INVESTIGATION OF THE ESTIMATES PROPERTIES AND GOODNESS-OF-FIT TEST STATISTICS FROM CENSORED SAMPLES WITH COMPUTER MODELING TECHNIQUE[1]

E.V. Chimitova, B.Yu. Lemeshko
Novosibirsk State Technical University
Novosibirsk, Russia
e-mail: chim@mail.ru

Abstract

The unbiased estimation of distribution parameters and goodness-of-fit hypothesis testing on the basis of strongly censored samples have been considered. The empirical bias corrections of maximum likelihood estimates (MLE) have been obtained for strongly censored samples, a number of distribution laws being under consideration. The investigation of Renyi and Kolmogorov statistic distributions has been carried out in the case of simple and composite hypothesis testing. For simple hypotheses the convergence rate of statistic distributions to the limiting distribution laws has been analysed depending on the censoring degree. For composite hypotheses the limiting statistic distribution laws become dependent on the distribution under test, parameter estimation method and type of parameters estimated. The approximations of the limiting statistic distributions for a number distribution laws under test have been developed in case of using MLE for unknown parameters.

1 Introduction

Statistical analysis of obtained observations is necessary almost in every field of research, connected with data registration. The problem of censored data processing often appears in life time data analysis, for example, in reliability tasks or in medical and biological investigations. It may follow from experiment time limitation (the first-type censoring) or limitating the number of observed failures (the second-type censoring). Under the censoring degree a is implied the probability to fall into the censoring interval in case of the first type or the ratio of the number of censored observations to the complete sample size in case of the second type.

In [1] and later papers we investigated Fisher information losses for different distributions depending on the censoring degree. For certain distributions a censored sample was shown to keep quite a lot of information, when the observed part of random variable domain being very small. So, for the exponential distribution law a censored sample contains more than 52% of complete sample information, the left censoring degree being as follows: $a = 0.95$. So it may be possible to determine good estimates of unknown parameters as well as to test goodness-of-fit of some theoretical distribution law to the empirical distribution by analysing strongly censored samples rather confidently.

[1] The research is supported by the Ministry of Education of the Russian Federation (project No. T02-3.3-3356)

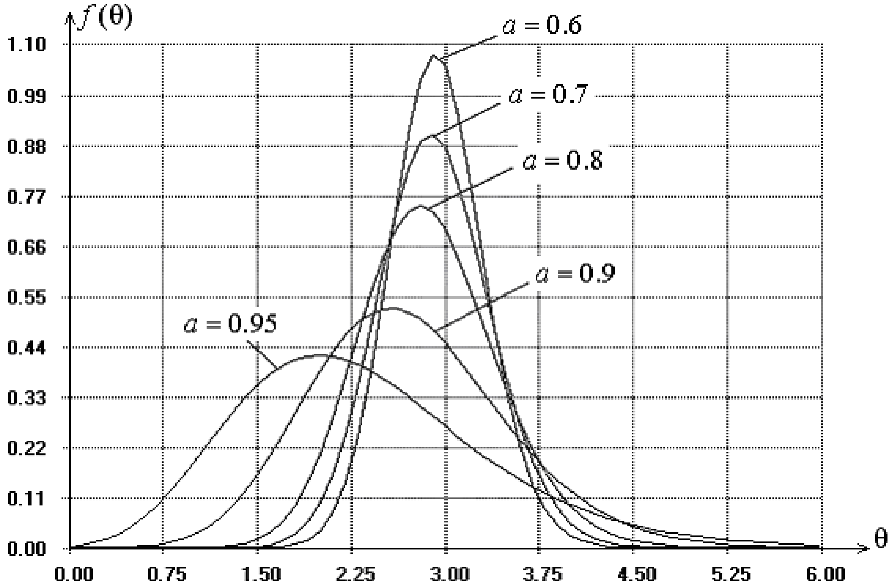

Figure 1: MLE density functions for the scale parameter of lognormal distribution for different censoring degerees a and the full sample size $n = 100$

2 The investigation of the MLE for censored samples

The maximum likelihood estimate (MLE) of unknown parameter by the left and right censored sample is the solution of the following system of likelihood equations

$$n_{(l)} \frac{\partial \ln P_{(l)}(\theta)}{\partial \theta_i} + \sum_{j=1}^{n-n_{(l)}-n_{(r)}} \frac{\partial \ln f(X_j, \theta)}{\partial \theta_i} + n_{(r)} \frac{\partial \ln P_{(r)}(\theta)}{\partial \theta_i} = 0, \;\; i = \overline{1, m}, \qquad (1)$$

where $P_{(l)}(\theta) = \int\limits_{-\infty}^{x_{(l)}} f(x, \theta) dx, P_{(r)}(\theta) = \int\limits_{x_{(r)}}^{\infty} f(x, \theta) dx, x_{(l)} < x_{(r)}, n_{(r)}$ and $n_{(l)}$ is the number of observations, fallen into the right and left censoring intervals correspondingly. In case of right censoring (or left) the first (or the third) item is eliminated from (1).

ML-estimates of distribution parameters by censored samples are asymptotically effective. But for the limited sample size and considerable censoring degree MLE distributions are far from the normal distribution law. Moreover, they turn out to be asymmetric and the estimates themselves become biased. For example, figure 1 illustrates the form of $\hat{\theta}_n^c$ estimate densities for the scale parameter of the lognormal distribution by analysing the right censored samples of size $n = 100$. The lognormal distribution has been simulated with the scale parameter equal to 3. With decreasing of n and increasing of the censoring degree the asymmetry of $\hat{\theta}_n^c$ estimate distribution increases.

