Vol. 34 (2014)
No. 5

数 学 杂 志
J. of Math. (PRC)

EMPIRICAL LIKELIHOOD INFERENCE FOR QUANTILES WITH RANDOM RIGHT CENSORED DATA

LIU Chang-sheng, LI Yong-xian

(*Department of Mathematics and Physics, Henan University of Urban Construction, Pingdingshan 467036, China*)

Abstract: This paper studies the construction of confidence region for quantiles with random right censored variables. By combining empirical likelihood method and censored values estimation method, we obtain an empirical log-likelihood ratio statistics about quantiles. In weaker conditions, that the limiting distribution of the statistics is Chi-square distribution with one degree of freedom is proved. The makes the inference of empirical likelihood for quantiles from complete data to incomplete data.

Keywords: empirical likelihood; right censored; confidence interval; Chi-square distribution

2010 MR Subject Classification: 62F25; 62G20

Document code: A **Article ID:** 0255-7797(2014)05-0849-07

1 Introduction

Quantile is an important population characteristic. In some instances the quantile approach is feasible and useful when other approaches are out of the question. For example, to estimate the parameter of a Cauchy distribution, with density $f(x) = 1/\pi[1 + (x - \mu)^2], -\infty < x < \infty$, the sample mean $\overline{X}$ is not a consistent estimate of the location parameter μ. However, the sample median $\theta_{1/2}$ is $AN(\mu, \pi^2/4n)$ and thus quite well-behaved.

Let $X_1, X_2, \cdots, X_n$ be a random sample from the unknown distribution $F(x)$ with density $f(x)$. Given $0 < q < 1$, we define the q-th quantile by $F^{-1}(q) = \inf\{x : F(x) > q\}$.

In this paper, we investigate how to apply empirical likelihood methods for inference about $\theta_q = F^{-1}(q)$ under right censorship. Assume that the variable X is censored randomly on the right by some censoring variable C and hence cannot be observed completely. One observes only

$$Y = \min\{X, C\},\ \delta = I(X \le C), \tag{1.1}$$

where $I(A)$ is the indicator function of event A. Supposed that, C is independent of X. The observations are $\{Y_i, \delta_i\}_{i=1}^n$, which is a random sample from the population (Y, δ).

* **Received date:** 2012-08-31 **Accepted date:** 2013-06-04

Foundation item: Supported by Science and Technology Project of Henan Province (112300410191).

Biography: Liu Changsheng(1976–), male, born at Luohe, Henan, lecturer, major in mathematics and statistics. E–mail:csliu@hncj.edu.cn.

Empirical likelihood methods were first used by Thomas and Grunkemeier [1] and popularized by Owen [2–3]. It is well-known that Owen's empirical likelihood is based on linear constraints and hence has very general applicability such as in smooth functions of means (see DiCiccio et al. [4]), quantile estimation (see Chen [5]), estimating equation (see Qin and Lawless [6]), empirical likelihood confidence interval (see [7–16]) and so on. For more details, we refer to Owen [17]. However, most of the references on empirical likelihood are concerned with complete data set. In practice, censoring data occurs in opinion polls, market research surveys, mail enquires, social-economic investigations, medical studies and other scientific experiments. Once the censoring values are imputed, the data set can then be analyzed using standard techniques for complete data.

The rest of this paper is arranged as follows. In Section 2, we propose a empirical likelihood method to quantiles. We obtain the empirical log-likelihood ratio statistics of the quantiles and show that it is asymptotically chi-square. The proof is arranged to Section 3.

2 Methodology and Main Results

If θ_q is q-quantile for $F(x)$, we know that θ_q coincides with the M-estimates defined by the equation

$$E[\phi(X-\theta_q)] = \int_{-\infty}^{\infty} \phi(x-\theta_q)dF(x) = 0 \tag{2.1}$$

with

$$\phi(z) = \begin{cases} -1, & z \le 0, \\ q/(1-q), & z > 0. \end{cases}$$

Let $U_i(\theta_q) = \frac{\phi(Y_i-\theta_q)\delta_i}{1-G(Y_i)}$, where $G(\cdot)$ is the cumulative distribution function of the censoring variable C. Obviously $\{U_i(\theta_q)\}_{i=1}^n$ are independent and identically distributed random variables. Furthermore,

$$E(U_i(\theta_q)) = E\left[\frac{\phi(Y_i-\theta_q)\delta_i}{1-G(Y_i)}\right] = E[\phi(X-\theta_q)] = 0.$$

