On the Nonparametric Statistics for the Distribution Function

Kim Hae-kyung

分布函數의 非母數統計量에 關한 硏究

金 海 琎

요 약

本 論文은 連續母分布函數의 推定 및 檢定을 위한 非母數統計量을 만들고 이 統計量의 近似確率分布를 求하여 適合度檢定에 이용될 수 있음을 보였다.

1. Introduction.

In the present note we shall study the problem of devising an optimum statistic for the statistical inference of an unknown population cumulative distribution function. In finding a statistic for distribution function, it is important to construct it as nonparametric so that its distribution does not depend on the functional form of the cumulative distribution function of the population. Kolmogorov and Sminov (1933) gave an asymptotic solution to this problem for large sample. This proposed an useful nonparametric statistic, so called Kolmogorov statistic, for the problem. In this note we gave another solution of the problem by means of constructing a statistic whose distribution is asymptotic chi-square distribution which is widely tabulated in order to avoid tedious computation of Kolmogorov statiistic.

Let $X_1, X_2, \ldots, X_n$ denote a random sample from an unknown cumulative distribution function $F_X(x)$. The sample cumulative distribution function, denoted by $F_n(x)$, is defined by

$$F_n(x) = \frac{1}{n}(\text{number of } X_i \text{ less than or equal to } x)$$
$$= \frac{1}{n}\sum_{i=1}^{n} I_{(-\infty, x)}(X_i)$$

for all real number x. For fixed x, $F_n(x)$ is a statistic since it is a function of the sample. It therefore has following probability distribution

$$P\left[F_n(x) = \frac{k}{n}\right] = \binom{n}{k}\left[F_X(x)\right]^k \cdot \left[1-F_X(x)\right]^{n-k}$$
$k=0, 1, 2, \cdots, n$

For fixed x, $F_n(x)$ is an unbiased and consistent estimator for $F_X(x)$, regardless of the form of $F_X(x)$. According to Borel's strong law of large numbers, the statistic $F_n(x)$ converges to $F_X(x)$ with probabilility one (Glivenko-Gantelli Theorem). Therefore, for sufficiently large sample size n, the deviations, $F_X(x)-F_n(x)$, should be small for all values of x.

2. Nonparametric statistcs.

For any real value x and $0<\delta<1$ let

(2·1) $$U_n^{\delta} = \inf_x \left[F_n(x)-F_X(x)+\delta\right]$$
$\delta \leq 1-F(x)$

(2·2) $$V_n^{\delta} = \sup_x \left[F_n(x)-F_X(x)-\delta\right]$$
$\delta \leq 1-F(x)$

Then for fixed n, δ, $U_n^{\delta}>0$ ($V_n^{\delta}<0$) implies that the curve of the sample cumulative distribution function $F_n(x)$ never overlaps the curveof the function $F_X(x)-\delta$ ($F_X(x)+\delta$). Note that $U_n^{\delta}>0$ if and only if $F_X(x)-F_n(x)<\delta$; and similarly ($V_n^{\delta}<0$) if and only if $F_n(x)-F_X(x)<\delta$.

These statistics are nonparametric, as is proved in the following theorem.

Theorem 1. The statistics U_n^+, V_n^+ are nonparametric for any continuous distribution function $F_X(x)$

Proof. The sample cumulative distribution function is constructed from the order statistics as

$$F_n(x)=\frac{k}{n} \quad \text{for } X_{(k)}\le x<X_{(k+1)} \quad k=0,1,2,\cdots,n$$

where $X_{(1)}, X_{(2)},\cdots, X_{(n)}$ are the order statistics of a random sample and $X_{(0)}=-\infty$, $X_{(n+1)}=\infty$

Therefore, we have

$$U_n^+=\inf_x\left[F_n(x)-F_X(x)+\delta\right]=\min_{0\le k\le n.}\ \inf_{X_{(k)}\le x<X_{(k+1)}}\left[F_n(x)-F_X(x)+\delta\right]$$

$$=\min_{0\le k\le n.}\ \inf_{X_{(k)}\le x<X_{(k+1)}}\left[\frac{k}{n}-F_X(x)+\delta\right]$$

$$=\min_{0\le k\le n}\left[\frac{k}{n}-\sup_{X_{(k)}\le x<X_{(k+1)}}F_X(x)+\delta\right]$$

$$=\min_{0\le k\le n.}\left[\frac{k}{n}-F_X(X_{(k+1)})+\delta\right]$$

Similarly, we have

$$V_n^+=\sup_x\left[F_n(x)-F_X(x)-\delta\right]=\max_{0\le k\le n,}\ \sup_{X_{(k)}\le x<X_{(k+1)}}\left[F_n(x)-F_X(x)-\delta\right]$$

$$=\max_{0\le k\le n}\ \sup_{X_{(k)}\le x<X_{(k+1)}}\left[\frac{k}{n}-F_X(x)-\delta\right]$$

$$=\max_{0\le k\le n}\left[\frac{k}{n}-\inf_{X_{(k)}\le x<X_{(k+1)}}F_X(x)-\delta\right]$$

$$=\max_{0\le k\le n.}\left[\frac{k}{n}-F_X(X_{(k)})-\delta\right]$$

Therefore, U_n^+ and V_n^+ are nonparameteric statistics since the probability distributions of these are depend only on the random variables $F_X(X_{(k)})$ which are the order statistics from the uniform distribution regardless of $F_X(x)$ as long as contineous on (0, 1).

