UNIVERSITY OF NORTH CAROLINA
Department of Statistics
Chapel Hill, N. C.

ON AN ANALOG OF REGRESSION ANALYSIS

by

P. K. Bhattacharya

August 1962

Contract No. AF 49(638)-213

In a multivariate distribution the nature of dependence of one variate on the others is studied through conditional quantiles. Estimators and test procedures for conditional quantile functions are given and some of their large sample properties investigated.

This research was supported by the Air Force Office of Scientific Research.

Institute of Statistics
Mimeo Series No. 335

ON AN ANALOG OF REGRESSION ANALYSIS.

by

P. K. Bhattacharya.[1]

1. Introduction.

Suppose $(X_1, \ldots, X_h, Y)$ follow an unknown multivariate distribution on which independent observations are made. The nature of dependence of Y on $(X_1, \ldots, X_h)$ is fully understood only when we know how the conditional distribution of Y given $X_1 = x_1, \ldots, X_h = x_h$, changes with $(x_1, \ldots, x_h)$. This can be attempted in two different ways. One approach is to make inference about the functional relation between the conditional moments of Y given $X_1 = x_1, \ldots, X_h = x_h$, and $(x_1, \ldots, x_h)$, and a special case of this (when the behavior of only the conditional expectation of Y is studied) is known in statistical literature as regression analysis. The classical methods of regression analysis are based on the assumption that $E[Y \mid X_1 = x_1, \ldots, X_h = x_h]$ is a linear function of $x_1, \ldots, x_h$. Mahalanobis [3] and Parthasarathy and Bhattacharya [4] have proposed some methods of regression analysis which do not involve any such linearity assumption. The methods proposed by Parthasarathy and Bhattacharya can be generalized in a straightforward manner for estimating and testing hypotheses about higher order conditional moments of Y, but in order to prove the consistency of these estimates and tests for the first m conditional moments, one should assume the existence of at least first 3m conditional absolute moments. A second approach to this problem is to make inference about the functional relation between the quantities of the conditional distribution of Y given $X_1 = x_1, \ldots, X_h = x_h$, and $(x_1, \ldots, x_h)$. This kind of an analog of the variance components analysis has been considered by Roy and Cobb [5], while Sathe [6] has given a test for the conditional median of Y given X under a restricted model.

1. This research was supported by the Air Force Office of Scientific Research.

In this paper, estimates have been proposed for conditional quantile functions of Y given $X_1, \ldots, X_h$, and the simultaneous uniform convergence of any number of such estimates to the corresponding conditional quantile functions has been studied. A large sample test for the hypothesis that certain conditional quantile functions are equal to specified functions, has been suggested and proved to be consistent. In order to avoid complicated notations, the methods and their properties will be discussed for the bivariate case in sections 2, 3 and 4, while in section 5, the corresponding methods for the multivariate case will be explained and their properties will be stated without using too many symbols.

2. Problem, assumptions and notations.

(X, Y) has a bivariate distribution. F is the marginal distribution function of X and G_x is the conditional distribution function of Y given X = x. For any $0 < p < 1$, $\phi_p(x)$ is the solution of the equation

$$G_x(\phi_p(x)) = p \quad . \tag{1}$$

On the basis of independent observations on (X, Y), we want to estimate the function ϕ_p for a given p, and for a specified real valued function μ defined on the range of X, we want to test the hypothesis $H_o: \phi_p = \mu$.

In what follows, $0 < p < 1$ is always a specified number for which we are interested in ϕ_p.

We make the following assumptions about F, $\{ G_x \}$ and ϕ_p:

(i) The range of X is bounded; for simplicity $0 \leq X \leq 1$.

(ii) F is continuous and strictly increasing.

(iia) For each x, $0 \leq x \leq 1$, G_x is continuous and strictly increasing.

(iii) ϕ_p is continuous (hence uniformly continuous).

(iv) For any given $\epsilon > 0$, there exists $\delta > 0$ (not depending on x) such that

$$|p' - p| < \delta \text{ implies } |\phi_p(x) - \phi_{p'}(x)| < \epsilon \text{ for all } x.$$

It should be noted that when I say that a univariate distribution function H is strictly increasing, we mean thereby, that H is strictly increasing on an interval (a, b), where

$$a = \begin{cases} \text{the g. l. b. of the set } \{x: H(x) = 0\} & \text{if this set is non-empty} \\ -\infty & \text{otherwise,} \end{cases}$$

and

$$b = \begin{cases} \text{the g. l. b. of the set } \{x: H(x) = 1\} & \text{if this set is non-empty} \\ \infty & \text{otherwise.} \end{cases}$$

Another point to note is that condition (iv) is satisfied if there exists a function ψ on $[0, 1]$ such that $G_x(y) = G_0(y + \psi(x))$.

$(X_1, Y_1), \ldots, (X_{nk}, Y_{nk})$ are independent observations on (X, Y). Let $X_{(1)} < \ldots < X_{(nk)}$ be the ordered values of $X_1, \ldots, X_{nk}$. $Y_{(i)} = Y_j$ if $X_{(i)} = X_j$. For $r = 1, \ldots, k$ and for $s = 1, \ldots, n$,

$$X_{(\overline{r-1}\cdot n+s)} = X_{rs} \quad , \quad Y_{(\overline{r-1}\cdot n+s)} = Y_{rs} \quad .$$

For any given integer k, define a set of random intervals as follows:

$$I_{k1} = [0, X_{1n}], \quad I_{kr} = (X_{r-1,n}, X_{rn}], \quad r = 2, \ldots, k-1, \text{ and}$$

$$I_{kk} = (X_{k-1,n}, 1] \quad .$$

Next let $Y_{r(1)} < \ldots < Y_{r(n)}$ be the ordered values of $Y_{r1}, \ldots, Y_{rn}$, and denote by $[a]$ the largest integer $\leq a$. Now define a random step-function f_{nk} on $[0,1]$ as follows:

$$f_{nk}(x) = Y_{r([np])} \quad \text{if} \quad x \in I_{kr} \quad , \quad r = 1, \ldots, k \quad .$$

In course of the analysis carried out in the next two sections, we shall make use of an upper bound for the tail probabilities of sums of independent and

bounded random variables due to Hoeffding [2]. We shall also make use of an upper bound for the error of approximation by the Central Limit Theorem and an upper bound of error involved in the usual approximation of the distribution function of "frequence χ^2" for a simple hypothesis regarding a multinomial distribution, both due to Esseen [1]. For the sake of completeness, these are stated below.

