SOME LIMIT THEOREMS IN REGRESSION THEORY

By K. R. PARTHASARATHY

and

P. K. BHATTACHARYA

Indian Statistical Institute

SUMMARY. (X,Y) follows an unknown bivariate distribution with $0 \leqslant X \leqslant 1$ and the regression of Y on X is continuous and a sequence of observations on (X,Y) are made. An estimate of the unknown regression function based on these observations and motivated by the method of Fractile Graphical Analysis has been suggested. Its large sample properties, viz., convergence in probability and almost sure uniform convergence to the true regression function have been investigated. Large sample tests for a specified regression function have also been proposed for the case when the conditional variance function of Y on X is known and for the case when it is unknown.

1. Introduction

Let us suppose that X and Y are real valued variables having a certain joint distribution function such that X takes values in the interval $[0, 1]$ and all conditional absolute moments of order up to $p \geqslant 3$ exist when X is fixed at any point x. If we can make a sequence of independent observations on (X, Y) the question naturally arises as to how we can construct from these observations an estimate of the unknown regression function of Y on X, possessing certain properties like convergence in probability, almost sure uniform convergence etc., to the true regression function. Another important problem is that of constructing at least a large sample test for a specified regression function. In Section 2 we shall make use of the technique of Fractile Graphical Analysis suggested by Mahalanobis (1958) to estimate the regression and analyse its large sample properties and in Section 3, construct a large sample test for the regression. The crucial part of our analysis consists of the utilization of certain results concerning the error of approximation by the central limit theorem and an upper bound of the tail probabilities in the distribution of sums of independent and bounded random variables. These results are given in the Appendix for reference.

2. Estimation of the regression function

Let (x_1, y_1), (x_2, y_2), ..., (x_{kn}, y_{kn}) be independent observations on (X, Y), $x_{(r)}$ the r-th order statistic in the set of observed values of X and $y_{(i)} = y_j$ if $x_{(i)} = x_j$. Throughout this section we shall assume that the distribution function of X is continuous and strictly increasing so that the probability of any two observations on X being

equal is zero. Then the statistics $x_{(r)}$ are defined unambiguously for almost all samples. In order to simplify notation we write for $r = 1, 2, \ldots, k$ and $s = 1, 2, \ldots, n$,

$$x_{(r-1n+s)} = x_{rs}, \quad y_{(r-1n+s)} = y_{rs},$$

$$E[Y \mid X = x] = \varphi(x), \quad E[Y^2 \mid X = x] = \xi(x),$$

$$E[\,|\,Y - \varphi(x)\,|^m \mid X = x] = \beta_m(x),\ m = 2, 3, \ldots, p,$$

$$\varphi(x_{rs}) = \varphi_{rs}, \quad \xi(x_{rs}) = \xi_{rs},\ \beta_m(x_{rs}) = \beta_{mrs},$$

$$\bar{y}_r = \sum_{s=1}^{n} y_{rs}/n, \qquad \overline{y^2}_r = \sum_{s=1}^{n} y_{rs}^2/n,$$

$$\bar{\varphi}_r = \sum_{s=1}^{n} \varphi_{rs}/n, \qquad \overline{\varphi^2}_r = \sum_{s=1}^{n} \varphi_{rs}^2/n,$$

$$\bar{\beta}_{mr} = \sum_{s=1}^{n} \beta_{mrs}/n, \quad \bar{\xi}_r = \sum_{s=1}^{n} \xi_{rs}/n. \qquad \ldots \quad (2.1)$$

Now we define two functions $f_{nk}(x)$ and $\hat{\varphi}_{nk}(x)$ as follows :

$$\begin{aligned} f_{nk}(x) &= \bar{y}_1 \quad \text{if} \quad 0 \leqslant x \leqslant x_{1n} \\ &= \bar{y}_r \quad \text{if} \quad x_{r-1n} < x \leqslant x_{rn}, \quad r = 2, \ldots k-1 \\ &= \bar{y}_k \quad \text{if} \quad x_{k-1n} < x \leqslant 1. \qquad \ldots \quad (2.2) \end{aligned}$$

$$\begin{aligned} \hat{\varphi}_{nk}(x) &= \bar{\varphi}_1 \quad \text{if} \quad 0 \leqslant x \leqslant x_{1n} \\ &= \bar{\varphi}_r \quad \text{if} \quad x_{r-1n} < x \leqslant x_{rn}, \quad r = 2 \ldots k-1 \\ &= \bar{\varphi}_k \quad \text{if} \quad x_{k-1n} < x \leqslant 1. \qquad \ldots \quad (2.3) \end{aligned}$$

Lemma 1 : *If the random variable X has a strictly increasing continuous distribution function $F(x)$, then*

$$P\left[\, F^{-1}\left(\frac{r}{k} - \frac{1}{2k}\right) \leqslant x_{rn} \leqslant F^{-1}\left(\frac{r}{k} + \frac{1}{2k}\right), r = 1, 2, \ldots k-1 \,\right]$$

$$\geqslant 1 - 4m \,.\, \exp\left[-\frac{3n}{16k}\right] - 2k.\, \exp\left[-\frac{n}{4k}\left(1 - \frac{1}{8m}\right)\right]$$

for every fixed positive integer m.