Theorem 2. Let U_n^+ and V_n^+ be nonparametric statistic from the sample distribution function $F_n(x)$ and true distribution function $F_X(x)$ as in (2.1) and (2.2) then

$$P\left[U_n^+>0\right]=P\left[V_n^+<0\right]$$

Proof. Let $X_{(1)}, X_{(2)},\cdots, X_{(n)}$ be the order statistics of a sample from a contineous distribution function $F_X(x)$.

Let

$$Y_{(i)}=F_X(X_{(i)}) \qquad \text{for } i=1,2,..,n$$

then

$Y_{(1)}, Y_{(2)},..,Y_{(n)}$ are also order statistics from the uniform distribution on(0, 1) with density function

$$f_{Y_{(1)},Y_{(2)},\cdots Y_{(n)}}(y_1,y_2,\cdots y_n)=n!\cdot I_A$$

where $A=\{(y_1, y_2,.., y_n)|0<y_1<y_2<\cdots<y_n<1\}$

One can easily check that $U_n^+>0$ if and only if

$$Y_{(i)}<Y_{(i+1)}<\frac{i}{n}+\delta \qquad \text{for } i=0,\ 1,\ ..,k$$

$$Y_{(i)}<Y_{(i+1)}<1 \qquad \text{for } i=k+1,\ k+2,..,n$$

where $K=[n(1-\delta)]$ and $Y_{(0)}=0$

similarly, $V_n^+<0$ ifand only if

$$\frac{i}{n}-\delta<Y_{(i)}<Y_{(i+1)} \text{ for } i=n-k,..,n$$

$$0<Y_{(i)}<Y_{(i+1)} \qquad \text{for } i=1,2,..,\ n-k-1$$

where $Y_{(n+1)}=1$. But if we take

$$\hat{Y}_{(i)}=1-Y_{(n-i+1)} \quad \text{for } i=1,2,..,n$$

then the order statistics $\hat{Y}_{(1)}, \hat{Y}_{(2)},..., \hat{Y}_{(n)}$ have the same distribution function as $Y_{(1)}Y_{(2)},..,$ $Y_{(n)}$ and the inequalities(2.4) reduce to(2.3). Therefore, we have $P[U_n^+>0]=P[V_n^+<0]$.

From the above theorem we see that $\lim_{n\to\infty}$ $P\left[U_n^+>0\right]=\lim_{n\to\infty}P\left[V_n^+<0\right]$ and let $u(\delta)$ denote the common limit probability distribution in the equality. We then have the followings.

Theorem 3. For every $h\ge 0$, $u\left(\frac{h}{\sqrt{n}}\right)=1-\exp[-2h^2]$

Proof. $P\left[U_n^+>0\right]=P\left[Y_{(i)}<Y_{(i+1)}<\frac{i}{n}+\delta\right.$ for $i=0,1,..,k$, $Y_{(i)}<Y_{(i+1)}<1$ for $i=k+1,$ $\left.k+2,..,n\right]$

$$=P\left[(0<Y_{(1)}<\delta)\cap(y_{(1)}<Y_{(2)}<\frac{1}{n}+\delta)\cap\cdots\cap\right.$$

$$(y_{(k-1)}<Y_{(k)}<\frac{k}{n}+\delta)\cap(y_{(k)}<Y_{(k+1)}<1)\cap(y_{(k+1)}$$

$$\left.<Y_{(k+2)}<1)\cap\cdots\cap(y_{(n-1)}<Y_{(n)}<1)\right].$$

Hence, we obtain the probability as follows:

$$(2.5)\quad n!\int_0^{\delta}\int_{y_{(1)}}^{\frac{1}{n}+\delta}\cdots\int_{y_{(k-1)}}^{\frac{k}{n}+\delta}\int_{y_{(k)}}^{1}\cdots\int_{y_{(n-1)}}^{1} dy_{(n)}\cdots dy_{(k)}\cdots dy_{(1)}$$

$$(2.6)\quad =n!\int_0^{\delta}\int_{y_{(1)}}^{\frac{1}{n}+\delta}\int_{y_{(k-2)}}^{\frac{k}{n}+\delta}\frac{(1-y_{(k)})^{n-k}}{(n-k)!}dy_{(k)}\cdots dy_{(1)}$$

$$=n!\left(\int_0^{\delta}\int_{y_{(1)}}^{\frac{1}{n}+\delta}\cdots\int_{y_{(k-2)}}^{\frac{k-1}{n}+\delta}\frac{(1+y^{n-k}_{(k-1)})}{(n-k+1)!}dy_{(k-1)}\cdots dy_{(1)}\right.$$

$$(2.7)\quad \left.-\frac{\left(1-\delta-\frac{k-1}{n}\right)^{n-k+1}}{(n-k+1)!}\int_0^{\delta}\int_{y_{(1)}}^{\frac{1}{n}+\delta}\cdots\int_{y_{(k-2)}}^{\frac{k-1}{n}+\delta}dy_{(k-1)}\cdots dy_{(1)}\right)$$