Theorem (Hoeffding). If $X_1, \ldots, X_n$ are independent random variables, $0 \leq X_i \leq 1$, $EX_i = \mu$, $S_n = X_1 + \ldots + X_n$, and if $0 < t < 1-\mu$, then

$$P[S_n - ES_n \geq nt] \leq e^{-2nt^2} \quad . \tag{2}$$

Theorem (Esseen). If $X_1, \ldots, X_n$ are a sequence of independent random variables with the same distribution function F, the mean value zero, the variance $\sigma^2 \neq 0$ and the finite absolute moments $\beta_1, \ldots, \beta_\nu$ (ν is an integer ≥ 3), then for $|x| < \sqrt{(1+\delta)(\nu-2)\log n}$,

$$|F_n(x) - \Phi(x)| \leq C_1(\delta, \beta)n^{-1/2}(1 + |x|^3)e^{-x^2/2} + C_2(\delta, \beta)n^{-(\nu-2)/2} \quad , \tag{3}$$

where F_n is the distribution function of $(X_1 + \ldots + X_n)/\sqrt{n}\,\sigma$, $\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\,dt$, $C_1(\delta, \beta)$ and $C_2(\delta, \beta)$ are finite constants, depending only on $0 < \delta < 1$, and the moments $\beta_1, \ldots, \beta_\nu$.

Theorem (Esseen). If n independent observations are taken on a multinomial distribution with $(m+1)$ classes and with class probabilities $q_1, \ldots, q_{m+1}$, $\sum_1^{m+1} q_i = 1$, and if $n_1, \ldots, n_{m+1}$ are the respective observed frequencies, $\sum_1^{m+1} n_i = n$, then

$$(4) \qquad P\left[\sum_{1}^{m+1} (n_i - nq_i)^2/nq_i \le \chi^2\right] = \frac{1}{2^{m/2}\Gamma(m/2)} \int_0^{\chi^2} e^{-w/2}\, w^{m/2-1}\, dw + \frac{Q(q_1, \ldots, q_{m+1})}{n^{m/(m+1)}},$$

where $Q(q_1, \ldots, q_{m+1})$ is a finite constant, depending only on $q_1, \ldots, q_{m+1}$.

3. Convergence of f_{nk}.

In this section we shall study the convergence in probability and almost sure convergence of f_{nk} to ϕ_p, uniformly. It is obvious that both n and k should tend to infinity for such convergences to take place, but the crucial point in our analysis is to find out how n and k should depend on each other as they tend to infinity.

We shall first prove a probability inequality for the event that the random variables $\{X_{rn}\}$, $r = 1, \ldots, k-1$, lie in some neighborhoods of $F^{-1}(r/k)$, $r = 1, \ldots, k-1$ respectively. The following lemma is a modification of a similar probability inequality given by Parthasarathy and Bhattacharya [4].

Lemma 1. Let $0 < a_k < 1$, $k = 1, 2, \ldots$ ad inf., be a sequence converging to zero. Under assumption (ii),

$$(5) \qquad P\left[F^{-1}\left(\frac{r}{k} - a_k\right) \le X_{rn} \le F^{-1}\left(\frac{r}{k} + a_k\right),\ r = 1, \ldots, k-1\right] > 1 - 2ke^{-2nka_k^2},$$

where $F^{-1}(a)$ for $a \le 0$ is defined to be 0 and $F^{-1}(a)$ for $a \ge 1$ is defined to be 1.

Proof. The left side of (5) is

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

$$= 1 - \sum_{r=1}^{k-1} \sum_{s \geq rn} \binom{kn}{s} \left(\frac{r}{k} - a_k\right)^s \left(1 - \frac{r}{k} + a_k\right)^{kn-s} - \sum_{r=1}^{k-1} \sum_{s \geq kn-rn} \binom{kn}{s} \left(1 - \frac{r}{k} - a_k\right)^s \left(\frac{r}{k} + a_k\right)^{kn-s}$$

$$= 1 - \sum_{r=1}^{k-1} \sum_{s \geq kn\left[\left(\frac{r}{k} - a_k\right) + a_k\right]} \binom{kn}{s} \left(\frac{r}{k} - a_k\right)^s \left(1 - \frac{r}{k} + a_k\right)^{kn-s}$$

$$- \sum_{r=1}^{k-1} \sum_{s \geq kn\left[\left(1 - \frac{r}{k} - a_k\right) + a_k\right]} \binom{kn}{s} \left(1 - \frac{r}{k} - a_k\right)^s \left(\frac{r}{k} + a_k\right)^{kn-s}$$

$$\geq 1 - 2(k-1)\, e^{-2nka_k^2}, \quad \text{by (2)}$$

$$> 1 - 2ke^{-2nka_k^2}.$$

We now find a probability inequality for the event that the length of each of the intervals $I_{k1}, \ldots, I_{kk}$ is less than a specified positive number. Let $L(I)$ denote the length of an interval I on the real line.

Lemma 2. Under assumptions (i) and (ii), for any $\delta > 0$,

$$P\left[L(I_{kr}) < \delta,\ r = 1, \ldots, k\right] > 1 - 2ke^{-2kna_k^2}$$

for sufficiently large k.

Proof. We first note that under assumptions (i) and (ii), for any given $\delta > 0$, there exists an integer k_o such that for $k > k_o$,

$$(6) \qquad \max\left[F^{-1}\left(\frac{1}{k} + a_k\right), F^{-1}\left(\frac{r+1}{k} + a_k\right) - F^{-1}\left(\frac{r}{k} - a_k\right),\ r = 1, \ldots, k-2,\right.$$

$$\left. 1 - F^{-1}\left(\frac{k-1}{k} - a_k\right)\right] < \delta.$$

We shall show that (6) along with

$$(7) \qquad F^{-1}(\tfrac{r}{k} - a_k) \leq X_{rn} \leq F^{-1}(\tfrac{r}{k} + a_k) \quad , \quad r = 1, \ldots, k-1 \quad ,$$

implies $L(I_{kr}) < \delta$, $r = 1, \ldots, k$.

