Lecture Notes in Mathematics

Edited by A. Dold and B. Eckmann

821

Statistique non Paramétrique Asymptotique

Proceedings, Rouen, France 1979

Edité par J. P. Raoult

Springer-Verlag
Berlin Heidelberg New York

Lecture Notes in Mathematics

Edited by A. Dold and B. Eckmann

821

Statistique non Paramétrique Asymptotique

Actes des Journées Statistiques,
Rouen, France, Juin 1979

Edité par J. P. Raoult

Springer-Verlag
Berlin Heidelberg New York 1980

Editeur
Jean-Pierre Raoult
Université de Rouen – Mathématiques
B. P. 67
76130 Mont-Saint-Aignan/France

AMS Subject Classifications (1980):
Primary: 60F05, 62E20, 62G05, 62G10, 62G30
Secondary: 60G57, 62E15, 62H30

ISBN 3-540-10239-6 Springer-Verlag Berlin Heidelberg New York
ISBN 0-387-10239-6 Springer-Verlag New York Heidelberg Berlin

CIP-Kurztitelaufnahme der Deutschen Bibliothek
Statistique non paramétrique asymptotique: actes des Journées Statistiques, Rouen, France, juin 1979 / éd. par J. P. Raoult. – Berlin, Heidelberg, New York: Springer, 1980.
(Lecture notes in mathematics; Vol. 821)
ISBN 3-540-10239-6 (Berlin, Heidelberg, New York) brosch.
ISBN 0-387-10239-6 (New York, Heidelberg, Berlin)

Printing and binding: Beltz Offsetdruck, Hemsbach/Bergstr.
2141/3140-543210

INTRODUCTION

Ce volume regroupe 7 des exposés qui ont été présentés aux "Journées Statistiques", consacrées à la Statistique non paramétrique asymptotique, qui se sont tenues à Rouen (France), les 13 et 14 juin 1979, avec le soutien de la Société Mathématique de France et du C.N.R.S.

Nous avons voulu, à l'occasion de ces journées, aider les chercheurs à confronter leurs méthodes et leurs résultats dans ce domaine; il est apparu que la plupart des chercheurs participant à ces Journées travaillent sur les statistiques de rang, et seul le dernier exposé publié ici (celui de G. Collomb) ne relève pas de ce domaine.

On sait qu'on distingue traditionnellement deux "approches" dans l'étude des propriétés asymptotiques des statistiques de rang : l'approche "historique" dite de Chernoff et Savage, et l'approche dite de Pyke et Shorack, qui repose sur l'étude des processus empiriques multidimensionnels ; proposées initialement pour des suites de variables aléatoires indépendantes, identiquement distribuées et à lois diffuses, ces deux approches ont été ensuite explorées dans des cadres plus généraux; nombre des textes regroupés ici participent de ce type d'exploration.

C'est ainsi que le texte de F.H. Ruymgaart se situe dans la ligne de Chernoff et Savage; il s'affranchit des hypothèses d'identique distribution et de continuité des fonctions de répartition; les difficultés qu'il rencontre sont donc en particulier (dans la ligne des travaux de Conover ou Vorlīčkova) celles dues à la présence d'ex-aequo; il propose un traîtement de ce problème valable à la fois pour les trois types classiques de tests non paramétriques (indépendance des composantes, "randomness", symétrie).

Par contre les textes de S. Balacheff et G. Dupont, de M. Harel et de L. Rüschendorf sont justifiés par l'approche de Pyke et Shorack .