Let $A_{n,\delta}(k)$ denote the integral (2.6),

$$(2.8)\quad B_{n,\delta}(k)=\int_0^{\delta}\int_{y_{(1)}}^{\frac{1}{n}+\delta}\cdots\int_{y_{(k-1)}}^{\frac{k}{n}+\delta}dy_{(k)}\cdots dy_{(1)}$$

$$(2.9)\quad f(k)=\frac{\left(1-\delta-\frac{k}{n}\right)^{n-k}}{(n-k)\ !}$$

We get

$$(2.10)\quad P\left[U_n^{\delta}>0\right]=n!\cdot A_{n,\delta}(k)=n!(A_{n,\delta}(k-1)-f(k-1)\cdot B_{n,\delta}(k-1)$$

On the other hand,

$$A_{n,\delta}(k)=A_{n,\delta}(k-1)-f(k-1)B_{n,\delta}(k-1)$$
$$=A_{n,\delta}(k-2)-f(k-2)B_{n,\delta}(k-2)-f(k-1)B_{n,\delta}(k-1)$$

Using finite induction we have

$$A_{n,\delta}(k)=A_{n,\delta}(0)-\left[f(0)B_{n,\delta}(0)+f(1)\cdot B_{n,\delta}(1)+\cdots+f(k-1)\cdot B_{n,\delta}(k-1)\right]$$
$$=A_{n,\delta}(0)-\left[\sum_{i=0}^{k-1}f(i)\ B_{n,\delta}(i)\right]$$

Where k is the largest integer in $n(1-\delta)$. Now let us consider the limit probability $\lim_{n\to\infty} P\left[U_n^{\delta}>0\right]$ when $\delta=\frac{h}{\sqrt{n}}$,

We can check that $u\left(\frac{h}{\sqrt{n}}\right)=1-\exp[-2h^2]$ The proof of the theorem is complete.

Remark. According to the above Theorem 3. we have

$$\lim_{n\to\infty} P\left[U_n^{\frac{1}{2}\sqrt{n}}>0\right]=\lim_{n\to\infty} P\left[V_n^{\frac{1}{2}\sqrt{n}}<0\right]=1-e^{-\frac{1}{2}d}$$

for all $d\geq 0$ which is the chi-square distribution with two degrees of freedom. Therefore $P\left[2\sqrt{n}\ (F_X(x)-F_n(x)<\sqrt{d}\right]=P\left[2\sqrt{n}\ (F_n(x)-F_X(x))>\sqrt{d}\right]$ is approximately chi-square distibution. This fact allows statistics, $H_n=2\sqrt{n}\ (F_X(x)-F_n(x))$ and $G_n=2\sqrt{n}(F_n(x)-F_X(x))$, to be broadly used as test statistics for goodness of fit. That is H_n and G_n are used to one tailed test that the distribution that is being sampled from is some specified continuous distribution ; a test null hypothesis $F_X(x)=F_0(x)$ against one sided alternative $F_X(x)>F_0(x)$ or $F_X(x)<F_0(x)$ respectively.

References

[1] Bell, C, B, and Haller, Smith H.: 1969. Bivariate Symmetry Tests; Parametric and Nonparametric, Ann. Math. Statist.: 259 ~269.

[2] Bickel, P. J.: 1969. A distribution free vers ion of the Smirov two sample test in the pvariate case, Ann. Math. statist.: 1~23.

[3] W. B. Davenport: 1970 Probability and random Processes, McGraw-Hill,

[4] T. S. Ferguson: 1967 Mathematical Statistics; A Decision Theoretic Approach, Academic Press,

[5] J. D. Gibbons: 1971 Nonparametric Statistical Inference, McGraw-Hill,

[6] E. L. Lehmann: 1959. Testing Statistical Hypotheses, John Wiley & Sons, Inc.,

[7] A. Levine: 1963. Theory of Probability, D. Van Nostrand Company, Inc.,

[8] M. Loeve: 1963. Probability Theory. 3rd ed., D. Van Nostrand Company, Inc.,

[9] A. M. Mood: 1974. Introduction to the Theory of Statistics, 3rd ed., McGraw-Hill,

[10] S. Siegel: 1956. Nonparametric Statistics, Mc Graw-Hill.

[11] S. S. Wilk: 1962. Mathematical Statistics, John Wiley & Sons.,