Proof : We have

$$P[x_{rn} < z] = \sum_{s=0}^{kn-rn} \binom{kn}{s} [1 - F(z)]^s\, [F(z)]^{kn-s}. \qquad \ldots \quad (2.4)$$

Applying Theorem (A 2) to the sum S_N of N independent binomial random variables each with probability for success equal to p, we have

$$P[S_N \geqslant N(p+\delta)] \leqslant \begin{cases} \exp\left[\, \dfrac{-N\delta}{2p} \log\left(\dfrac{\delta}{2q} + \sqrt{1 + \dfrac{\delta^2}{4q^2}}\,\right)\right] \text{if } p \geqslant q \\ \exp\left[\, \dfrac{-N\delta}{2q} \log\left(\dfrac{\delta}{2p} + \sqrt{1 + \dfrac{\delta^2}{4p^2}}\,\right)\right] \text{if } p \leqslant q, \end{cases}$$

which, if $\frac{\delta}{2p}$ and $\frac{\delta}{2q}$ are less than unity, becomes

$$P[S_N \gtrless N(p+\delta)] \leqslant \begin{cases} \exp\left[-N\delta^2\left(1-\frac{\delta}{4q}\right)\right] \text{ if } p \geqslant q \\ \exp\left[-N\delta^2\left(1-\frac{\delta}{4p}\right)\right] \text{ if } p \leqslant q. \end{cases} \quad \ldots (2.5)$$

We have

$$P\left[F^{-1}\left(\frac{r}{k}-\frac{1}{2k}\right) \leqslant x_{rn} \leqslant F^{-1}\left(\frac{r}{k}+\frac{1}{2k}\right), r = 1, 2, \ldots k-1\right]$$

$$\geqslant 1-\sum_{r=1}^{k-1} P\left[x_{rn} < F^{-1}\left(\frac{r}{k}-\frac{1}{2k}\right)\right]-\sum_{r=1}^{k-1} P\left[x_{rn} > F^{-1}\left(\frac{r}{k}+\frac{1}{2k}\right)\right]$$

which, after using (2.4) and an application of (2.5) for $\delta = \frac{1}{2k}$ and two binomial sums with probability for successes $\frac{r}{k}-\frac{1}{2k}$ and $1-\frac{r}{k}-\frac{1}{2k}$ and number of summands equal to nk, becomes greater than $4m.\exp\left[-\frac{3n}{16k}\right]+2k.\exp\left[\frac{-n}{4k}\left(1-\frac{1}{8m}\right)\right]$ for every fixed integer m.

Lemma 2 : *Under the conditions of Lemma* 1 *if* $\varphi(x)$, *the regression function of* Y *on* X *is continuous then for any* $\epsilon > 0$

$$P\left[\sup_{0\leqslant x\leqslant 1} |\hat{\varphi}_{nk}(x)-\varphi(x)| > \epsilon\right] < c(\epsilon, \varphi)\left[m\exp\left[-\frac{3n}{16k}\right]+k\exp\left[-4k\left(1-\frac{1}{8m}\right)\right]\right]$$

for every fixed integer m, *and where* $c(\epsilon, \varphi)$ *is a constant which depends on* ϵ *and* φ *only.*

Proof : Let $\delta[a, b] = \max_{a\leqslant x_1, x_2\leqslant b} |\varphi(x_1)-\varphi(x_2)|$

$$\delta_k = \max\left\{\delta\left[0, F^{-1}\left(\frac{1}{2k}\right)\right], \delta\left[F^{-1}\left(\frac{1}{k}-\frac{1}{2k}\right), F^{-1}\left(\frac{1}{k}+\frac{1}{2k}\right)\right], \ldots, \delta\left[F^{-1}\left(1-\frac{1}{2k}\right), 1\right]\right\}.$$

If $$F^{-1}\left(\frac{r}{k}-\frac{1}{2k}\right) \leqslant x_{rn} \leqslant F^{-1}\left(\frac{r}{k}+\frac{1}{2k}\right), r=1, 2, \ldots k-1$$

then $$d_{nk}(\varphi) = \sup_{0\leqslant x\leqslant 1} |\hat{\varphi}_{nk}(x)-\varphi(x)| \leqslant \delta_k. \quad \ldots (2.6)$$

Thus

$$P[d_{nk}(\varphi) > \epsilon] \leqslant \left[P[d_{nk} > \epsilon | F^{-1}\left(\frac{r-\frac{1}{2}}{k}\right) \leqslant x_{rn} \leqslant F^{-1}\left(\frac{r+\frac{1}{2}}{k}\right), r = 1, 2, \ldots k-1\right]$$

$$+1-P\left[F^{-1}\left(\frac{r-\frac{1}{2}}{k}\right) \leqslant x_{rn} \leqslant F^{-1}\left(\frac{r+\frac{1}{2}}{k}\right)\right], r = 1, 2, \ldots, k_{-1}. \quad \ldots (2.7)$$