S. Balacheff et G. Dupont, de même que M. Harel, s'appuyant en particulier sur les résultats de Bickel et Wichura, considèrent des suites de variables aléatoires multidimensionnelles mélangeantes, non nécessairement identiquement distribuées, mais à fonctions de répartition continues. Conformément à une idée de Rüschendorf, idée qui permet d'élargir le cadre d'emploi de l'approche de Pyke et Shorack, ils s'intéressent aux processus empiriques et aux processus de rang tronqués, c'est-à-dire dont l'espace des "temps" possède, en plus des k dimensions liées à la nature de l'espace dans lequel sont effectuées les observations, une dimension supplémentaire, caractérisée par un paramètre, compris entre 0 et 1 , qui caractérise la proportion d'observations qui sont retenues dans les sommations à effectuer. S. Balacheff et G. Dupont démontrent la convergence de ces processus par une méthode fondée sur l'emploi du module de continuité multidimensionnel . M. Harel prolonge leur travail en s'intéressant aux processus tronqués et "corrigés" (en anglais "weighted"), c'est-à-dire où les trajectoires sont modifiées par division par une fonction du "temps" qui s'annulle là où s'annulent presque sûrement les trajectoires des processus multidimensionnels considérés; une telle procédure de correction est indispensable pour l'application de ces résultats à l'étude des statistiques linéaires de rang; elle est traitée par M. Harel selon des techniques introduites (pour des processus non tronqués)par Fears et Mehra (Ann. Stat. 1974), puis Mehra et M. Sudhakara Rao (seule la première partie -processus "semi-corrigés"- du travail de M. Harel est publiée dans le présent volume) .

L. Rüschendorf, pour sa part, étudie les processus empiriques multidimensionnels corrigés (weighted) dans le cadre "réduit", c'est-à-dire relativement à des suites de variables aléatoires i.i.d. et à loi uniforme sur le cube unité multidimensionnel ; l'intérêt de son travail tient non au résultat, classique, mais à une méthode originale et simple de démonstration, fondée sur une représentation de type Poissonnien (utilisant un résultat de Rosenblatt, 1975) .

Dans un cadre très voisin, P. Deheuvels poursuit ici l'étude d'un test d'indépendance(pour des suites de variables aléatoires p-dimensionnelles i.i.d. à marges diffuses), fondé sur le processus de dépendance empirique (qui ne dépend que des rangs), dont on mesure la distance à la fonction de dépendance théorique caractérisant l'indépendance (c'est-à-dire l'application "produit" de $[0,1]^p$ dans $[0,1]$); cette méthode peut -être considérée comme résultant de la conjonction des techniques standards de réduction de toute probabilité à marges diffuses sur $\mathbb{R}^p$ en une probabilité à marges uniformes sur $[0,1]^p$, et des idées présentées par Blum, Kiefer et Rosenblatt en 1961. P. Deheuvels fournit ici à la fois un théorème de convergence vers un processus gaussien et des exemples de calculs explicites permettant, pour certaines valeurs de p et de la taille de l'échantillon, la mise en oeuvre de ce test.

Tous les travaux que nous venons de citer visent essentiellement à obtenir des théorèmes de convergence vers des processus gaussiens; de plus en plus, les statisticiens souhaitent avoir des indications sur la puissance des tests qu'ils fondent sur de tels théorèmes; en particulier, ils souhaitent connaître des développements asymptotiques locaux de ces puissances; à ce sujet, on se réfère en particulier, pour ce qui est des tests de symétrie unidimensionnelle, au travail publié en 1976 par Albers, Bickel et Van Zwet . W. Albers le prolonge ici en s'intéressant,toujours pour tester la symétrie, aux tests linéaires de rang "adaptatifs", c'est-à-dire dans lesquels les fonctions de score utilisées sont déduites des observations pour s'adapter à la forme des probabilitées, symétriques autour d'un point autre que l'origine, qui sont supposées constituer la contre-hypothèse; les outils d'estimation employés sont suggérés par les travaux de Hogg, ainsi que de Shapiro, Wilk et Chen.

Enfin le texte de G. Collomb est le seul à ne pas porter sur les statistiques de rang, mais sur l'estimation de la régression; il est évident que, pour estimer la régression, en un point x (multidimensionnel), d'une variable aléatoire (unidimensionnelle), on doit faire intervenir les valeurs observées de la variable explicative avec un poids d'autant plus faible qu'elles sont loin de x ; dans les méthodes usuelles de noyaux (proposées par Nadaraya et Watson), ces poids font intervenir seulement, de manière "non adaptative", la distance à x; de son côté, Stone (en 1977) a étudié un estimateur reposant plus fortement sur la répartition empirique observée, car il s'agit de faire la moyenne des valeurs de la variable expliquée en les k valeurs de la variable explicative qui sont les plus proches de x . Le travail de G. Collomb consiste à réunir ces deux approches en introduisant des noyaux tels que le poids de chaque observation dépende de son rang, dans le classement, en ordre croissant, des valeurs le la variable aléatoire explicative.