For $r = 2, \ldots, k-1$,

$$L(I_{kr}) = X_{rn} - X_{r-1,n}$$

$$\leq F^{-1}(\tfrac{r}{k} + a_k) - F^{-1}(\tfrac{r-1}{k} - a_k) \quad , \quad \text{by (7)}$$

$$< \delta \quad , \quad \text{by (6)} \quad .$$

Also,
$$L(I_{k1}) < X_{1n}$$

$$\leq F^{-1}(\tfrac{1}{k} + a_k) \quad , \quad \text{by (7)}$$

$$< \delta \quad , \quad \text{by (6)} \quad ,$$

and
$$L(I_{kk}) \leq 1 - F^{-1}(\tfrac{k-1}{k} - a_k) \quad , \quad \text{by (7)}$$

$$< \delta \quad , \quad \text{by (6)} \quad .$$

An application of lemma 1 now completes the proof.

Lemma 3. Under assumptions (iia), (iii) and (iv), for sufficiently large k and for any given $\eta > 0$,

$$P\Big[\sup_{x \in I_{kr}} \phi_{p+\epsilon}(x) - \inf_{x \in I_{kr}} \phi_{p-\epsilon}(x) < \eta, \; r = 1, \ldots, k \Big] > 1 - 2ke^{-2nka_k^2} \quad ,$$

if $\epsilon > 0$ is such that

$$| p' - p | < \epsilon \text{ implies } | \phi_p(x) - \phi_{p'}(x) | < \eta/3 \text{ for all } x.$$

(By assumption (iv) such an $\epsilon > 0$ exists for any given $\eta > 0$.)

Proof. Choose $\delta > 0$ such that

$$x_1 \epsilon [0, 1] \quad , \quad x_2 \epsilon [0, 1] \quad , \; | x_1 - x_2 | < \delta$$

together imply

$$| \phi_p(x_1) - \phi_p(x_2) | < \eta/3 \quad .$$

By assumption (iii) such a $\delta > 0$ exists for any given $\eta > 0$. Now,

$$P[\sup_{x \epsilon I_{kr}} \phi_{p+\epsilon}(x) - \inf_{x \epsilon I_{kr}} \phi_{p-\epsilon}(x) < \eta, \; r = 1, \ldots, k]$$

$$\geq P[\sup_{x \epsilon I_{kr}} \phi_{p+\epsilon}(x) - \inf_{x \epsilon I_{kr}} \phi_{p-\epsilon}(x) < \eta, \; L(I_{kr}) < \delta, \; r = 1, \ldots, k] \; .$$

But $P[\sup_{x \epsilon I_{kr}} \phi_{p+\epsilon}(x) - \inf_{x \epsilon I_{kr}} \phi_{p-\epsilon}(x) < \eta, \; r = 1, \ldots, k \,|\, L(I_{kr}) < \delta, r = 1, \ldots, k]$

$$= P[\sup_{x_1, x_2 \epsilon I_{kr}} \{ \phi_{p+\epsilon}(x_1) - \phi_{p-\epsilon}(x_2) \} < \eta, \; r = 1, \ldots, k \,|\, L(I_{kr}) < \delta, \; r = 1, \ldots, k]$$

$$\geq P[\sup_{x_1 x_2 \epsilon I_{kr}} |\phi_p(x_1) - \phi_p(x_2)| < \eta/3, \; \sup_{x \epsilon I_{kr}} |\phi_{p+\epsilon}(x) - \phi_p(x)| < \eta/3,$$

$$\sup_{x \epsilon I_{kr}} |\phi_{p-\epsilon}(x) - \phi_p(x)| < \eta/3, \; r = 1, \ldots, k \,|\, L(I_{kr}) < \delta, \; r = 1, \ldots, k] = 1,$$

by the choice of δ and ϵ. An application of lemma 2 now completes the proof.

Lemma 4. Under assumptions (iia), (iii) and (iv), for arbitrary $\epsilon > 0$ and for large n,

$$P[\inf_{x \epsilon I_{kr}} \phi_{p-\epsilon}(x) < Y_{r([np])} < \sup_{x \epsilon I_{kr}} \phi_{p+\epsilon}(x) \,|\, X_1 = x_1, \ldots, X_{nk} = x_{nk}]$$

$$\geq 1 - 2ke^{-n\epsilon^2} \quad ,$$

for any given $x_1, \ldots, x_{nk}$ in $[0, 1]$.

Proof. Suppose $x_1, \ldots, x_n$ are points in an interval $[a, b] \subset [0, 1]$. Let $Y_1, \ldots, Y_n$ be mutually independent random variables, Y_i having distribution function G_{x_i}, $i = 1, \ldots, n$. Also let $Y_{(1)} < \ldots < Y_{(n)}$ be the ordered values of $Y_1, \ldots, Y_n$. Then

$$P[Y_{([pn])} \leq \inf_{x \in [a,b]} \phi_{p-\epsilon}(x)]$$

$$= P[\text{at least } [np] \text{ of } Y_1, \ldots, Y_n \text{ are} \leq \inf_{x \in [a,b]} \phi_{p-\epsilon}(x)]$$

$$= P\Big[U_1 + \ldots U_n \geq [np] - \sum_{i=1}^{n} G_{x_i}\big(\inf_{x \in [a,b]} \phi_{p-\epsilon}(x)\big)\Big] \quad ,$$

where $U_1, \ldots, U_n$ are mutually independent random variables with

$$P[U_i = 1 - G_{x_i}(\inf_{x \in [a,b]} \phi_{p-\epsilon}(x))] = G_{x_i}(\inf_{x \in [a,b]} \phi_{p-\epsilon}(x))$$

and

$$P[U_i = -G_{x_i}(\inf_{x \in [a,b]} \phi_{p-\epsilon}(x))] = 1 - G_{x_i}(\inf_{x \in [a,b]} \phi_{p-\epsilon}(x)),$$

$i = 1, \ldots, n$.