Because of (2.6) the first term on the right side of (2.7) becomes less than $P(\delta_k > \epsilon)$. Since F is continuous and strictly increasing F^{-1} is continuous. Since

$\varphi(x)$ is uniformly continuous we have $\delta_k < \varepsilon$ for $k > k_0(\varepsilon, \varphi)$. Thus the first term vanishes for $k > k_0(\varepsilon, \varphi)$. The second term becomes less than $4m \exp\left[\frac{-3n}{16k}\right] + 2k \exp\left[\frac{-n}{4k}\left(1-\frac{1}{8m}\right)\right]$ because of Lemma 1. Combining the two we get the required inequality.

Hereafter, for simplicity, we shall assume that $\beta_s(x) > c_1 > 0$ and $\beta_m(x) < c_2$ for all $m = 2, \ldots, p$ and all x. However from the proofs we can see that all our results hold good under more general conditions.

Lemma 3 : *If* $\beta_s(x) > c_1 > 0$ *and* $\beta_m(x) < c_2$ *for all* $m = 1, 2, \ldots, p$ *and all* x, *then for any* $\varepsilon > 0$

$$P\left[\sup_{0 \leq x \leq 1} |f_{nk}x - \hat{\varphi}_{nk}(x)| > \varepsilon\right] < C.\frac{k}{n^{\frac{p-1}{2}}}$$

where C *is a constant depending only on* ε, p, c_1 *and* c_2.

Proof : We have

$$P(\varepsilon) = P\left[\sup_{0 \leq x \leq 1} |f_{nk}(x) - \hat{\varphi}_{nk}(x)| > \varepsilon | x_1, \ldots, x_{kn}\right]$$

$$= P\left[\sup_{r=1,2,\ldots,k} |\bar{y}_r - \bar{\varphi}_r| > \varepsilon | x_1 \ldots, x_{kn}\right]$$

$$\leqslant \sum_{r=1}^{k} P\left[\frac{|\bar{y}_r - \bar{\varphi}_r|}{\sqrt{\bar{\beta}_{pr}/n}} > \frac{\varepsilon\sqrt{n}}{\sqrt{c_2}}\right]. \qquad \ldots (2.8)$$

Making use of (A7) and noting that $T_{pn} < \frac{\sqrt{n}}{4}$ we have

$$\frac{\varepsilon\sqrt{n}}{\sqrt{c_2}} > \sqrt{4(p-2)\log T_{pn}} \text{ for } n > n(\varepsilon).$$

Hence by Theorem A1 we obtain

$$P(\varepsilon) \leqslant 2k\Phi\left(\frac{-\varepsilon\sqrt{n}}{\sqrt{c_2}}\right) \frac{\theta_p c_2}{1 + \left(\frac{\varepsilon\sqrt{n}}{\sqrt{c_2}}\right)^p} \cdot \frac{k}{n^{\frac{p-2}{2}}} \qquad \ldots (2.9)$$

where $c_2 = \frac{4c_2^{3/p}}{c_1^{3/2}}$ and θ_p is a positive constant which depends only on p and $\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\, dt$. Since $\Phi(-x) \leqslant \frac{1}{\sqrt{2\pi}\, x}\, e^{-x^2/2}$, $x > 0$, (2.9) gives, for $n > n(\varepsilon)$.

$$P(\varepsilon) \leqslant \frac{k}{\varepsilon\sqrt{n/c_2}}\, e^{\frac{-\varepsilon^2}{2c_2}.n} + \theta_p.\frac{c_2^{p/2}}{\varepsilon^p}\,\frac{c_2 k}{n^{\frac{p-1}{2}}}$$

$$= c_4(\varepsilon, p, c_1, c_2)\left[\frac{k}{\sqrt{n}}\, e^{\frac{-\varepsilon^2}{2c_2}.n} + \frac{k}{n^{\frac{p-1}{2}}}\right]$$

which completes the proof.

Now we shall state and prove the main result of this section.

Theorem 1 : *If the p-th conditional absolute moment of Y exists when X is fixed at any point x, $p \geqslant 3$, X has a strictly increasing continuous distribution function in $[0, 1]$, the regression $\phi(x)$ of Y on X is continuous and $\beta_2(x) > c_1 > 0$, $\beta_m(x) < c_2$ for all $m = 1, 2, \ldots, p$ and for all x, then for $n \geqslant (4+\delta)\, k \log k$, $\delta > 0$, $d_{nk} = \sup_x | f_{nk}(x) - \phi(x) |$ converges to zero in probability as $k \to \infty$ and for $n \geqslant (8+\delta)\, k \log k$, $\delta > 0$, d_{nk} converges almost surely to zero as $k \to \infty$.*

Proof : The first part of the theorem is an immediate consequence of Lemmas 2 and 3 and the second part follows from the same lemmas and Borel-Cantelli lemma.