Jean-Pierre RAOULT

<u>Exposés présentés lors des Journées des 13 et 14 juin 1979 à Rouen, et ne figurant pas dans ce volume</u> :

M. BERTRAND (Le Havre) Résultats récents sur l'estimation des densités

I. IBRAGIMOV (Léningrad) Asymptotic properties of some non parametric estimations in Gaussian white noise.

J.F. INGENBLECK (Bruxelles) Normalité asymptotique de certaines statistiques de rang sérielles.

G. NEUHAUS (Hamburg) Computing the distribution of Cramer-Von Mises statistics when parameters are present.

M.C.A. VAN ZUIJLEN (Nijmegen) Properties of the empirical distribution function for independent non identically random variables .

TABLE DES MATIERES

F.H. RUYMGAART. A unified approach to the asymptotic distribution theory of certain midrank statistics. 1

S. BALACHEFF, G. DUPONT. Normalité asymptotique des processus empiriques tronques et des processus de rang (Cas multidimensionnel melangeant) 19

M. HAREL. Convergence en loi pour la topologie de Skorohod du processus empirique multidimensionnel normalisé tronque et semi-corrigé (Etude au voisinage de la frontière inférieure de $[0,1]^{1+k}$) 46

L. RÜSCHENDORF. Weak convergence of weighted multiparameter empirical process 86

P. DEHEUVELS. Non parametric tests of independence 95

W. ALBERS. Asymptotic expansions for the power of adaptive rank tests in the one-sample problem 108

G. COLLOMB. Estimation de la regression par la methode des k points les plus proches avec noyeau: quelque propriétés de convergence ponctuelle. 159

1. INTRODUCTION

The purpose of this paper is to give a unified account of the asymptotic distribution theory of rank statistics for the classical problems of testing symmetry, randomness and independence, in the case of independent but not necessarily identically distributed sample elements (referred to as the non-i.i.d. case). In the case of continuous underlying distribution functions (d.f.'s) a unification of this kind has already been obtained in Ruymgaart and van Zuijlen (1978): more general in the sense that the sample elements are allowed to be of arbitrary finite dimension, on the other hand more restrictive because rank statistics for testing symmetry are not included. Here we shall emphasize the non-i.i.d. case for purely discrete underlying d.f.'s which is of practical interest. The non-i.i.d. case covers the asymptotic distribution theory under the hypothesis and under local as well as fixed alternatives as special cases. Our approach is patterned on the one first adopted in Chernoff and Savage (1958), to prove asymptotic normality of two-sample rank statistics under fixed alternatives in such a way that local alternatives could be dealt with too. Therefore we shall call this approach the Chernoff-Savage approach.

When the underlying d.f.'s are discrete, ties will in general be present and the ordinary notion of rank becomes unsatisfactory. Among the various ways to deal with ties (see e.g. Hájek (1969), Vorlíčková (1970,1972), Behnen (1973,1976,1978) and Conover (1973)) we shall restrict ourselves to the method of midranks which appears to be particularly well suited to the Chernoff-Savage approach. This is because using midranks we leave the score functions unaltered: the only modification is that the original ranks are replaced by the midranks. The Chernoff-Savage approach strongly hinges on certain properties of empirical d.f.'s and it turns out that the midranks are generated by a not altogether unnatural modification of the empirical d.f.. The desired properties of this modified empirical d.f. follow almost trivially from the results in van Zuijlen (1978); Section 3 of that basic paper contains all the essential information on empirical d.f.'s in the non-i.i.d. case for arbitrary underlying d.f.'s. The statistics obtained by replacing the ranks by the midranks will be called midrank statistics.