Now, $G_{x_i}(\inf_{x \in [a,b]} \phi_{p-\epsilon}(x))$

$$\leq G_{x_i}(\phi_{p-\epsilon}(x_i)) \qquad \text{since } x_i \in [a,b]$$

$$= p - \epsilon \qquad \text{by (1)} \quad .$$

Hence $[np] - \sum_{i=1}^{n} G_{x_i}(\inf_{x \in [a,b]} \phi_{p-\epsilon}(x))$

$$\geq np - 1 - n(p - \epsilon) = n\epsilon - 1 > n\epsilon/\sqrt{2} \qquad \text{for large } n.$$

Thus, $$P\lfloor U_1 + \ldots + U_n \geq \lceil np \rceil - \sum_{i=1}^{n} G_{x_i}\left(\inf_{x \in \lfloor a,b \rceil} \phi_{p-\epsilon}(x)\right) \rfloor$$

$$\leq P\lfloor U_1 + \ldots + U_n \geq n\epsilon/\sqrt{2} \rfloor$$

$$\leq e^{-n\epsilon^2} \quad , \quad \text{by (2)} \quad .$$

Similarly, $P\lfloor Y_{(\lceil np \rceil)} \geq \sup_{x \in \lfloor a,b \rceil} \phi_{p+\epsilon}(x) \rfloor \leq e^{-2n\epsilon^2}$.

The desired probability inequality is now obtained.

We are now in a position to prove the following theorem about uniform convergence of f_{nk} to ϕ_p.

Theorem 1. If F, $\{G_x\}$ and ϕ_p satisfy conditions (i) - (iv), then for $n \geq k^{\gamma}$, $\gamma > 0$, $d_{nk} = \sup_{0 \leq x \leq 1} |f_{nk}(x) - \phi_p(x)|$ converges to zero in probability as $k \longrightarrow \infty$ and for $n \geq k$, d_{nk} converges to zero with probability one as $k \longrightarrow \infty$.

Proof. Let $\eta > 0$ be given.

$$P\lfloor \sup_{0 \leq x \leq 1} |f_{nk}(x) - \phi_p(x)| < \eta \rfloor$$

$$\geq P\lfloor \inf_{x \in I_{kr}} \phi_{p-\epsilon}(x) < Y_{r(\lceil np \rceil)} < \sup_{x \in I_{kr}} \phi_{p+\epsilon}(x), \sup_{x \in I_{kr}} \phi_{p+\epsilon}(x) - \inf_{x \in I_{kr}} \phi_{p-\epsilon}(x)$$

$$< \eta, \ r = 1, \ldots, k \rfloor, \text{ for arbitrary } \epsilon > 0,$$

$$= P\lfloor \inf_{x \in I_{kr}} \phi_{p-\epsilon}(x) < Y_{r(\lceil np \rceil)} < \sup_{x \in I_{kr}} \phi_{p+\epsilon}(x), \ r = 1, \ldots, k \mid \sup_{x \in I_{kr}} \phi_{p+\epsilon}(x)$$

$$- \inf_{x \in I_{kr}} \phi_{p-\epsilon}(x) < \eta, \ r = 1, \ldots, k \rfloor \times P\lfloor \sup_{x \in I_{kr}} \phi_{p+\epsilon}(x) - \inf_{x \in I_{kr}} \phi_{p-\epsilon}(x) < \eta, \ r = 1, \ldots, k \rfloor$$

$$> (1 - 2ke^{-n\epsilon^2}) \cdot (1 - 2ke^{-2nka_k^2}) \quad ,$$

by lemmas 3 and 4 if $\epsilon > 0$ is so chosen as to satisfy the condition of lemma 3. Both the factors in the last term tend to one as $k \longrightarrow \infty$ if $n \geq k^{\gamma}$, $\gamma > 0$ and $a_k = k^{-\gamma/2}$, which proves the first part of the theorem. Again,

$$\sum_{k=1}^{\infty} \left[1 - (1 - 2ke^{-n\epsilon^2})(1 - 2ke^{-2nka_k^2}) \right]$$

converges if $n \geq k$ and $a_k = k^{-1/2}$. The second part of the theorem now follows from Borel-Cantelli lemma.

If we have $0 < p_1 < \ldots < p_m < 1$, and if we define

$$f_{nk}^{(i)}(x) = Y_{r([np_i])} \quad \text{for } x \in I_{kr},\ r = 1, \ldots, k \quad , i = 1, \ldots, m ,$$

then theorem 1 can be immediately extended for the simultaneous uniform convergence of $f_{nk}^{(1)}, \ldots, f_{nk}^{(m)}$ to $\phi_{p_1}, \ldots, \phi_{p_m}$ respectively.

4. Large Sample Tests for Specified Conditional Quantile Functions.

Let μ be a specified real-valued function on $[0, 1]$ and consider the problem of testing the hypothesis $H_0: \phi_p = \mu$ against the alternative $H_1: \phi_p \neq \mu$.

We define random variables $\{U_{rs}\}$, $r = 1, \ldots, k$; $s = 1, \ldots, n$, as follows:

$$U_{rs} = \begin{cases} 1 & \text{if } Y_{rs} \leq \mu(X_{rs}) \\ 0 & \text{otherwise} \end{cases} .$$

These are mutually independent random variables and under H_0, each of them takes values 1 and 0 with probabilities p and 1-p respectively. Let $\bar{U}_r = \sum_{s=1}^{n} U_{rs}/n$, $r = 1, \ldots, k$, and

$$\tau_{nk} = \sup_{r = 1, \ldots, k} \frac{|\bar{U}_r - p|}{\sqrt{p(1-p)/n}} .$$

If we can find the limiting distribution of τ_{nk} (suitably standardized) under H_0, then we can test H_0 at any given level of significance $0 < \alpha < 1$, as follows:

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

$$\lim_{k \longrightarrow \infty} P[\tau_{nk} \leq \tau_{nk}(\alpha) \mid H_0] = 1 - \alpha \quad .$$

Again, suppose $0 < p_1 < \ldots < p_m < 1$ are given numbers and $\mu_1, \ldots, \mu_m$ are specified real-valued functions on $[0, 1]$, satisfying the condition that $\mu_1(x) < \ldots < \mu_m(x)$ for all $x \in [0, 1]$. Consider the problem of testing the hypothesis $H_0^{(m)}: (\phi_{p_1} = \mu_1, \ldots, \phi_{p_m} = \mu_m)$ against the alternative $H_1^{(m)}$ that $\phi_{p_i} \neq \mu_i$ for at least one i, $i = 1, \ldots, m$.