3. Large sample test of a specified regression function

In this section we shall consider the problem of testing the null hypothesis H_0 that the regression function $\phi(x)$ is equal to a specified function $\mu(x)$.

Consider the statistic

$$\tau_{nk} = \sup_{r=1, \ldots, k} \frac{|\bar{y}_r - \bar{\mu}_r|}{\sqrt{\bar{\beta}_{2r}/n}},$$

where $\bar{y}_r$ and $\bar{\beta}_{2r}$ are as defined in (2.1) and $\bar{\mu}_r = \dfrac{\sum_{s=1}^{n} \mu(x_{rs})}{n}$. If we had known the probability distribution of τ_{nk} under the null hypothesis, then for any given level of significance $0 < \alpha < 1$ we could apply the following test:

Reject H_0 if and only if $\tau_{nk} > \tau_{nk}(\alpha)$

where $$P[\tau_{nk} > \tau_{nk}(\alpha) \,|\, H_0] = \alpha.$$

Theorem 2 enables us to apply such a test at least in the large sample.

Theorem 2 : *If $\beta_2(x) > c_1 > 0$ and $\beta_m(x) < c_2 < \infty$ for all x and $m = 1, 2, \ldots, p$, $p \geqslant 3$ then we have for $n \geqslant (k \log k)^{2/p-2}$*

$$\lim_{k \to \infty} P[\tau_{nk} \leqslant \lambda_k \,|\, H_0] = \exp\left[-\frac{1}{\sqrt{\pi}}\, e^{-\theta} \right]$$

where $$\lambda_k = \sqrt{2(\theta + \log k - \tfrac{1}{2} \log \log k)}.$$

Proof: Under H_0, we have

$$P(\lambda_k) = P[\tau_{nk} \leqslant \lambda_k \,|\, x_1, \ldots, x_{kn}]$$

$$\prod_{r=1}^{k} P\left[\frac{|\bar{y}_r - \bar{\phi}_r|}{\sqrt{\bar{\beta}_{2r}/n}} \leqslant \lambda_k \,|\, x_1, \ldots, x_{kn} \right] = \prod_{r=1}^{k} [\Phi(\lambda_k) - \Phi(-\lambda_k) + e_{rk}]$$

where because of (A8) and the fact that $\lambda_k < \sqrt{4(p-2) \log T_{kn}}$ for $k \geqslant k_0(\theta)$ we get

$$|e_{rk}| \leqslant c_5 \left[\frac{(1 + \lambda_k^3)}{\sqrt{n}}\, e^{-\lambda_k^2/2} + \frac{1}{n^{\frac{p-2}{2}}} \right]$$

where c_5 is a constant depending on p, c_1, c_2. Thus

$$\log P(\lambda_k) = k \log [\Phi(\lambda_k)-\Phi(-\lambda_k)]+Z_k,$$

where

$$|Z_k| \leqslant \sum_{r=1}^{k} \left|\log \left[1+ \frac{e_{rk}}{\Phi(\lambda_k)-\Phi(-\lambda_k)}\right]\right|.$$

Since for sufficiently large k

$$\frac{|e_{rk}|}{\Phi(\lambda_k)-\Phi(-\lambda_k)} < \tfrac{1}{2}$$

and $\log (1+x) = x+vx^2, |v| < 1$ for $|x| < \frac{1}{2}$,

$$|Z_k| \leqslant c_6 k \left[\frac{1+\lambda_k^2}{\sqrt{n}} e^{-\lambda_k^2/2} + \frac{1}{n^{\frac{p-2}{2}}}\right] \qquad \ldots (3.1)$$

Substituting for λ_k in (3.1) we have

$$|Z_k| \leqslant c_6 \left[\frac{1+\lambda_k^2}{\sqrt{n}} \cdot \sqrt{\log k}.\, e^{-\theta} + \frac{k}{n^{\frac{p-2}{2}}}\right] \qquad \ldots (3.2)$$

Since $n \geqslant (k \log k)^{2/p-2}$ it is easy to see from (3.2) that $|Z_k| \to 0$ as $k \to \infty$. This completes the proof.

The statistic τ_{nk} defined above can be used only when $\beta_2(x)$ is known. The natural way of modifying this in the case of unknown variance function is to replace β_{2r} by $y_r^2-\bar{y}_r^2$ where y_r^2 and $\bar{y}_r$ are as in (2.1). Since this leads to certain complications in evaluating the limiting distribution we replace it by

$$s_r^2 = \frac{1}{n}\sum_{i=1}^{n} [y_{ri}-\mu(x_{ri})]^2 \text{ (where } x_{ri} \text{ is as in (2.1)) and write}$$

$$t_{nk} = \sup_{r=1,2,\ldots,k} \sqrt{n}\,.\, \frac{|\bar{y}_r-\bar{\mu}_r|}{s_r}$$

To show that this replacement does not change τ_{nk} effectively in large samples we prove the following lemma.