Thus, following the idea of Owen [2] an empirical likelihood-ratio function can similarly be defined for θ_q as

$$R(\theta_q) = \sup_{p_1,\cdots,p_n}\left\{\prod_{i=1}^n (np_i) : \sum_{i=1}^n p_i = 1, \sum_{i=1}^n p_i U_i(\theta_q) = 0\right\}. \tag{2.2}$$

However, we cannot use $R(\theta_q)$ directly to make inference on θ_q since the distribution function $G(\cdot)$ of $\{Z_i\}_{i=1}^n$ is unknown. To do this, it is natural to replace $G(\cdot)$ by the Kaplan-Meier estimator

$$\hat{G}_n(y) = 1 - \prod_{i=1}^n \left[\frac{n-i}{n-i+1}\right]^{I\{Y_{(i)}\le y, \delta_i=0\}}, \tag{2.3}$$

where $Y_{(1)} \le Y_{(2)} \le \cdots \le Y_{(n)}$ is the order statistics of Y_i's. Let $\hat{U}_i(\theta_q) = \frac{\phi(Y_i-\theta_q)\delta_i}{1-\hat{G}_n(Y_i)}$, then an estimated empirical log-likelihood ratio function can be defined as

$$\log \hat{R}(\theta_q) = \sup_{p_1,\cdots,p_n} \left\{ \sum_{i=1}^{n} \log(np_i) : \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i \hat{U}_i(\theta_q) = 0 \right\}. \tag{2.4}$$

By the method of Lagrange multiplier for (2.4), we may prove that the maximization point occurs with

$$p_i = \frac{1}{n} \{1 + \lambda(\theta_q)\hat{U}_i(\theta_q)\}^{-1}, \ \ i = 1, \cdots, n, \tag{2.5}$$

where $\lambda(\theta_q)$ is the solution to

$$\frac{1}{n} \sum_{i=1}^{n} \frac{\hat{U}_i(\theta_q)}{1 + \lambda(\theta_q)\hat{U}_i(\theta_q)} = 0. \tag{2.6}$$

By (2.4) and (2.5), we can obtain

$$\log \hat{R}(\theta_q) = -\sum_{i=1}^{n} \log\{1 + \lambda(\theta_q)\hat{U}_i(\theta_q)\}. \tag{2.7}$$

Theorem 2.1 If $E(U_i^2(\theta_q)) < \infty$ and θ_q is the true q-quantile of $F(\cdot)$, we have

$$-2 \log \hat{R}(\theta_q) \longrightarrow^{L} \chi_1^2, \tag{2.8}$$

where $\rightarrow^{L}$ represents the convergence in distribution, χ_1^2 is standard Chi-square random variable with 1 degree of freedom.

Remark On the basis of Theorem 2.1, $-2 \log \hat{R}(\theta_q)$ can be used to construct a confidence region for θ_q,

$$\hat{I}_\alpha(\theta_q) = \{\theta_q : -2 \log \hat{R}(\theta_q) \le c_\alpha\},$$

with $P(\chi_1^2 \le c_\alpha) = 1 - \alpha$. Then by Theorem 2.1, $\hat{I}_\alpha(\theta_q)$ gives a confidence interval for θ_q with asymptotically correct coverage probability $1 - \alpha$.

3 Proof of Theorem 2.1

Throughout this section, we use $c > 0$ to represent any constant which may take different values for each appearance.

Lemma 3.1 Under the assumptions of Theorem 2.1, if θ_q is the true q-quantile of $F(\cdot)$, we have

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \hat{U}_i(\theta_q) \longrightarrow^{L} N(0, v_q^2), \tag{3.1}$$

where $v_q^2 = E(U_i^2(\theta_q))$.

Proof By the definition of $\hat{U}_i(\theta_q)$, it is easy to show that

$$\begin{aligned}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{U}_i(\theta_q) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}U_i(\theta_q)+\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\hat{U}_i(\theta_q)-U_i(\theta_q))\\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}U_i(\theta_q)+\frac{1}{\sqrt{n}}\sum_{i=1}^{n}U_i(\theta_q)\frac{\hat{G}_n(Y_i)-G_n(Y_i)}{1-\hat{G}_n(Y_i)}.\end{aligned}$$

Since $\{U_i(\theta_q)\}_{i=1}^n$ are independent and identically distributed random variables and $E(U_i^2(\theta_q)) < \infty$. They imply $\frac{1}{\sqrt{n}}\sum\limits_{i=1}^{n}|U_i(\theta_q)| = O(1)$. Zhou [18] proved

$$\sup_{Y_i\le Y_{(n)}}\left|\frac{\hat{G}_n(Y_i)-G_n(Y_i)}{1-\hat{G}_n(Y_i)}\right| = O_p(n^{-1/2}), \tag{3.2}$$

where $Y_{(n)} = \max\limits_{1\le i\le n} Y_i$. So we have

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}U_i(\theta_q)\frac{\hat{G}_n(Y_i)-G_n(Y_i)}{1-\hat{G}_n(Y_i)} = o_p(1).$$

By the central limit theory of independent and identically distributed random variables,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}U_i(\theta_q) \longrightarrow^{L} N(0, v_q^2).$$

This completes the proof.

