The Korean Communications
in Statistics Vol. 4, No. 1, 1997
pp. 311-315

# Confidence Intervals for Distribution Function

**Choi, J. R.[1], Kang, M. K.[2] and Chu, I. S.[3]**

## Abstract

In this note we consider confidence interval based on Kolmogorov-Smirnov statistic. In order to obtain confidence interval we need percentage points of the statistics. Bootstrap method is examined whether it is useful to determine the points. It is concluded that the method is useful for observations with many ties, whereas it gives less conserbative points for continuous distributions.

## 1. Introduction

Let $X_1, \cdots, X_n$ be a random sample with distribution function $F(x)$. Suppose we want to construct confidence interval for $F(x)$. The confidence interval plays important role in estimating percentage points of $F(x)$ and in goodness-of-fit test.

In this paper, we discuss the construction of confidence interval for $F(x)$ and the role of bootstrap in the confidence interval for $F(x)$ of a population for simple random sampling.

It is reasonable to construct a confidence interval around an adequate estimator of the function $F(x)$. As an estimator of $F(x)$, it is shown by many authors that a kernal-type estimator

$$\widehat{F_n}(x) = \frac{1}{n} \sum_{i=1}^{n} K(\frac{x - X_i}{h_n}),$$

where $K(x)$ is a known distribution function, is better than the useful empirical distribution function

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} u(x - X_i)$$

in the sense of mean integrated squared error(MISE). Here $u(x) = 1$ or 0 according as $x \geq 0$ or $x < 0$.

1) Professor, Department of Mathematics, Dong-A University, Pusan 604-714, Korea
2) Professor, Department of Computer Science and Statistics, Dong-Eui University, Pusan 614-714, Korea.
3) Lecturer, Department of Mathematics, Dong-A University, Pusan 604-714, Korea

However, it is shown in Shirahata and Chu(1992) that $\widehat{F_n(x)}$ and $F_n(x)$ are asymptotically equivalent in the sense of integrated squared error, because the criterion MISE ignores random term and compare them between only higher order terms. Furthermore, the kernel-type estimator can not be applied for discrete distributions. Hence, we adopt $F_n(x)$ as an estimator of $F(x)$.

## 2. Confidence Intervals

A confidence interval is usually constructed based on Kolmogorov-Smirnov statistic

$$K_n = n \sup_{-\infty<x<\infty} \{F_n(x)-F(x)\}^2$$

and a $100(1-\alpha)\%$ confidence interval for $F(x)$ is given by

$$\left(F_n(x)-\frac{k_n(\alpha)}{n},\ F_n(x)+\frac{k_n(\alpha)}{n}\right) \tag{1}$$

where $k_n(\alpha)$ is determined from

$$k_n(\alpha) = \inf\{k : P(K_n > k) < \alpha\}.$$

However, the width of interval (1) is constant and seems to be too wide for $x$ with $F(x)$ being close to 1 or 0.

In order to overcome this property let us consider a generalized Kolmogorov-Smirnov statistic

$$A_n = n \sup_{-\infty<x<\infty} \frac{\{F_n(x)-F(x)\}^2}{F(x)(1-F(x))},$$

and a $100(1-\alpha)\%$ confidence interval for $F(x)$ is given by

$$\left(\frac{F_n(x)+\frac{1}{2n}a_n(\alpha)-C_n}{1+\frac{1}{n}a_n(\alpha)},\ \frac{F_n(x)+\frac{1}{2n}a_n(\alpha)+C_n}{1+\frac{1}{n}a_n(\alpha)}\right) \tag{2}$$

where $a_n(\alpha)$ is determined from

$$a_n(\alpha) = \inf\{a : P(A_n > a) < \alpha\}$$

and

$$C_n = \left(\frac{1}{n}a_n(\alpha)F_n(x)\{1-F_n(x)\}+\frac{1}{4n^2}a_n^2(\alpha)\right)^{\frac{1}{2}}.$$

When $F_n(x)=1$, the confidence intervals (1) and (2) are

$$\left(1-\frac{k_n(\alpha)}{n},\ 1+\frac{k_n(\alpha)}{n}\right)$$

and

$$\left(\left(1+\frac{a_n(\alpha)}{n}\right)^{-1},\ 1\right)$$

respectively. From the form of above confidence intervals, the intervals based on $A_n$ is expected to shaper for large $F(x)$. We note that the interval based on $K_n$ is inadequate for $x$ near 1 or 0 whereas the interval based on $A_n$ is reasonable for $x$ near 1 or 0.