Lemma 4 : *If X has a continuous distribution function and Y is a bounded random variable, $\beta_2(x)>c_1>0$ and if $n \geqslant (\log k)^{3+\delta}$, $\delta>0$ then under the null hypothesis H_0*

$$\lim_{k\to\infty} \log k\,.\, \sup_r \left|\sqrt{\frac{s_r^2}{\bar{\beta}_{2r}}}-1\right| = 0,$$

with probability one.

Proof: We have for any $\epsilon > 0$,

$$P\left[\sup\left|\sqrt{\frac{s_r^2}{\bar{\beta}_{2r}}}-1\right| > \frac{\epsilon}{\log k} \,|H_0\right] \leqslant \sum_{r=1}^{k} P\left[\left|\frac{s_r^2}{\bar{\beta}_{2r}}-1\right| > \frac{\epsilon}{\log k}\right]$$

$$\leqslant \sum_{r=1}^{k} P\left[\,|s_r^2-\bar{\beta}_{2r}| > \frac{\epsilon c_1}{\log k}\right]$$

$$\leqslant \sum_{r=1}^{k} P\left[\left|\bar{y}_r^2-\xi_r\right| > \frac{\epsilon c_1}{2\log k}\right] + \sum_{r=1}^{k} P\left[\left|\sum_{i=1}^{n} \varphi_{ri}\, y_{ri}-\varphi_{ri}^2\right| > \frac{n\epsilon c_1}{4\log k}\right].$$

Then from Theorem A2 we get

$$\sum_{r=1}^{k} P\left[\left|\bar{y}_r^2-\bar{\xi}_r\right|>\frac{\epsilon c_1}{2\log k}\,|x_1,\ldots,x_{kn}\right]\leqslant 2k\exp\left[\frac{-b_1 n}{(\log k)^2}\right]$$

and

$$\sum_{r=1}^{k} P\left[\left|\sum_{i=1}^{n}(\varphi_{ri}\,y_{ri}-\varphi_{ri}^2)\right|>\frac{n\epsilon c_1}{4\log k}\right]\leqslant 2k\exp\left[-\frac{b_2 n}{(\log k)^2}\right]$$

where b_1 and b_2 are constants depending on c_1, ϵ and the upper bound of $|Y|$. Thus whenever $n\geqslant(\log k)^{2+\delta}$ the series

$$\sum_{k=1}^{\infty} P\left[\sup_r\left|\sqrt{\frac{\bar{\delta}_r^2}{\bar{\beta}_{2r}}}-1\right|>\frac{\epsilon}{\log k}\right]$$

converges. An application of Borel-Cantelli lemma completes the proof of the lemma.

Remark : It is possible to show that if $n\geqslant k\,(\log k)^{2+\delta}$, $\delta>0$ then $\log k\,\sup_r\left|\sqrt{\frac{\bar{\delta}_r^2}{\bar{\beta}_{2r}}}-1\right|$ converges to zero in probability even when Y is not a bounded random variable.

The following theorem enables us to use the statistic t_{nk} for testing a specified regression function when the conditional variance function is not known.

Theorem 3 : *If X is a random variable with a continuous distribution in $[0,1]$, Y is a bounded random variable and $\beta_2(x)>c_1>0$ then for $n\geqslant(k\log k)^{2/p-2}$ for some $p\geqslant 3$ we have*

$$\lim_{k\to\infty} P[t_{nk}<\lambda_k\mid H_0]=\exp\left[-\frac{1}{\sqrt{\pi}}\,e^{-\theta}\right]$$

where

$$\lambda_k=\sqrt{2(\theta+\log k-\tfrac{1}{2}\log\log k)}.$$

Proof: It is easily seen that

$$P[\tau_{nk}<\lambda_k(1+V_{nk})\,|\,H_0]\leqslant P[t_{nk}<\lambda_k|H_0]\leqslant P[\tau_{nk}<\lambda_k(1+U_{nk})\,|\,H_0]\quad\ldots\quad(3.3)$$

where τ_{nk} is as in Theorem 2, $U_{nk}=\sup_r\left(\sqrt{\frac{\bar{\delta}_r^2}{\bar{\beta}_{2r}}}-1\right)$ and $V_{nk}=\inf_r\left(\sqrt{\frac{\bar{\delta}_r^2}{\bar{\beta}_{2r}}}-1\right)$.

Further, $P[\tau_{nk}<\lambda_k(1+V_{nk})]=P\left[\frac{\tau_{nk}^2-2\log k+\log\log k}{2}.\,Z_{nk}+v_{nk}\leqslant\theta\right]$

where Z_{nk} converges to unity in probability and v_{nk} converges to zero inprobability because of Lemma 4. This fact together with an application of Theorem 2 leads to the result

$$\lim_{k\to\infty} P[\tau_{nk}<\lambda_k(1+V_{nk})]=\exp\left[-\frac{1}{\sqrt{\pi}}\,e^{-\theta}\right].\qquad\ldots\quad(3.4)$$

A similar procedure leads to the result

$$\lim_{k\to\infty} P[\tau_{nk} < \lambda_k(1+U_{nk})] = \exp\left[-\frac{1}{\sqrt{\pi}}\, e^{-\theta}\right]. \qquad \ldots \quad (3.5)$$

(3.3), (3.4) ad (3.5) complete the proof.