Define $$U_{rs}^{(1)} = \begin{cases} 1 & \text{if } Y_{rs} \leq \mu_1(X_{rs}) \\ 0 & \text{otherwise} \end{cases} ,$$

$$U_{rs}^{(i)} = \begin{cases} 1 & \text{if } \mu_{i-1}(X_{rs}) < Y_{rs} \leq \mu_i(X_{rs}) \\ 0 & \text{otherwise} \end{cases} , \quad i = 2, \ldots, m \quad ,$$

$$U_{rs}^{(m+1)} = \begin{cases} 1 & \text{if } Y_{rs} > \mu_m(X_{rs}) \\ 0 & \text{otherwise} \end{cases} ,$$

and $$U_{ro}^{(i)} = \sum_{s=1}^{n} U_{rs}^{(i)} \quad , \quad r = 1, \ldots, k \quad , \quad i = 1, \ldots, m+1 \quad .$$

Let $q_1 = p_1$, $q_i = p_i - p_{i-1}$, $i = 2, \ldots, m$; $q_{m+1} = 1 - p_m$, and consider the statistic

$$\tau_{nk}^{(m)} = \sup_{r=1,\ldots,k} \sum_{i=1}^{m+1} \frac{(U_{ro}^{(i)} - nq_i)^2}{nq_i} \quad .$$

If we can find the limiting distribution of $\tau_{nk}^{(m)}$ (suitably standardized) under $H_0^{(m)}$, then we can test $H_0^{(m)}$ with the help of $\tau_{nk}^{(m)}$ in the same way as we hope to test H_0 with the help of τ_{nk}.

The following theorem gives us the limiting distributions of τ_{nk} and $\tau_{nk}^{(m)}$ under H_0 and $H_0^{(m)}$ respectively.

For any real Θ and for any integer k, define

$$\lambda_k(\Theta) = \begin{cases} \sqrt{2(\Theta + \log k - \frac{1}{2}\log\log k)} & \text{if } \Theta + \log k - \frac{1}{2}\log\log k > 0 \\ 0 \quad \text{otherwise} & , \end{cases}$$

and

$$\lambda_k^{(m)}(\Theta) = \begin{cases} 2\left\{\Theta + \log k + (m/2-1)\log\log k\right\} & \text{if } \Theta + \log k + (m/2-1)\log\log k > 0 \\ 0 \quad \text{otherwise} & . \end{cases}$$

Theorem 2. (a) If $n \geq k^{\gamma}$, $\gamma > 0$, then

$$\lim_{k \longrightarrow \infty} P\left[\tau_{nk} \leq \lambda_k(\Theta) \mid H_0\right] = \exp\left[-\frac{1}{\sqrt{\pi}} e^{-\Theta}\right] .$$

(b) If $n \geq (k \log k)^{1+1/m}$, then

$$\lim_{k \longrightarrow \infty} P\left[\tau_{nk}^{(m)} \leq \lambda_k^{(m)}(\Theta) \mid H_0^{(m)}\right] = \exp\left[-\frac{1}{\Gamma(m/2)} e^{-\Theta}\right] /$$

Proof. (a) Under H_0, $\left\{U_{rs}\right\}$ are independently and identically distributed with mean p, variance p(1-p), and with finite moments of all orders which are functions of p only. Let ν be so large that

$$\lambda_k(\Theta) < \sqrt{(1+\delta)(\nu-2)\gamma \log k} \quad , \; 0 < \delta < 1 \quad , \quad \text{and} \tag{8}$$

$$\gamma(\nu - 2) > 2 \quad . \tag{9}$$

By (8) and (3),

$$P\left[\frac{|\bar{U}_r - p|}{\sqrt{p(1-p)/n}} \le \lambda_k(\varrho) \,\middle|\, H_0\right] = \Phi(\lambda_k(\varrho)) - \Phi(-\lambda_k(\varrho)) + e_{kr} \quad ,$$

$$\text{where } |e_{kr}| \le C\left[\left\{1 + (\lambda_k(\varrho))^3\right\} n^{-1/2} e^{-(\lambda_k(\varrho))^2/2} + n^{-(\nu-2)/2}\right]$$

$$\le C'\left[e^{-\varrho(\log k)^2} k^{-(1+\gamma/2)} + k^{-\gamma(\nu-2)/2}\right] \quad ,$$

since $n \ge k^{\gamma}$, C' being a finite constant, depending on p and p and δ. Now,

$$\log P\left[\tau_{nk} \le \lambda_k(\varrho) \,|H_0\right] = \sum_{r=1}^{k} \log P\left[\frac{|\bar{U}_r - p|}{\sqrt{p(1-p)/n}} \le \lambda_k(\varrho) \,|H_0\right]$$

$$= \sum_{r=1}^{k} \log\left[\Phi(\lambda_k(\varrho)) - \Phi(-\lambda_k(\varrho)) + e_{kr}\right]$$

$$= k \log\left[\Phi(\lambda_k(\varrho)) - \Phi(-\lambda_k(\varrho))\right] + Z_k \quad ,$$

$$\text{where } |Z_k| \le \sum_{r=1}^{k} \left| \log\left[1 + \frac{e_{kr}}{\Phi(\lambda_k(\varrho)) - \Phi(-\lambda_k(\varrho))}\right] \right| \quad .$$