In the case of continuous underlying d.f.'s the midrank statistics that we shall consider are equivalent to the usual rank statistics and will consequently be ancillary for the relevant hypotheses. If the underlying d.f.'s are purely discrete, however, the midrank statistics are no longer ancillary but we shall prove them to be asymptotically ancillary.

Let us now be more specific and introduce the basic notation and conventions in order to give a summary of the results. For each $N \in \mathbb{N}$ let be given 2-dimensional mutually independent random vectors $(X_{1N}, Y_{1N}) = (X_1, Y_1)$, $(X_{2N}, Y_{2N}) = (X_2, Y_2), \ldots, (X_{NN}, Y_{NN}) = (X_N, Y_N)$. All the random vectors are supposed to be defined on the same probability space $(\Omega, \mathcal{A}, P)$. The bivariate d.f. of (X_n, Y_n) is $H_{nN} = H_n$ with marginals $F_{nN} = F_n$ (the d.f. of X_n) and $G_{nN} = G_n$ (the d.f. of Y_n). Observe that the triangular array $\{H_n\}$ completely specifies the asymptotic model. For each $N \in \mathbb{N}$ we define the averages $\bar{H}_N(x,y) = \bar{H}(x,y) = N^{-1} \Sigma_{n=1}^{N} H_n(x,y)$, $\bar{F}_N(x) = \bar{F}(x) = N^{-1} \Sigma_{n=1}^{N} F_n(x)$ and $\bar{G}_N(y) = \bar{G}(y) = N^{-1} \Sigma_{n=1}^{N} G_n(y)$ for $x, y \in \mathbb{R}$, of these d.f.'s. The empirical d.f. based on these N random vectors is defined in the usual way by $\hat{H}_N(x,y) = N^{-1}(\#\{n \le N: X_n \le x, Y_n \le y\})$; the marginal empirical d.f.'s of course satisfy $\hat{F}_N(x) = N^{-1}(\#\{n \le N: X_n \le x\})$, $\hat{G}_N(y) = N^{-1}(\#\{n \le N: Y_n \le y\})$ for $x,y \in \mathbb{R}$.

Let us next introduce the modified empirical d.f.'s

$$(1.1) \quad \hat{\Phi}_N(x) = \tfrac{1}{2}[\hat{F}_N(x)+\hat{F}_N(x-)], \quad \hat{\Psi}_N(y) = \tfrac{1}{2}[\hat{G}_N(y)+\hat{G}_N(y-)],$$

for $x,y \in \mathbb{R}$. The midrank $R_{nN} = R_n$ of X_n in the sample $X_1,X_2,\ldots,X_N$ and the midrank $Q_{nN} = Q_n$ of Y_n in the sample $Y_1,Y_2,\ldots,Y_N$ are defined by

$$(1.2) \quad R_n = N\hat{\Phi}_N(X_n), \quad Q_n = N\hat{\Psi}_N(Y_n), \quad n \in \{1,2,\ldots,N\}.$$

It will be convenient to dispose of the nonrandom functions

$$(1.3) \quad \overline{\Phi}_N(x) = \overline{\Phi}(x) = \tfrac{1}{2}[\overline{F}(x)+\overline{F}(x-)], \quad \overline{\Psi}_N(y) = \overline{\Psi}(y) = \tfrac{1}{2}[\overline{G}(y)+\overline{G}(y-)],$$

for $x,y \in \mathbb{R}$.

Let us note that

$$(1.4) \quad R_n = \#\{m \le N : X_m < X_n\} + \tfrac{1}{2}(\#\{m \le N : X_m = X_n\}),$$

$$(1.5) \quad R_n/N = \hat{\Phi}_N(X_n) \in \{1/(2N), 3/(2N), \ldots, (2N-1)/(2N)\},$$

for $n = 1,2,\ldots,N$. If no ties are present the R_n/N assume each of the values on the right in (1.5) exactly once, so that also in this case the midrank is still different from the ordinary rank defined as $N\hat{F}_N(X_n)$. We shall, however, exclusively use the midranks defined in (1.2). If there are ties the midranks are more natural than the ranks, and even if no ties are present the midranks are more pleasant to work with since they never assume the value N. Of course $\hat{\Psi}_N$ and the Q_n have similar properties.