Finally we shall state a theorem concerning the order of change in the level of significance if we apply the test 'reject H_0 if and only if $\tau_{nk} > \lambda_k$' where $[\Phi(\lambda_k)-\Phi(-\lambda_k)]^k = 1-\alpha$ instead of the test 'reject H_0 if and only if $\tau_{nk} < \tau_{nk}(\alpha)$ with $P[\tau_{nk} \leqslant \tau_{nk}(\alpha)|H_0] = 1-\alpha$.

Theorem 4: *Let λ_k satisfy $\Phi[(\lambda_k)-\Phi(-\lambda_k)]^k = 1-\alpha$ and $\alpha < 1-\left[1-\sqrt{\frac{2}{e\pi}}\right]^k$. Then under the conditions of Theorem 2 we have*

$$\left| P[\tau_{nk} \leqslant \lambda_k|H_0]-(1-\alpha) \right| \leqslant c.\left[\frac{\left(\log\frac{k}{\alpha}\right)^2}{\sqrt{n}}\cdot\alpha + \frac{k}{n^{\frac{p-2}{2}}}\right]$$

where c is a constant depending on p, c_1 and c_2.

4. REMARKS

In this paper we have not given any consideration to the power of the τ_{nk} and t_{nk} tests proposed in Section 3. It would be very interesting if lower bounds for the power of these tests could be given in terms of the supremum distance between the regression functions under the null hypothesis and the alternative hypothesis. Though this has not been done, we can at least easily verify that if the regression functions under null hypothesis and the alternative hypothesis are both continuous, the τ_{nk} and t_{nk} tests are consistent, i.e., for any given level of significance $0 < \alpha < 1$, the probability of rejecting H_0 tends to unity as $k \to \infty$.

For the validity of the above theorems we have imposed certain conditions on the regression function φ and the variance function β_2 separately. In some cases (e.g. Binomial on Possion distribution) β_2 is a function of φ. This however does not affect our analysis in any way so long as the conditions on φ and β_2 remain valid separately.

For simplicity in proof we assumed that $\beta_2(x)$ and $\beta_m(x)$ are bounded away from zero and bounded above respectively. However, it is easy to show that theorems 1, 2 and 4 are valid under the weaker assumption that $E[\beta_2(x)]^{-2(p-2)/2}$ and $E[\beta_3(x)]^{6(p-2)/p}$ are finite.

---

*$\Phi(x)$ is the usual normal distribution function.

Appendix

In this section we shall state and sketch the proof of some of the results which were utilised in our paper.

Let $X_1, X_2, X_3 \ldots$ be a sequence of independent random variables with $EX_n = 0$ and $EX_n^2 = \sigma_n^2$ and $\beta_{\nu n}$ the ν-th absolute moment of X_n. We shall suppose that the moments of order $k > 3$ exist. We write

$$B_{\nu n} = \frac{1}{n}(\beta_{\nu 1} + \ldots + \beta_{\nu n}), \quad \rho_{\nu n} = \frac{B_{\nu n}}{B_{2n}^{\nu/2}}, \quad T_{kn} = \frac{\sqrt{n}}{4\rho_{kn}^{3/k}} \qquad \ldots \text{(A1)}$$

Then
$$1 \leqslant \rho_{\nu n} \leqslant \rho_{kn}^{\nu/k}, \quad \nu = 2, 3, \ldots, k. \qquad \ldots \text{(A2)}$$

Let $f_n(t)$ and $F_n(x)$ be the characteristic function and distribution function of $\frac{X_1 + \ldots + X_n}{\sqrt{nB_n}}$. We now state two lemmas one of which is due to Cramér (1937) and the other due to Esseen and Berry.

Lemma A1 : (Cramér). *For* $|t| \leqslant T_{kn}^{1/3}$ *we have*

$$\left| f_n(t) - e^{-t^2/2} - \sum_{\nu=1}^{k-3} \frac{P_{\nu n}(it)}{n^{\nu/2}} e^{-t^2/2} \right| \leqslant \frac{c_k}{T_{kn}^{k-2}} [|t|^k + |t|^{3(k-2)}] . e^{-t^2/2}$$

where
$$P_{\nu n}(it) = \sum_{j=1}^{\nu} c_{j\nu n} (it)^{\nu+2j},$$

$c_{j\nu n} = c_k' . \rho_{kn}^{\frac{\nu+2j}{k}}$ *and* c_k *and* c_k' *are constants depending only on* k, c_k *being positive.*

Lemma A2 : (Berry-Esseen). *For* $|t| \leqslant T_{3n}$

$$\left| f_n(t) - e^{-t^2/2} \right| \leqslant \frac{4}{T_{3n}} . \ |t|^3 . e^{-t^2/2} .$$