In order to apply the confidence intervals in section 3, we need the percentage points of the statistics. When $F(x)$ is continuous, the asymptotic distribution of $K_n$ is well known and given by

$$\lim_{n\to\infty} P(K_n\le u) = \sum_{j=-\infty}^{\infty}(-1)^j e^{-2j^2u}. \tag{3}$$

The convergence to the asymptotic distribution (3) is not fast. It is convenient to use Stephens (1986) empirical modification that

$$\left(1+\frac{0.12}{n^{1/2}}+\frac{0.11}{n}\right)^2 K_n$$

approximately follows the asymptotic distribution (3). The percentage point $k_n(\alpha)$ can be determined when $F(x)$ is continuous.

The asymptotic distribution of $A_n$ does not exist, see Anderson and Darling (1952). However, the distribution of $A_n$ exists for finite $n$ and we can expect that (2) is useful for not too small $\alpha$.

If the population distribution is discrete, then (3) does not hold and the asymptotic percentage pointage points determined by (3) gives conservative intervals. Hence, we must estimate percentage points to obtain accurate intervals.

## 3. Estimate Percentage Points $k_n(\alpha)$ and $a_n(\alpha)$

In order to apply the confidence intervals (1) and (2), we must determine $k_n(\alpha)$ and $a_n(\alpha)$. If $F(x)$ is continuous, then $k_n(\alpha)$ is already tabulated, e.g., Owen (1962). However, there are many ties in practical data. Hence we need to estimate them.

Bickel and Krieger (1989) proposed bootstrap method discussed at first by Efron (1979) to estimate $k_n(\alpha)$. They proved that the bootstrap method is consistent. Further they performed Monte Carlo studies for many discrete distributions. Here we also performed Monte

Carlo studies on $k_n(\alpha)$ and $a_n(\alpha)$.

In order to investigate the accuracy of bootstrap estimate, we need true percentage points. Table 1 and 2 give simulated values of $k_n(\alpha)$ and $a_n(\alpha)$ for (a) continuous distribution, (b) discrete uniform distribution on integers 0-10, (c) Binomial distribution B(10, 0.5), (d) Poission distribution with mean 5 and (e) geometric distribution with probability of success $p=0.2$ where the sample sizes are $n=20$ and $n=40$ and numbers of replications are 10,000. We regard that the entries in percentage points. The entries at "bootstrap (a)" are the bootstrapped percentage points for continuos distribution. If $F(x)$ is continuous or there are no ties in the observations, we may regard that the observations are $\{1/n, 2/n, \cdots, (n-1)/n, 1\}$ for bootstrap experiments and hence we do not compute bootstrapped sample for each data.

Table 1. Simulated percentage points of $K_n$

| Case | $n$ | 50% | 90% | 95% | 99% |
|---|---|---|---|---|---|
| (a) | 20 | 0.624 | 1.400 | 1.737 | 2.548 |
| | 40 | 0.649 | 1.458 | 1.784 | 2.505 |
| bootstrap | 20 | 0.450 | 1.250 | 1.250 | 2.450 |
| (a) | 40 | 0.625 | 1.225 | 1.600 | 2.500 |
| (b) | 20 | 0.423 | 1.117 | 1.390 | 2.142 |
| | 40 | 0.437 | 1.131 | 1.423 | 2.108 |
| (c) | 20 | 0.322 | 0.995 | 1.041 | 1.547 |
| | 40 | 0.375 | 0.924 | 1.198 | 1.990 |
| (d) | 20 | 0.359 | 0.933 | 1.157 | 1.949 |
| | 40 | 0.331 | 1.024 | 1.355 | 1.866 |
| (e) | 20 | 0.407 | 1.133 | 1.373 | 1.946 |
| | 40 | 0.422 | 1.063 | 1.399 | 2.201 |

Table 2. Simulated percentage points of $A_n$

| Case | $n$ | 50% | 90% | 95% | 99% |
|---|---|---|---|---|---|
| (a) | 20 | 4.957 | 20.480 | 40.168 | 188.878 |
| | 40 | 5.414 | 21.768 | 41.797 | 187.815 |
| bootstrap | 20 | 2.813 | 6.275 | 9.474 | 13.889 |
| (a) | 40 | 3.600 | 8.100 | 9.231 | 15.172 |
| (b) | 20 | 2.366 | 6.125 | 7.088 | 10.580 |
| | 40 | 2.335 | 5.760 | 7.606 | 10.453 |
| (c) | 20 | 2.901 | 8.169 | 14.994 | 49.268 |
| | 40 | 2.637 | 11.200 | 23.662 | 23.662 |
| (d) | 20 | 2.945 | 9.064 | 22.854 | 69.705 |
| | 40 | 2.923 | 10.485 | 12.380 | 45.706 |
| (e) | 20 | 2.998 | 10.856 | 18.774 | 96.667 |
| | 40 | 2.998 | 10.989 | 18.267 | 75.069 |

Tables 3 and 4 are the results of simulated data where the numbers of replications are 5,000. The entries with "S" are the numbers of samples which are contained in the confidence intervals the confidence limits of which are given by Tables 1 and 2. The entries "B" are the numbers of samples which are contained in the confidence intervals the percentage points of which are estimated by the bootstrap method where the number of bootstrap replications per sample is 1,000. The number of bootstrap replications for the case (a) is 10,000 and the confidence limits are given by Tables 1 and 2.