Lemma 3.2 Under the assumptions of Theorem 2.1, if θ_q is the true q-quantile of $F(\cdot)$, we have

$$\frac{1}{n}\sum_{i=1}^{n}\hat{U}_i^2(\theta_q) \longrightarrow^{P} v_q^2. \tag{3.3}$$

Proof By the definition of $\hat{U}_i(\theta_q)$, it is easy to show that

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\hat{U}_i^2(\theta_q) &= \frac{1}{n}\sum_{i=1}^{n}U_i^2(\theta_q)+\frac{1}{n}\sum_{i=1}^{n}(\hat{U}_i(\theta_q)-U_i(\theta_q))^2\\ &\quad+\frac{2}{n}\sum_{i=1}^{n}U_i(\theta_q)(\hat{U}_i(\theta_q)-U_i(\theta_q)) =: I_1+I_2+I_3.\end{aligned}$$

Next, from the law of large numbers and (3.2), we can get $n^{-1}\sum\limits_{i=1}^{n}U_i^2(\theta_q) \to E(U_i^2(\theta_q))$, a.s., then we have

$$I_2 = \frac{1}{n}\sum_{i=1}^{n}(\hat{U}_i(\theta_q)-U_i(\theta_q))^2 = \frac{1}{n}\sum_{i=1}^{n}U_i^2(\theta_q)\left|\frac{\hat{G}_n(Y_i)-G_n(Y_i)}{1-\hat{G}_n(Y_i)}\right|^2 = o_p(1).$$

By using the similar arguments as $I_2 = o_p(1)$, we can also obtain $I_3 = o_p(1)$. So, the law of large numbers implies that

$$\frac{1}{n}\sum_{i=1}^{n}\hat{U}_i^2(\theta_q) = \frac{1}{n}\sum_{i=1}^{n}U_i^2(\theta_q) + o_p(1) \longrightarrow^p v_q^2.$$

This completes the proof.

Lemma 3.3 Under the assumptions of Theorem 2.1, if θ_q is the true q-quantile of $F(\cdot)$, we have

$$\max_{1\le i\le n}|\hat{U}_i(\theta_q)| = o_p(n^{1/2}), \quad \lambda(\theta_q) = O_p(n^{-1/2}). \tag{3.4}$$

Proof It is well known that for any sequence of independent and identically distributed random variables $\{\xi\}_{i=1}^n$ with $E(\xi_i^2) < \infty$, we have

$$\max_{1\le i\le n}\frac{|\xi_i|}{\sqrt{n}} \longrightarrow 0.$$

This implies that $\max\limits_{1\le i\le n}|U_i(\theta_q)| = o_p(n^{1/2})$. From (3.2), we have

$$\max_{1\le i\le n}|\hat{U}_i(\theta_q)| = \max_{1\le i\le n}|U_i(\theta_q)|\cdot\left|\frac{\hat{G}_n(Y_i)-G_n(Y_i)}{1-\hat{G}_n(Y_i)}\right| + \max_{1\le i\le n}|U_i(\theta_q)| = o_p(n^{1/2}).$$

Next, we prove $\lambda(\theta_q) = O_p(n^{-1/2})$. Let $\lambda(\theta_q) = \alpha|\lambda(\theta_q)|$, where $\alpha = 1$ or -1. Note $\bar{\Lambda} = \frac{1}{n}\sum\limits_{i=1}^{n}\hat{U}_i(\theta_q)$, $\Lambda^* = \max\limits_{1\le i\le n}|\hat{U}_i(\theta_q)|$, $S = \frac{1}{n}\sum\limits_{i=1}^{n}\hat{U}_i^2(\theta_q)$. From (2.6), we have