Let $K,L : (0,1) \to \mathbb{R}$, and for each $N \in \mathbb{N}$ let $\phi_N, \psi_N : \mathbb{R} \to \mathbb{R}$ be measurable. Apart from a standardization of location and scale, the statistics that we are going to consider are

(1.6) $$S_N = \iint \phi_N(x)K(\hat{\Phi}_N(x)).\psi_N(y)L(\hat{\Psi}_N(y))d\hat{H}_N(x,y) =$$

$$= N^{-1}\Sigma_{n=1}^{N}\phi_N(X_n)K(R_n/N).\psi_N(Y_n)L(Q_n/N).$$

To standardize location and scale we shall use statistics that, in the case of continuous underlying d.f.'s, reduce to nonrandom numbers. Properly standardized these statistics turn out to be at least asymptotically ancillary for most of the purely discrete asymptotic models $\{H_n\}$ satisfying the generalized hypothesis, i.e. for which

(1.7) $\bar{H} = \bar{F} \times \bar{G} \quad \forall N \in \mathbb{N}$.

First let us show that the three standard hypotheses and the appropriate midrank statistics for testing these hypotheses can be obtained by suitable choices of $\{H_n\}$ and S_N.

To describe the problem of testing symmetry, let us consider a triangular array $\{\xi_{nN}\} = \{\xi_n\}$ of mutually independent random variables (r.v.'s). Choosing

(1.8) $(X_n, Y_n) = (|\xi_n|, \text{sgn}(\xi_n))$,

we obtain the model for the symmetry problem. The hypothesis of symmetry is that

(1.9) $\begin{cases} L(\xi_n) \text{ is independent of } n \text{ and } N, \text{ symmetric about } 0, \\ \text{with } P(\xi_n = 0) = 0. \end{cases}$

It is well known that in the continuous case (1.9) entails

(1.10) $\bar{H} = F \times G \quad \forall N \in \mathbb{N}$,

where $F = L(|\xi_n|)$, $G = L(\text{sgn}(\xi_n))$. In the purely discrete case relation (1.7) holds still true, provided the condition $P(\xi_n = 0) = 0$ is fulfilled. For the sake of uniformity this condition is therefore supposed to be part of the hypothesis of symmetry.

Assuming the model (1.8), statistics for testing the hypothesis (1.9) are obtained from (1.6) by setting $\phi_N = 1$ on $\mathbb{R}$ for all $N \in \mathbb{N}$, $\psi_N = \text{sgn}$ on $\mathbb{R}$ for all $N \in \mathbb{N}$, $L = 1$ on $(0,1)$. Then S_N reduces to

(1.11) $$S_N = N^{-1}[\Sigma_{n:\xi_n>0} K(R_n/N) - \Sigma_{n:\xi_n<0} K(R_n/N)],$$

which are the usual midrank statistics for testing symmetry.

For the problem of testing randomness let us take the triangular array $\{\xi_n\}$ as before and $\eta_{nN} = \eta_n$ the r.v. which is degenerate at the integer n. This entails independence of all the r.v.'s involved. If we choose

(1.12) $$(X_n, Y_n) = (\xi_n, \eta_n) = (\xi_n, n),$$

we obtain a model which is equivalent to the usual model for the problem of randomness since the Y_n are dummy variables. The <u>hypothesis of randomness</u> can be expressed by the assumption that

(1.13) $L(\xi_n)$ <u>is independent of</u> n <u>and</u> N.

Let us observe that if (1.13) is fulfilled we have

(1.14) $$\overline{H} = F \times \overline{G} = F \times (N^{-1}\Sigma_{n=1}^{\infty} 1_{[n,\infty)}) \qquad \forall N \in \mathbb{N},$$

where $F = L(\xi_n)$.

