Dai (1996) and Yu and Guo (2001). The parameter estimates and lower confidence limit for reliability can then be obtained using standard methods for complete data. Three commonly used lifetime distribution models are applied to illustrate this approach. Simulation and real data examples are presented to demonstrate the feasibility and effectiveness of the QF algorithm. It is observed that the QF algorithm works quite well even when the sample size is small. We have observed that the group size, *n*, is more important in determining the estimation accuracy than the grouping number, *k*. Therefore we suggest to reduce the grouping number in order to augment the group size in small sample size situation.

It should be noted that QF algorithm is very flexible that it can be extended to different incomplete data types and different lifetime models. On model selection, we might assume different lifetime distributions for a given data set and conduct QF algorithm with each model. The ideal model is the one that gives the most precise predictions. The QF algorithm we suggested here belongs to single completion. Multiple completions can also be implemented. For example, we can combine conditional quantile with unconditional or conditional mean. These issues could be investigated later.

Acknowledgement: The work described in this paper was partially supported by a grant from City University of Hong Kong (Project No. 9380058), and a grant from National Center for Mathematics and Interdisciplinary Sciences, Chinese Academy of Sciences.

Reference

Andersen, P.K. and Perme, M.P., 2010. Pseudo-observations in survival analysis. Statistical Methods in Medical Research 19, 71-99.

Bassetti, F., Bodini, A. and Regazzini, E., 2007. Consistency of minimum divergence estimators based on grouped data. Statistics & Probability Letters 77, 937–941.

Birnbaum, Z.W. and Saunders, S.C., 1958. A statistical model for life-length of materials. Journal of the American Statistical Association 53, 151–160.

Buckley, J. and James, I., 1979. Linear regression with censored data. Biometricka 66, 429-436.

Davison, A.C. and Tsai, C.L., 1992. Regression model diagnostics. International Statistical Review 60, 337-353.

Dumonceaux, R. and Antle, C.E., 1973. Discriminating between the lognormal and Weibull distribution. Technometrics 15, 923-926.

Heard, A. and Pensky, M., 2006. Confidence intervals for reliability and quantile functions with application to NASA space flight data. IEEE Transactions on Reliability 55, 591-601.

Holan, S.H., Toth, D., Ferreira, M.A.R. and Karr, A.F., 2010. Bayesian multiscale multiple imputation with implications to data confidentiality. Journal of the American Statistical Association 105, 564–577.

Jiang, N.N., 2008. Quantile-Filling Algorithm for Weibull Distribution and its Application (in Chinese). Journal of Systems Science and Mathematical Sciences 28, 662-668.

Kabaila, P. and Leeb, H., 2006. On the large-sample minimal coverage probability of confidence intervals after model selection. Journal of the American Statistical Association 101, 619-629.

Kalbfleisch, J.D. and Prentice, R.L., 1973. Marginal likelihoods based on Cox's regression and life model. Biometrika 60, 267-278.

Lai, T.L. and Ying, Z., 1991b. Large sample theory of a modified Buckley-James estimator for regression analysis with censored data. Annals of Statistics 19, 1370-1402.

Lambert, P., 2011. Smooth semi- and nonparametric Bayesian estimation of bivariate densities from bivariate histogram data. Computational Statistics & Data Analysis 55, 429–445.

Lee, E.T. and Wang, J.W., 2003. Statistical methods for survival data analysis. New York: Wiley.

Lieblein, J. and Zelen, M., 1956. Statistical investigation of the fatigue life of deep-groove ball bearing. Journal of Research National Bureau of Standards 57, 273-316.

Liu, J. and Lindsay, B.G., 2009. Building and using semiparametric tolerance regions for parametric multinomial models. Annals of Statistics. 37, 3644–3659.

Liu, L.X., Murray, S. and Tsodikov, A., 2011. Multiple imputation based on restricted mean model for censored data. Statistics in Medicine 30, 1339-1350.

McKane, S.W., Escobar, L.A. and Meeker, W.Q., 2005. Sample size and number of failure requirements for demonstration tests with log-location-scale distributions and failure censoring. Technometrics 47, 182-190.

Meister, A., 2007. Optimal convergence rates for density estimation from grouped data. Statistics and Probability Letters 77, 1091–1097.

Murthy, D.N.P., Xie, M. and Jiang, R., 2003. Weibull Models. New York: Wiley.

Pavur, R.J., Edgeman, R.L. and Scott, R.C., 1992. Quadratic statistics for the goodness-of-fit test of the inverse Gaussian distribution. IEEE Transactions on Reliability 41, 118-123.

Reiter, J.P., 2007. Small-sample degrees of freedom for multi-component significance tests for multiple imputation for missing data. Biometrika 94, 502–508.

Reiter, J.P. and Raghunathan, T.E., 2007. The multiple adaptations of multiple imputation. Journal of the American Statistical Association 102, 1462–1471.

Rivero, C. and Valdes, T., 2008. An algorithm for robust linear estimation with grouped data. Computational Statistics & Data Analysis 53, 255–271.

Rubin, D.B., 1976. Inference and missing data. Biometrika 63, 581-590.

Ryan, A.G., Wells, L.J. and Woodall, W.H., 2011. Methods for monitoring multiple proportions when inspecting continuously. Journal of Quality Technology 43, 237–248.

Upadhyay, S.K. and Mukherjee, B., 2008. Assessing the value of the threshold parameter in the Weibull distribution using Bayes paradigm. IEEE Transactions on Reliability 57, 489-497.

Upadhyay, S.K. and Mukherjee, B., 2010. Bayes analysis and comparison of accelerated Weibull and accelerated Birnbaum-Saunders models. Communications in Statistics-Theory and Methods 39, 195-213.

Wang, H., 2008. Coverage probability of prediction intervals for discrete random variables. Computational Statistics & Data Analysis 53, 17-26.

Wang, S., Nan, B., Zhu, J. and Beer, D., 2008. Doubly penalized Buckley-James method for survival data with high-dimensional covariates. Biometrics 64, 132-140.

Weerahandi, S., 1993. Generalized confidence intervals. Journal of the American Statistical Association 88, 899-905.

Yang, M.S. and Yu, N.Y., 2005. Estimation of parameters in latent class models using fuzzy clustering algorithms. European Journal of Operational Research 160, 515-531.

Yu, D. and Dai, S.S., 1996. Study on the synthetic evaluation method for the storage reliability of missile system (in Chinese). Academy of Mathematics and System Science, Chinese Academy of Science. Research Report.

Yu, D., 2010. Data transformation (in Chinese). Academy of Mathematics and System Science, Chinese Academy of Science. Research Report.

Yu, D. and Guo, K., 2001. Research on long-life satellite’s electromechanical equipment reliability evaluation (in Chinese). Academy of Mathematics and System Science, Chinese Academy of Science. Research Report.