Let
$$g(t) = e^{-t^2/2} \left[1 + \sum_{\nu=1}^{k-3} \frac{P_{\nu n}(it)}{n^{\nu/2}} \right]. \qquad \ldots \text{(A3)}$$

Hereafter we shall denote by θ_k any positive constant which depends only on k. Then by using (A1), (A2) and Lemma A1 we have

$$\left| \sum_{\nu=1}^{k-3} \frac{P_{\nu n}(it)}{n^{\nu/2}} \right| \leqslant \theta_k \sum_{\nu=1}^{k-3} \sum_{j=1}^{\nu} \left| \frac{c_{j\nu n}}{n^{\nu/2}} \right| |t|^{\nu+2} \leqslant \theta_k . \sum_{\nu=1}^{k-3} \sum_{j=1}^{\nu} \frac{\rho_{kn}^{\frac{\nu+2j}{k}}}{n^{\nu/2}} . \ |t|^{\nu+2j} .$$

$$\leqslant \theta_k . \sum_{\nu=1}^{k-3} \sum_{j=1}^{\nu} \frac{\rho_{kn}^{\frac{2\nu}{k}}}{(\sqrt{n})^{\nu}} \ |t|^{\nu+2j} \leqslant \theta_k \sum_{\nu=1}^{k-3} \sum_{j=1}^{\nu} \frac{|t|}{T_{kn}^{\nu}}. \qquad \ldots \text{(A4)}$$

From (A3) and (A4) it is seen that the inequality of Lemma A2 can be rewritten as

$$\left| f_n(t)-g_n(t) \right| \leqslant \frac{4}{T_{3n}} |\ t\ |^3 . e^{-t^2/3} + \theta_k . \sum_{r=1}^{k-3} \sum_{j=1}^{r} \frac{|t|^{r+2j}}{T_{kn}^r} e^{-t^2/3} \qquad \ldots \quad (A5)$$

Lemma A3 : $$\int_{-T_{kn}}^{+T_{kn}} \left| \frac{f_n(t)-g_n(t)}{t} \right| dt \leqslant \frac{\theta_k}{T_{kn}^{k-2}}.$$

Case 1 : Let $T_{kn} \leqslant 1$, then $T_{kn} \leqslant T_{kn}^{1/3}$. Hence applying Lemma A1 we get

$$\int_{-T_{kn}}^{+T_{kn}} \left| \frac{f_n(t)-g_n(t)}{t} \right| dt \leqslant \frac{\theta_k}{T_{kn}^{k-2}} .$$

Case 2 : Let $T_{kn} > 1$. From (A2) it is easy to see that $1 < T_{kn} \leqslant T_{3n}$:

Thus $$\int_{-T_{kn}}^{T_{kn}} \left| \frac{f_n(t)-g_n(t)}{t} \right| dt = 2 \int_{T_{kn}^{1/3}}^{T_{kn}} + \int_{-T_{kn}^{1/3}}^{T_{kn}^{1/3}} \left| \frac{f_n(t)-g_n(t)}{t} \right| dt.$$

$$= 2I_1 + I_2 \text{ say.}$$

In the region of integration of I_1 Lemma A2 is applicable and for I_2 Lemma A1 is applicable. Thus from (A5) we have

$$\int_{-T_{kn}}^{T_{kn}} \left| \frac{f_n(t)-g_n(t)}{t} \right| dt \leqslant \frac{\theta_k}{T_{kn}^{k-2}} + \frac{\theta_k}{T_{kn}} e^{-T_{kn}^{2/3}/4} \leqslant \theta_k / T_{kn}^{k-2}.$$

Let $$Q(x) \geqslant 0, 0 \leqslant q(t) \leqslant 1, \int_{-\infty}^{+\infty} Q(x)dx = 1.$$

$$q(t) = \int_{-\infty}^{+\infty} e^{itx} Q(x)dx,\ q(t) \leqslant \frac{c}{|t|^{3(k-2)}}$$

$$\int_{-\infty}^{+\infty} |t|^{3k-7} q(t)dt < \infty. \qquad \ldots \quad (A6)$$

Lemma A4 : *If $q(t)$ satisfies conditions* (A6) *then for* $|x| > 1$

$$P(T_{kn}) = \int_{-T_{kn}}^{+T_{kn}} \left| \frac{f_n(t)-g_n(t)}{t} \right| q(xt)dt < \frac{\theta}{1+|x|^k} \cdot \frac{1}{T_{kn}^{k-2}}$$

Proof: *Case* 1 : Let $T_{kn} \leqslant 1$, then $T_{kn} \leqslant T_{kn}^{1/3}$. Applying Lemma A1.