Assuming the model (1.12) and choosing $\phi_N = 1$ on $\mathbb{R}$ for all $N \in \mathbb{N}$ the S_N reduce to

(1.15) $$S_N = N^{-1}\Sigma_{n=1}^{N} c_n K(R_n/N),$$

where

(1.16) $$c_n = c_{nN} = \psi_N(n) L((2n-1)/(2N)).$$

These are the usual midrank statistics for testing randomness.

For a description of the problem of independence let $\{(\xi_{nN},\eta_{nN})\} = \{(\xi_n,\eta_n)\}$ be any triangular array of random vectors that are mutually independent in each row. The choice

$$(1.17) \quad (X_n,Y_n) = (\xi_n,\eta_n)$$

leads at once to the model for the independence problem. Note that in this case we could do without the auxiliary r.v.'s ξ_n and η_n. The <u>hypothesis of independence</u> is that

$$(1.18) \begin{cases} L((\xi_n,\eta_n)) \text{ is independent of } n \text{ and } N;\ \xi_n \text{ and } \eta_n \\ \text{are mutually independent.} \end{cases}$$

Assuming (1.17) and letting $\phi_N = \psi_N = 1$ on $\mathbb{R}$ for all $N \in \mathbb{N}$ we find

$$(1.19) \quad S_N = N^{-1}\Sigma_{n=1}^{N} K(R_n/N)L(Q_n/N),$$

the usual midrank statistics for testing independence.

It might be worthwhile to note that the class (1.6) also contains <u>linear combinations of functions of order statistics</u>. Let $\{\xi_n\}$ be an array of r.v.'s and let

$$(1.20) \quad (X_n,Y_n) = (\xi_n,\xi_n).$$

We shall also assume that

$$(1.21) \quad L(\xi_n) \text{ is continuous for all } n \text{ and } N.$$

Choosing $K = 1$ on $(0,1)$ and $\psi_N = 1$ on $\mathbb{R}$ for all $N \in \mathbb{N}$ leads to

$$(1.22) \quad S_N = N^{-1}\Sigma_{n=1}^{N} \phi_N(\xi_n)L(Q_n/N) = $$
$$= N^{-1}\Sigma_{n=1}^{N} L((2n-1)/(2N))\phi_N(\xi_{n:N}),$$

by the continuity of the $L(\xi_n)$. To study the rank statistics it will be convenient to impose a boundedness condition on the ϕ_N and ψ_N, a condition which is too restrictive to lead to sufficiently general results for the statistics in (1.22). We shall therefore exclusively deal with the theory of rank statistics. For the asymptotic distribution theory of linear combinations of functions of order statistics along the lines sketched above we refer to e.g. Moore (1968) and Ruymgaart and van Zuijlen (1977).

Next let us turn to the standardization under the generalized hypothesis (1.7) of the S_N in their general form (1.6). Let

$$(1.23) \quad \overline{\mu}_N = \overline{\mu} = \int\int \phi_N(x) K(\overline{\Phi}(x)) . \psi_N(y) L(\overline{\Psi}(y)) d\overline{H}(x,y),$$

and note that

$$(1.24) \quad \overline{\mu} = \overline{\kappa}\,\overline{\lambda} \quad \forall N \in \mathbb{N}, \text{ \underline{if} (1.7) \underline{is} \underline{satisfied},}$$

where

$$(1.25) \quad \overline{\kappa}_N = \overline{\kappa} = \int \phi_N K(\overline{\Phi}) d\overline{F}, \quad \overline{\lambda}_N = \overline{\lambda} = \int \psi_N (L(\overline{\Psi}) d\overline{G}.$$

Even if $\phi_N = \psi_N = 1$ on $\mathbb{R}$ the quantities $\overline{\kappa}$ and $\overline{\lambda}$ depend on $\overline{F}$ and $\overline{G}$ unless the underlying d.f.'s are continuous, an assumption that we are just trying to avoid. If $\overline{F}$ and $\overline{G}$ were known the location could be standardized by $\overline{\kappa}\,\overline{\lambda}$ under (1.7). The kind of problem that we wish to consider entails that we cannot assume any specific knowledge about $\overline{F}$ and $\overline{G}$ so that $\overline{\kappa}\,\overline{\lambda}$ has to be considered as unknown. We can, however, replace $\overline{\kappa}$ and $\overline{\lambda}$ by the estimators

$$(1.26) \quad \hat{\kappa}_N = \int \phi_N K(\hat{\Phi}_N) d\hat{F}_N, \quad \hat{\lambda}_N = \int \psi_N L(\hat{\Psi}_N) d\hat{G}_N.$$