Since $e_{kr} \longrightarrow 0$ and $\Phi(\lambda_k(\varrho)) - \Phi(-\lambda_k(\varrho)) \longrightarrow 1$ as $k \longrightarrow \infty$, for sufficiently large k,

$$\frac{|e_{kr}|}{\Phi(\lambda_k(\varrho)) - \Phi(-\lambda_k(\varrho))} < \frac{1}{2} \quad .$$

Now $\log(1 + x) = x + vx^2$, $|v| < 1$ for $|x| < 1/2$. Hence,

$$|Z_k| \le C'' \cdot k\left[e^{-\varrho(\log k)^2} k^{-(1+\gamma/2)} + k^{-\gamma(\nu-2)/2}\right] \quad ,$$

where C'' is a finite constant depending on p and δ. It now follows from (9) that $Z_k \longrightarrow 0$ as $k \longrightarrow \infty$. To complete the proof we have only to verify that

$$\lim_{k \longrightarrow \infty} k \log \left[\Phi(\lambda_k(\varrho)) - \Phi(-\lambda_k(\varrho)) \right] = - \frac{1}{\sqrt{\pi}} e^{-\varrho} \quad .$$

(b) It can be shown in the same way as in proving (a) that

$$\log \left[\tau_{nk}^{(m)} \leq \lambda_k^{(m)}(\varrho) \mid H_0^{(m)} \right] = k \log \left[\frac{1}{2^{m/2}\Gamma(m/2)} \int_0^{\lambda_k^{(m)}(\varrho)} e^{-w/2} w^{m/2-1} dw \right] + Z_k^{(m)},$$

where $\lim_{k \longrightarrow \infty} Z_k^{(m)}$ can be shown to be zero if $n \geq (k \log k)^{1+1/m}$, by an application of (4). The rest of the proof follows from the fact that

$$\lim_{k \longrightarrow \infty} k \log \left[\frac{1}{2^{m/2}\Gamma(m/2)} \int_0^{\lambda_k^{(m)}(\varrho)} e^{-w/2} w^{m/2-1} dw \right] = - \frac{1}{\Gamma(m/2)} e^{-\varrho} \quad .$$

For any $0 < \alpha < 1$, let

$$\tau_{nk}(\alpha) = \sqrt{2 \left\{ \log(k/\sqrt{\pi}) - \frac{1}{2} \log \log k - \log \log \left(\frac{1}{1-\alpha}\right) \right\}}$$

and $$\tau_{nk}^{(m)}(\alpha) = 2 \left\{ \log(k/\Gamma(m/2)) + (m/2 - 1) \log \log k - \log \log\left(\frac{1}{1-\alpha}\right) \right\} \quad .$$

Theorem 3. Under assumptions (i) - (iv),

(a) If $n \geq k^{\gamma}$, $\gamma > 0$, then the test with critical region

$$\tau_{nk} > \tau_{nk}(\alpha) \quad , \quad 0 < \alpha < 1 \quad ,$$

is a large sample size α test for the null hypothesis H_0, and is consistent against the alternative H_1.

(b) If $n \geq (k \log k)^{1+1/m}$, then the test with critical region

$$\tau_{nk}^{(m)} > \tau_{nk}^{(m)}(\alpha) \quad , \quad 0 < \alpha < 1 \quad ,$$

is a large sample size α test for the null hypothesis $H_0^{(m)}$, and is consistent against the alternative $H_1^{(m)}$.

Proof. The first parts of (a) and (b) are immediate consequences of theorem 2. We shall prove here the second part of (b), and the proof for the second part of (a) will follow on exactly similar lines.

Since we are assuming ϕ_{p_i}, $i = 1, \ldots, m$ to be continuous, we need consider only the case when $\mu_1, \ldots, \mu_m$ are continuous. Suppose $H_0^{(m)}$ does not hold. Then $\phi_{p_i} \neq \mu_i$ for at least one integer i between 1 and m. Let j be the smallest integer for which $\phi_{p_j} \neq \mu_j$. Since ϕ_{p_j} and μ_j are both continuous, there exists some $\delta > 0$ and an interval $(c_1, c_2) \subset [0, 1]$, such that

$$\text{either} \quad \mu_j(x) > \phi_{p_j}(x) + \delta \qquad \text{for all } x\epsilon(c_1, c_2)$$

$$\text{or} \quad \mu_j(x) < \phi_{p_j}(x) - \delta \qquad \text{for all } x\epsilon(c_1, c_2) \quad .$$

Suppose the first is true. (The other case can be treated similarly.) By assumption (iv), for given $\delta > 0$, there exists an $\epsilon > 0$ such that

$$p_j < p < p_j + 2\epsilon \text{ implies } \phi_p(x) - \phi_{p_j}(x) < \delta \text{ for all } x \quad .$$

Choose $p = p_j + \epsilon$. Then

$$\phi_{p_j+\epsilon}(x) < \phi_{p_j}(x) + \delta \quad .$$

Hence, $$G_x(\phi_{p_j}(x) + \delta) > G_x(\phi_{p_j+\epsilon}(x)) = p_j + \epsilon \quad .$$

For convenience of notation, let $p_o = 0$, and $\phi_{p_o}(x) = \mu_o(x) = -\infty$ for all x. Now for any $x\epsilon(c_1, c_2)$,

$$P[U_{rs}^{(j)} = 1 \mid X_{rs} = x] = P[\mu_{j-1}(x) < Y_{rs} \leq \mu_j(x) \mid X_{rs} = x]$$

$$= P[\phi_{p_{j-1}}(x) < Y_{rs} \leq \mu_j(x) \mid X_{rs} = x], \text{ since } j \text{ is the smallest integer for which } \phi_{p_j} \neq \mu .$$

$$\geq P[\phi_{p_{j-1}}(x) < Y_{rs} \leq \phi_{p_j}(x) + \delta \mid X_{rs} = x]$$