$$P(T_{kn}) \leqslant \frac{\theta_k}{T_{kn}^{k-2}} \int_{-\infty}^{+\infty} [|t|^k + |t|^{3(k-2)}] e^{-t^2/2} \cdot \frac{q(xt)dt}{|t|}$$

$$\leqslant \frac{\theta_k}{T_{kn}^{k-2}} \cdot \frac{1}{|x|^k} \cdot \int_{-\infty}^{+\infty} [|t|^{k-1} + |t|^{3k-7}] e^{-t^2/2} q(t)dt \leqslant \frac{\theta_k}{(1+|x|^k)} \frac{1}{T_{kn}^{k-2}}.$$

Case 2 : Let $T_{kn} > 1$. Then as before $1 < T_{kn} \leqslant T_{3n}$.

$$P(T_{kn}) = 2 \int_{T_{kn}^{1/3}}^{T_{kn}} + \int_{-T_{kn}^{1/3}}^{T_{kn}^{1/3}} \left| \frac{f_n(t) - g_n(t)}{t} \right| q(xt)dt = 2I_1 + I_2 \text{ say.}$$

As before by applying Lemma A1 for I_2 and Lemma A2 for I_1 and proceeding as in Case 1 we obtain

$$P(T_{kn}) \leqslant \frac{\theta_k}{1+|x|^k} \cdot \frac{1}{T_{kn}^{k-2}}.$$

Utilising Lemmas A3 and A4 and proceeding along the same lines as Esseen (1944) we can prove the following results.

Theorem A1 : *Let $X_1, X_2, \ldots, X_n$ be a sequence of random variables with mean zero and finite absolute moments of order $k (k \geqslant 3)$. Then for $x \geqslant \lambda_n$*

$$\left| F_n(x) - \Phi(x) - \sum_{\nu=1}^{k-3} \frac{P_{\nu n}(-\Phi)}{n^{\nu/2}} \right| \leqslant \frac{\theta_k}{1+|x|^k} \cdot \frac{1}{T_{kn}^{k-2}} \text{ and for } x < \lambda_n$$

$$\left| F_n(x) - \Phi(x) - \sum_{\nu=1}^{k-3} \frac{P_{\nu n}(-\Phi)}{n^{\nu/2}} \right| \leqslant \theta_k' \left[\frac{(1+|x|^3)e^{-x^2/2}}{T_{kn}} + \frac{1}{T_{kn}^{k-2}} \right]$$

where $\lambda_n = 1+\delta$ if $T_{kn} \leqslant 1$ and $\max [1+\delta, \sqrt{2.(1+\delta)(k-2)\log T_{kn}}]$ if $T_{kn} > 1$.

θ_k and θ_k' are constants which depends only on k and δ, $P_{\nu n}(-\Phi) = \sum_{j=1}^{\nu} (-1)^{\nu+2j} c_{j\nu n} \Phi^{(\nu+2j)}(x)$ and $c_{j\nu n}$ are as in Lemma A 1.

Remark: From the above theorem it can be easily deduced that

$$\left| F_n(x) - \Phi(x) \right| \leqslant \frac{\theta_k}{1+|x|^k} \cdot \frac{1}{T_n^{k-2}} \text{ for } x > \lambda_n \qquad \ldots \text{ (A 7)}$$

$$\left| F_n(x) - \Phi(x) \right| \leqslant \theta_k' \left[\frac{(1+|x|^3)e^{-x^2/2}}{T_{kn}} + \frac{1}{T_{kn}^{k-2}} \right] \text{ for } x \leqslant \lambda_n. \qquad \text{(A 8)}$$

Let $\xi_1, \xi_2, \ldots, \xi_n$ be n independently distributed random variables with $E(\xi_i) = 0$, $|\xi_i| \leqslant c$, $i = 1, 2, \ldots n$, $E(\xi_1 + \ldots + \xi_n)^2 = \sigma^2$ and $\xi = \xi_1 + \ldots + \xi_n$. Then we have the following theorem due to Prohorov (1959).

Theorem A2 : *Under the conditions stated above, for* $x > 0$ *we have*

$$P[\xi > x] \leqslant \exp\left[-\frac{x}{2c} \sinh^{-1} \frac{xc}{2\sigma^2}\right].$$

REFERENCES

CRAMÉR, H. (1937) : Random variables and probability distributions. *Cambridge Tracts in Mathematics*, No. 36.

ESSEEN, C. G. (1944) : Fourier analysis of distribution functions. A mathematical study of the Laplace-Gaussian Law. *Acta Math.*, **77**, 1.

MAHALANOBIS, P. C. (1958) : Lectures in Japan : Fractile Graphical Analysis. Indian Statistical Institute.

PROHOROV, YU. V. (1959) : An extremal problem in probability theory. *Theoriia Veroyatnostei i ee Primenenia* **4**, 211.

Paper received : May, 1960.

ACKNOWLEDGEMENT

Thanks are due to the following for their kind help in the editing of the papers published in this issue.

Shri S. John	Dr. C. R. Rao
Professor D. B. Lahiri	Shri R. Ranga Rao
Dr. Sujit Kumar Mitra	Shri J. Sethuraman

Editor,
Sankhyā : The Indian Journal of Statistics.