From Tables 3 and 4, we can find that, the bootstrap method is not effective when $F(x)$ is continuous. However, when $F(x)$ is continuous, we adopt confidence limits determined by simulation studies for $a_n(\alpha)$ and by Stephens(1986) for $k_n(\alpha)$. If $F(x)$ is discrete, the bootstrap estimate of confidence limit is effective for $k_n(\alpha)$. For $a_n(\alpha)$, the bootstrap method is inadequate in the cases (b) and (c) and is reasonable in the cases (d) and (e). The distribution (b) and (c) are well approximated by uniform distribution over the unit interval and

a normal distribution, respectively. Thus, we can conclude that $a_n(\alpha)$ with bootstrap method is useful for discrete distribution which are not well approximated by a continuous distribution.

Table 3. Number of coveraged samples based on $K_n$

| Case | $n$ | | 50% | 90% | 95% | 99% |
|---|---|---|---|---|---|---|
| (a) | 20 | B | 2933 | 8559 | 8559 | 9537 |
| | 40 | B | 4853 | 8451 | 9274 | 9717 |
| (b) | 20 | S | 2596 | 4585 | 4771 | 4960 |
| | | B | 2593 | 4230 | 4663 | 4914 |
| | 40 | S | 2575 | 4549 | 4796 | 4957 |
| | | B | 2269 | 4413 | 4660 | 4937 |
| (c) | 20 | S | 2885 | 4536 | 4798 | 4953 |
| | | B | 1980 | 4390 | 4730 | 4942 |
| | | S | 2609 | 4606 | 4764 | 4967 |
| | | B | 2307 | 4457 | 4712 | 4934 |
| (d) | 20 | S | 2398 | 4493 | 4730 | 4965 |
| | | B | 2755 | 4565 | 4838 | 4978 |
| | 40 | S | 2605 | 4530 | 4738 | 4960 |
| | | B | 2581 | 4540 | 4758 | 4942 |
| (e) | 20 | S | 4221 | 4945 | 4984 | 4996 |
| | | B | 3120 | 4674 | 4863 | 4979 |
| | 40 | S | 4655 | 4966 | 4988 | 4998 |
| | | B | 3036 | 4633 | 4839 | 4978 |

Table 4. Number of coveraged samples based on $A_n$

| Case | $n$ | | 50% | 90% | 95% | 99% |
|---|---|---|---|---|---|---|
| (a) | 20 | B | 1867 | 6154 | 7421 | 8352 |
| | 40 | B | 2677 | 6747 | 7241 | 8381 |
| (b) | 20 | S | 2670 | 4627 | 4771 | 4961 |
| | | B | 2116 | 4322 | 4730 | 4898 |
| | 40 | S | 2606 | 4543 | 4784 | 4952 |
| | | B | 2558 | 4418 | 4752 | 4948 |
| (c) | 20 | S | 3261 | 4583 | 4807 | 4997 |
| | | B | 1495 | 4186 | 4354 | 4611 |
| | 40 | S | 2706 | 4522 | 4981 | 4981 |
| | | B | 1876 | 3741 | 4367 | 4511 |
| (d) | 20 | S | 2523 | 4556 | 4860 | 4981 |
| | | B | 2666 | 4533 | 4773 | 4961 |
| | 40 | S | 2605 | 4505 | 4727 | 4943 |
| | | B | 2543 | 4471 | 4722 | 4943 |
| (e) | 20 | S | 4008 | 4746 | 4784 | 4968 |
| | | B | 3124 | 4650 | 4849 | 4976 |
| | 40 | S | 4022 | 4631 | 4895 | 4925 |
| | | B | 2809 | 4586 | 4791 | 4973 |

## References


[1] Anderson, T. W. and Darling, D. A. (1952). Asymptotic theory of certain goodness of fit criteria based on stochastic processes. Ann. Math. Statist., 23, 193-212

[2] Bickel, P. J. and Krieger, A. M. (1989). Confidence bands for a distribution function using the bootstrap. J. Amer. Statist. Soc., 84, 95-100.

[3] Efron, B.(1979). Bootstrap methods: another look at the jadkknife. Ann. statist., 7, 1-26.

[4] Owen, D. B. (1962). Handbook of statistical Tables. Addison-Wesley, Reading, Mass.

[5] Shirahata, S. and Chu, I. S. (1992). Integrated squared error of kernel-type estimator of distribution function. Ann. Inst. statist. Math., 44, 579-591

[6] Stephens, M. A. (1986). Tests Based on EDF statistics. Goodness-of-Fit Techniques (ED. D'Agostino and Stephens). 97-193.