By means of statistical regularities simulation technique the bias of distribution parameter ML-estimates has been investigated depending on the sample size n and the censoring degree a. The distribution laws frequently used in "life time" data analysis, such as lognormal, exponential, gamma, Rayleigh, Weibull distributions have been considered. The estimates of mathematical expectation of relative MLE biases in the form of empirical functions of sample size and censoring degree have been obtained. The bias of shift parameter MLE has been shown to be directly proportional to the value of the scale parameter when estimating both shift and scale parameters; estimating both scale and form parameters provides the scale parameter MLE bias to be in inverse proportion to the value of the form parameter.

If the values obtained are used as corrections the estimate bias essentially decreases.

3 The investigation of the Renyi and Kolmogorov goodness-of-fit tests

For testing goodness-of-fit of the theoretical distribution $F(x)$ to the empirical one $F_n(x)$ by censored data are used the Renyi test and the Kolmogorov test [2].

The two-sided statistic of the Renyi test for left censoring is represented by $S_R^c = \sqrt{\frac{na}{1-a}} \cdot \sup\limits_{F(x) \geq a} \frac{|F_n(x)-F(x)|}{F(x)}$, and in case of right censoring: $S_R^c = \sqrt{\frac{na}{1-a}} \cdot \sup\limits_{F(x) \leq 1-a} \frac{|F_n(x)-F(x)|}{1-F(x)}$, where $a \in (0,1)$ is the censoring degree.

The Kolmogorov statistic for censored data is defined by $S_K^c = \sup\limits_{M} |F_n(x) - F(x)|$, where $M = \{x : F(x) \geq a\}$ for left censoring and $M = \{x : F(x) \leq 1-a\}$ for right censoring. The limiting distributions of these statistics for simple goodness-of-fit hypotheses are given in [2].

Unlike the Kolmogorov statistic distribution for censored samples, the limiting distribution of the Renyi statistic doesn't depend on the censoring degree a and thus theoretically it is more convenient to be used in practice. However, the problem of the rate of statistic distribution convergence to the corresponding limiting distribution laws is still urgent.

The results of statistical simulation have shown the Renyi statistic distributions to be essentially dependent on the censoring degree and type. For example, figure 2 illustrates the empirical distributions of the Renyi statistic in case of simple hypothesis testing, their goodness-of-fit to the exponential law for the censoring degree $a = 0.9$ being considered.

As figure 2 shows, the Renyi statistic distributions essentially differ from the limiting law for the sample size $n < 1000$ and $a = 0.9$. The investigation of Renyi statistic distributions depending on the censoring degree has shown that the best goodness-of-fit to the limiting law $L(S)$ is achieved for $a = 0.5$. For small or, on the contrary, high censoring degrees the empirical statistic distributions essentially differ from $L(S)$.

In contrast to the Renyi statistic, Kolmogorov statistic distributions have shown their good convergence to the corresponding limiting functions. The minimal sample sizes providing a good agreement of empirical distributions of the Kolmogorov statistic

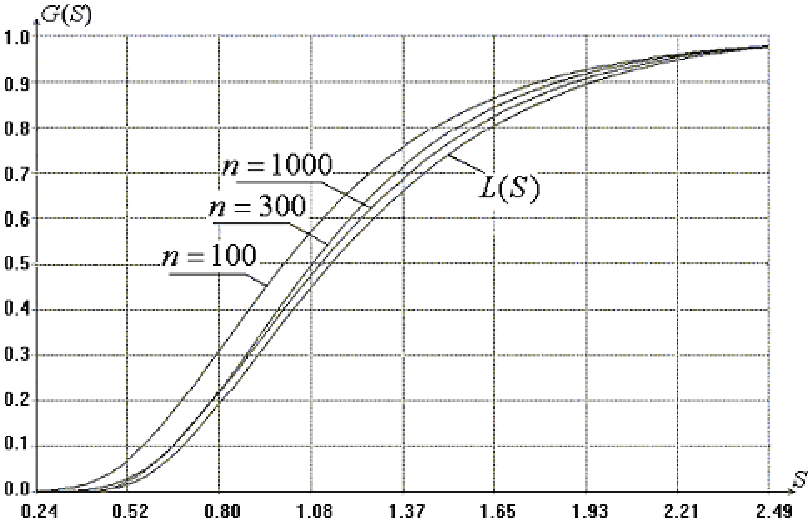

Figure 2: Distributions of S_R^c statistic, $a = 0.9$, left censoring

with the corresponding limiting laws have been obtained for different censoring degrees.

It should be pointed out that these criteria are intended for testing simple hypotheses H_0. But in practice of statistical data analysis it is frequently needed to test goodness-of-fit after estimating parameters of the assumed distribution law by analysing the same sample. And in this case (i.e. composite hypothesis testing) nonparametric goodness-of-fit tests become "distribution-dependent".

By means of computer simulation technique the approximations of the limiting Kolmogorov test statistic distributions have been developed, MLE for censored samples being used. Different distribution laws frequently used in life-time data analysis have been considered as a null hypothesis H_0. The limiting distributions have been approximated with the family of lognormal distributions for different censoring degrees.

References

[1] B.Yu. Lemeshko, S.Ya. Gildebrant, S.N. Postovalov. (1998). On the estimation of distribution parameters by analysing incomplete samples. *Proceedings of IV International Conference "Actual Problems of Electronics and Instrument Engineering" APEIE-98.* Vol.**160**, pp. 17-23.

[2] Mania G.M. (1974). *Statistical estimation of distributions.* Publishing house of TSU, Tbilisi. (in Russian)