$$\begin{aligned}
0 &= \frac{1}{n}\sum_{i=1}^{n}\frac{\hat{U}_i(\theta_q)}{1+\lambda(\theta_q)\hat{U}_i(\theta_q)} = \frac{1}{n}\sum_{i=1}^{n}\frac{\alpha\hat{U}_i(\theta_q)}{1+|\lambda(\theta_q)|\alpha\hat{U}_i(\theta_q)}\\
&= \frac{1}{n}\sum_{i=1}^{n}\alpha\hat{U}_i(\theta_q) - |\lambda(\theta_q)|\cdot\frac{1}{n}\sum_{i=1}^{n}\frac{\hat{U}_i^2(\theta_q)}{1+|\lambda(\theta_q)|\alpha\hat{U}_i(\theta_q)}\\
&\le \alpha\bar{\Lambda} - \frac{|\lambda(\theta_q)|}{1+|\lambda(\theta_q)|\Lambda^*}\cdot\frac{1}{n}\sum_{i=1}^{n}\hat{U}_i^2(\theta_q) = \alpha\bar{\Lambda} - \frac{|\lambda(\theta_q)|}{1+|\lambda(\theta_q)|\Lambda^*}\cdot S,
\end{aligned}$$

where we have used $0 < 1+\lambda(\theta_q)\hat{U}_i(\theta_q) \le 1+|\lambda(\theta_q)|\Lambda^*$, which yields from

$$p_i = \frac{1}{n}(1+\lambda(\theta_q)\hat{U}_i(\theta_q))^{-1} \ge 0.$$

Therefore, $|\lambda(\theta_q)|(S-\alpha\bar{\Lambda}\Lambda^*) \le |\alpha\bar{\Lambda}|$. By Lemma 3.1 and Lemma 3.2, we know $\Lambda^* = o_p(n^{1/2})$ and $\Lambda = O_p(n^{-1/2})$, we have $|\lambda(\theta_q)|(S+o_p(1)) \le |\alpha\bar{\Lambda}|$. Lemma 3.2 implies that $S = \Sigma+o_p(1)$, hence $|\lambda(\theta_q)| = O_p(n^{-1/2})$. This completes the proof.

Proof of Theorem 2.1 Applying a Taylor expansion to equation (2.6) and (2.7), we can obtain

$$\begin{aligned}
-2\log\hat{R}(\theta_q) &= 2\sum_{i=1}^{n}\log\{1+\lambda(\theta_q)\hat{U}_i(\theta_q)\}\\
&= 2\lambda(\theta_q)\sum_{i=1}^{n}\hat{U}_i(\theta_q) - \lambda^2(\theta_q)\sum_{i=1}^{n}\hat{U}_i^2(\theta_q) + r_n,
\end{aligned} \tag{3.5}$$

where

$$\begin{aligned} |r_n| &= C\sum_{i=1}^{n}|\lambda(\theta_q)\hat{U}_i(\theta_q)|^3 \le |\lambda(\theta_q)|^3 \max_{1\le i\le n}|\hat{U}_i(\theta_q)|\sum_{i=1}^{n}\hat{U}_i^2(\theta_q) \\ &= O_p(n^{-3/2})\cdot o_p(n^{1/2})\cdot O_p(n^{-1}) = o_p(1). \end{aligned}$$

From (2.6), we can get

$$\begin{aligned} 0 &= \sum_{i=1}^{n}\frac{\hat{U}_i\theta_q)}{1+\lambda(\theta_q)\hat{U}_i(\theta_q)} = \sum_{i=1}^{n}\left\{1-\lambda(\theta_q)\hat{U}_i(\theta_q)+\frac{[\lambda(\theta_q)\hat{U}_i(\theta_q)]^2}{1+\lambda(\theta_q)\hat{U}_i(\theta_q)}\right\}\hat{U}_i(\theta_q) \\ &= \sum_{i=1}^{n}\hat{U}_i(\theta_q)-\lambda(\theta_q)\sum_{i=1}^{n}\hat{U}_i^2(\theta_q)+\sum_{i=1}^{n}\frac{\lambda^2(\theta_q)\hat{U}_i^3(\theta_q)}{1+\lambda(\theta_q)\hat{U}_i(\theta_q)}. \end{aligned} \tag{3.6}$$

From Lemma 3.2 and Lemma 3.3, by simple calculation, we have

$$\sum_{i=1}^{n}\frac{\lambda^2(\theta_q)\hat{U}_i^3(\theta_q)}{1+\lambda(\theta_q)\hat{U}_i(\theta_q)} = o_p(n^{1/2}).$$

It follows that

$$\lambda(\theta_q) = \frac{\sum\limits_{i=1}^{n}\hat{U}_i(\theta_q)}{\sum\limits_{i=1}^{n}\hat{U}_i^2(\theta_q)} + o_p(n^{-1/2}).$$