$$= G_x(\phi_{p_j}(x) + \delta) - G_x(\phi_{p_{j-1}}(x))$$

$$> p_j + \epsilon - p_{j-1}$$

$$= q_j + \epsilon \quad .$$

Now let $c_1'' < c_1' < c_2' < c_2''$ be points in $(c_1, (c_1 + c_2)/2)$ and choose k_o sufficiently large such that for $k \geq k_o$, $F(c_1) > 1/k$, $1 - F(c_2) > 1/k$, $F(c_2') - F(c_1') > 1/k$, $F(c_1') - F(c_1'') > a_k$ and $F(c_2'') - F(c_2') > a_k$, where $\{a_k\}$ is a sequence as in lemma 1. Then for $k \geq k_o$, $F^{-1}(r/k) \leq c_1' < c_2' \leq F^{-1}(\frac{r+1}{k})$ implies $F(c_2') - F(c_1') \leq 1/k$, which is a contradiction. Hence for each $k \geq k_o$, there exists at least one integer $r(k)$ such that $c_1' < F^{-1}(\frac{r(k)}{k}) < c_2'$. Also for $k \geq k_o$, $F^{-1}(\frac{r(k)}{k} - a_k) \leq c_1''$ implies $F(c_1') - F(c_1'') \leq a_k$; which is a contradiction, and $F^{-1}(\frac{r(k)}{k}) + a_k) \geq c_2''$ implies $F(c_2'') - F(c_2') \leq a_k$ which is a contradiction. Hence for each $k \geq k_o$, there exists a smallest integer $r_1(k)$ such that

$$c_1 < F^{-1}(\frac{r_1(k)}{k} - a_k) < F^{-1}(\frac{r_1(k)}{k} + a_k) < (c_1 + c_2)/2 \quad .$$

Similarly, for each value of k greater than some integer, there exists a largest integer $r_2(k) > r_1(k)$ such that

$$(c_1 + c_2)/2 < F^{-1}(\frac{r_2(k)}{k} - a_k) < F^{-1}(\frac{r_2(k)}{k} + a_k) < c_2 \quad .$$

Hence for such large values of k,

$$F^{-1}\left(\frac{r_1(k)}{k} - a_k\right) < X_{r_1(k),n} < F^{-1}\left(\frac{r_1(k)}{k} + a_k\right)$$

and $$F^{-1}\left(\frac{r_2(k)}{k} - a_k\right) < X_{r_2(k),n} < F^{-1}\left(\frac{r_2(k)}{k} + a_k\right)$$

imply that $I_{kr} \subset (c_1, c_2)$ for $r = r_1(k) + 1, \ldots, r_2(k)$, and this set of integers is non-empty. Thus for sufficiently large values of k,

$$P\left[I_{kr} \subset (c_1, c_2),\ r = r_1(k) + 1, \ldots, r_2(k)\right]$$

$$\geq P\left[F^{-1}\left(\frac{r_i(k)}{k} - a_k\right) < X_{r_i(k),n} < F^{-1}\left(\frac{r_i(k)}{k} + a_k\right),\ i = 1, 2\right]$$

$$\geq 1 - 2ke^{-2nka_k^2}, \text{ by lemma 1.}$$

For each k greater than some sufficiently large integer, now choose and fix an integer r(k) such that $r_1(k) + 1 \leq r(k) \leq r_2(k)$. Then

$$P\left[\tau_{nk}^{(m)} > \tau_{nk}^{(m)}(\alpha)\right] \geq P\left[\tau_{nk}^{(m)} > \tau_{nk}^{(m)}(\alpha),\ I_{k,r(k)} \subset (c_1, c_2)\right]$$

$$\geq P\left[\tau_{nk}^{(m)} > \tau_{nk}^{(m)}(\alpha) \mid I_{k,r(k)} \subset (c_1, c_2)\right] x\left[1 - 2ke^{-2nka_k^2}\right]$$

$$\geq P\left[U_{r(k),o}^{(j)} - nq_j)^2/nq_j > \tau_{nk}^{(m)}(\alpha) \mid I_{k,r(k)} \subset (c_1, c_2)\right]$$

$$x\left[1 - 2ke^{-2nka_k^2}\right]$$

$$\geq P\left[U_{r(k),o}^{(j)} > nq_j + \sqrt{nq_j\tau_{nk}^{(m)}(\alpha)} \mid I_{k,r(k)} \subset (c_1, c_2)\right]$$

$$x\left[1 - 2ke^{-2nka_k^2}\right]$$

$$\geq \left[1 - e^{-2n\left(\epsilon - \sqrt{q_j\tau_{nk}^{(m)}(\alpha)/n}\right)^2}\right] x\left[1 - 2ke^{-2nka_k^2}\right],$$

by (2). To complete the proof we have only to note that if $n \geq (k \log k)^{1+1/m}$, then with the choice of $a_k = k^{-1/2}$, both factors in the final expression tend to 1.

6. Generalization to the case when X is vector-valued.

Suppose $(X, Y) = (X_1, \ldots, X_h, Y)$ follows an unknown multivariate distribution. Let F be the distribution function of $X = (X_1, \ldots, X_h)$ and for any $x = (x_1, \ldots, x_h)$, let G_x be the conditional distribution of Y given $X = x$. For given $0 < p < 1$, let $\phi_p(x)$ be the p-quantile of G_x.

The assumptions we make about F, $\{G_x\}$ and ϕ_p are the same as (i) - (iv) given in section 2, with some modifications. We shall only state the modified version of assumption (ii), the modifications on the other assumptions being obvious.

Modified assumption (ii). The marginal distribution of X_1 and the conditional distribution of X_i given $X_1 = x_1, \ldots, X_{i-1} = x_{i-1}$ $(0 \leq x_j \leq 1, j = 1, \ldots, i-1)$, $i = 2, \ldots, h$, are all continuous and strictly increasing.

In sections 2, 3 and 4, for a sample of size nk from a bivariate (X, Y), we defined a random division of $[0, 1]$ into k sub-intervals with the help of the fractiles of X observations. Studying the convergence of the sample fractiles to the corresponding population fractiles, we were able to give a probability inequality for the event that the length of each of these random intervals is less than a specified quantity. Due to the uniform continuity and some other regular nature of ϕ_p, the rest was accomplished by investigating the behavior of the p-quantile of the Y-observations corresponding to the X-observations belonging to each one of these random intervals, separately.