Furthermore, from (3.6), we can get that

$$\lambda(\theta_q)\sum_{i=1}^{n}\hat{U}_i(\theta_q)-\lambda^2(\theta_q)\sum_{i=1}^{n}\hat{U}_i^2(\theta_q) = o_p(1). \tag{3.7}$$

By (3.5), (3.6) and (3.7), we have

$$\begin{aligned} -2\log\hat{R}(\theta_q) &= \lambda(\theta_q)\sum_{i=1}^{n}\hat{U}_i(\theta_q)+o_p(1) \\ &= \left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{U}_i(\theta_q)\right)^2\times\left(\frac{1}{n}\sum_{i=1}^{n}\hat{U}_i^2(\theta_q)\right)^{-1}+o_p(1)\to^L\chi_1^2. \end{aligned}$$

This completes the proof.

References

[1] Thomas D R, Grunkemeier G L. Confidence interval estimation of survival probabilities for censored data[J]. Journal of the American Statistical Association, 1975, 70(352): 865–871.

[2] Owen A B. Empirical likelihood ratio confidence intervals for a single function[J]. Biometrika, 1988,75(2): 237–249.

[3] Owen A. Empirical likelihood ratio confidence regions[J]. The Annals of Statistics, 1990, 18(1): 90–120.

[4] DiCiccio T, Hall P, Romano J. Empirical likelihood is Bartlett-correctable[J]. The Annals of Statistics, 1991, 19(2): 1053–1061.

[5] Chen S X, Hall P. Smoothed empirical likelihood confidence intervals for quantiles[J]. The Annals of Statistics, 1993, 21: 1166–1181.

[6] Qin J, Lawless J F. Empirical likelihood and general estimating equations[J]. The Annals of Statistics, 1994, 22(1): 300–325.

[7] DiCiccio T J, Romano J P. Nonparametric confidence limits by resampling methods and least favorable families[J]. Int Statist Rev., 1990, 58: 59–76.

[8] Jing B Y, Wood A T A. Exponential empirical likelihood is not Bartlett correctable[J]. The Annals of Statistics, 1996, 24(1): 365–369.

[9] Xue L G, Zhu L. Empirical likelihood for single-index models[J]. Journal of Multivariate Analysis, 2006, 97(6): 1295–1312.

[10] Xue L G, Zhu L X. Empirical likelihood for a varying coefficient model with longitudinal data[J]. Journal of the American Statistical Association, 2007,102(478): 642–654.

[11] Zhu L X, Xue L G. Zhu L, Xue L. Empirical likelihood confidence regions in a partially linear single-index model[J]. Journal of the Royal Statistical Society: Series B, 2006, 68(3): 549–570.

[12] Wu C. Some algorithmic aspects of the empirical likelihood method in survey sampling[J]. Statistica Sinica, 2004, 14(4): 1057–1067.

[13] Shi J, Lau T S. Shi J, Lau T S. Empirical likelihood for partially linear models[J]. Journal of Multivariate Analysis, 2000, 72(1): 132–148.

[14] Wang Q H, Jing B Y. Empirical likelihood for partial linear models[J]. Annals of the Institute of Statistical Mathematics, 2003, 55(3): 585–595.

[15] Zhou M. Some properties of the Kaplan-Meier estimator for independent nonidentically distributed random variables[J]. The Annals of Statistics, 1991, 19: 2266–2274.

[16] Cheng P E. Nonparametric estimation of mean functionals with data missing at random[J]. Journal of the American Statistical Association, 1994, 89(425): 81–87.

[17] Owen A B. Empirical likelihood[M]. London: Chapman and Hall/CRC, 2010.

[18] Zhou M. Asymptotic normality of the synthetic estimator for censored survival data[J]. The Annals of Statistics, 1992, 20(2): 1002–1021.

具有随机右删失随机变量分位数的经验似然推断

刘常胜 , 李永献

(河南城建学院数理系, 河南 平顶山 467036)

摘 要: 本文研究了具有随机右删失随机变量分位数的置信域的构造. 利用经验似然和截尾值估算相结合的方法, 给出了分位数的对数经验似然比统计量, 在较少的条件下证明了该统计量的极限分布为自由度为1 的χ^2 分布. 使得完全数据下的分位数的经验似然推断方法应用到非完全数据中.

关键词: 经验似然; 右删失; 置信区间; χ^2分布

MR(2010)**主题分类号**: 62F25; 62G20 **中图分类号**: O212.1