When X is vector-valued, there is only a partial ordering among the X's. We therefore modify our procedure as follows. Let $(X_{11}, \ldots, X_{h1}, Y_1), \ldots, (X_{1N}, \ldots, X_{hN}, Y_N)$ be N independent observations on $(X_1, \ldots, X_h, Y)$, where

$N = nk^h$. Let $N_j = nk^{h-j}$, $j = 1, \ldots, h$. First, let us arrange th X_1-coordinates of all observations in increasing order of magnitude and let $X_{1(i)}$ be the i-th order statistic obtained from $X_{11}, \ldots, X_{1N}$. Let us divide the interval $[0, 1]$ into random sub-intervals as follows. $I_1 = [0, X_{1(N_1)}]$, $I_{r_1} = (X_{1(\overline{r_1-1}\cdot N_1)}, X_{1(r_1 N_1)}]$, $r_1 = 2, \ldots, k-1$, and $I_k = (X_{1(\overline{k-1}\cdot N_1)}, 1]$. For each $r_1 = 1, \ldots, k$, consider all samples $(X_{1i}, \ldots, X_{hi}, Y_i)$ for which $X_{1i} \in I_{r_1}$ and call them the samples belonging to the r_1-th fractile group. Now for each $j = 1, \ldots, h-1$ and for each $r_1 = 1, \ldots, k; \ldots; r_j = 1, \ldots, k$, arrange the X_{j+1}-coordinates of all samples belonging to the $(r_1, \ldots, r_j)$-th fractile group, in increasing order of magnitude, and with the N_{j+1}-th, $2N_{j+1}$-th, $\ldots$, $(k-1)N_{j+1}$-th order statistics obtained from the X_{j+1}-coordinates of these samples, divide the interval $[0, 1]$ into random sub-intervals in the same way as $I_1, \ldots, I_k$ were defined, and call these intervals $I_{r_1 \ldots r_j 1}, \ldots, I_{r_1 \ldots r_j k}$. For each $r_1 = 1, \ldots, k; \ldots; r_{j+1} = 1, \ldots, k$, consider all samples $(X_{1i}, \ldots, X_{hi}, Y_i)$ which belong to the $(r_1, \ldots, r_j)$-th fractile group and for which $X_{j+1,i} \in I_{r_1 \ldots r_j r_{j+1}}$, as belonging to the $(r_1, \ldots, r_j, r_{j+1})$-th fractile group. Finally, when for each $r_1 = 1, \ldots, k; \ldots, r_h = 1, \ldots, k$, we have exactly n observations belonging to the $(r_1, \ldots, r_h)$-th fractile group, let us arrange the Y-coordinates of all samples belonging to the $(r_1, \ldots, r_k)$-th fractile group, in increasing order of magnitude, and suppose the order statistics obtained from the Y-coordinates of these samples are, $Y_{r_1 \ldots r_h(1)} < \ldots < Y_{r_1 \ldots r_h(n)}$. Also, for any specified real valued function μ defined on the h times Cartesian product of $[0, 1]$ with itself, let

$$U_i(r_1, \ldots, r_h) = \begin{cases} 1 & \text{if } (X_{1i}, \ldots, X_{hi}, Y_i) \epsilon I_{r_1} x I_{r_1 r_2} x \ldots x I_{r_1 r_2 \ldots r_h}, \\ & \text{and if } Y_i \leq \mu(X_{1i}, \ldots, X_{hi}) \\ 0 & \text{otherwise} \end{cases},$$

and let $\bar{U}(r_1, \ldots, r_h) = \sum_{i=1}^{N} U_i(r_1, \ldots, r_h)/n$.

Now we define a random function

$$f_{nk}(x_1, \ldots, x_h) = Y_{r_1 \ldots r_h(\lceil np \rceil)} \text{ if } x_1 \epsilon I_{r_1}, x_2 \epsilon I_{r_1 r_2}, \ldots, x_h \epsilon I_{r_1 r_2 \ldots r_h},$$
$$r_1 = 1, \ldots, k; \ldots, r_h = 1, \ldots, k,$$

and a statistic

$$\tau_{nk} = \sup_{\substack{r_1 = 1, \ldots, k; \ldots, \\ r_h = 1, \ldots, k.}} \frac{\left| \bar{U}(r_1, \ldots, r_h) - p \right|}{\sqrt{p(1-p)/n}} .$$

It can be easily verified that statements regarding the convergence of f_{nk} to ϕ_p, the limiting distribution of τ_{nk} (properly standardized) under the null hypothesis H_0: $(\phi_p = \mu)$ and the consistency of the test " reject H_0 if and only if $\tau_{nk} > \tau_{nk}(\alpha)$, where $\tau_{nk}(\alpha)$, $0 < \alpha < 1$, is such that

$$\lim_{k \longrightarrow \infty} P\lceil \tau_{nk} \leq \tau_{nk}(\alpha) \mid H_0 \rceil = 1 - \alpha \text{ " },$$

made as in theorems 1, 2 and 3 with k replaced by $K = k^h$, are valid.

For the case of several conditional quantile functions, the results of sections 3 and 4 can be extended for the multivariate case in the same way.

Thanks are due to S. N. Roy and Sudhindra N. Ray for making some helpful suggestions.

REFERENCES.

[1] Esseen,C. G. (1945). Fourier analysis of distribution functions. A mathematical study of the Laplace-Gaussian law. Acta Math. 77, 1-125.

[2] Hoeffding, Wassily (1962). Probability inequalities for sums of bounded random variables. Institute of Statistics, University of North Carolina, Mimeograph Series No. 326.

[3] Mahalanobis, P. C. (1958). Lectures in Japan: Fractile graphical analysis. Indian Statistical Institute Monograph.

[4] Parthasarathy, K. R. and Bhattacharya, P. K. (1961). Some limit theorems in regression theory. Sankhyā, Series A. 23, 91-102.

[5] Roy, S. N. and Cobb, Whitfield (1960). Mixed model variance analysis with normal error and possibly non-normal other random effects: Part I: The univariate case. Ann. Math. Statist. 31, 939-957.

[6] Sathe, Y. S. (1962). Studies of some problems in nonparametric inference. Institute of Statistics, University of North Carolina, Mimeograph Series No. 325.