This leads to considering the statistics

$$(1.27) \quad T_N = S_N - \hat{\kappa}_N \hat{\lambda}_N.$$

Under (1.7) these statistics turn out to be well standardized as far as location is concerned. To standardize the scale we are again confronted with a parameter that depends on the model. If in addition to (1.7) we impose a further technical condition this parameter is equal to $\bar{\sigma}^2\,\bar{\tau}^2$, where

$$(1.28)\quad \bar{\sigma}^2 = \int \phi_N^2 K^2(\bar{\Phi})\,d\bar{F}-\bar{\kappa}^2, \quad \bar{\tau}^2 = \int \psi_N^2 L^2(\bar{\Psi})\,d\bar{G}-\bar{\lambda}^2;$$

under this condition it can, moreover, be consistently estimated. Summarizing, it will appear that the statistics

$$(1.29)\quad \hat{S}_N = N^{\frac{1}{2}}(S_N-\hat{\kappa}_N\hat{\lambda}_N)/(\hat{\sigma}_N\hat{\tau}_N)$$

are asymptotically ancillary for the (suitably restricted) generalized hypothesis, where

$$(1.30)\quad \hat{\sigma}_N^2 = \int \phi_N^2 K^2(\hat{\Phi}_N)\,d\hat{F}_N-\hat{\kappa}_N^2, \quad \hat{\tau}_N^2 = \int \psi_N^2 L^2(\hat{\Psi}_N)\,d\hat{G}_N-\hat{\lambda}_N^2.$$

For almost arbitrary asymptotic models that do not necessarily satisfy the generalized hypothesis (1.7), the location of T_N in (1.27) is suitably standardized as

$$(1.30)\quad T_N-\bar{\mu}+\bar{\kappa}\,\bar{\lambda} = (S_N-\bar{\mu})-(\hat{\kappa}_N\hat{\lambda}_N-\bar{\kappa}\,\bar{\lambda}) .$$

Of course the location parameter now depends on the model, except when (1.7) is satisfied. The same holds true for the scale parameter.

In Section 2 we give the assumptions and a decomposition of $N^{\frac{1}{2}}(T_N-\bar{\mu}+\bar{\kappa}\,\bar{\lambda})$ which is of basic importance. Asymptotic standard normality of $\hat{S}_N$ in (1.29) is described in Section 3 for most asymptotic models satisfying the generalized hypothesis; this settles the asymptotic ancillarity. Section 4 is devoted to asymptotic normality under almost arbitrary asymptotic models. Finally in the Appendix we gather some useful lemmas on the modified empirical d.f..

2. ASSUMPTIONS AND BASIC DECOMPOSITION

To formulate the assumptions we have to introduce some notation. Throughout the number $c \in (0,\infty)$ will denote a generic constant; in particular this number is independent of the asymptotic model $\{H_n\}$ and also does not depend on the sample size N. Let us define the functions

(2.1) $r(s) = [s(1-s)]^{-1}, \qquad s \in (0,1);$

(2.2) $\delta(z) = 0$ <u>for</u> $z < 0$, $= \frac{1}{2}$ <u>for</u> $z = 0$, <u>and</u> $= 1$ <u>for</u> $z > 0$.

Observe that for $x,y \in \mathbb{R}$ we have

(2.3) $\hat{\Phi}_N(x) = N^{-1}\Sigma_{n=1}^{N} \delta(x-X_n), \quad \hat{\Psi}_N(y) = N^{-1}\Sigma_{n=1}^{N} \delta(y-Y_n),$

from which easily follows

(2.4) $E(\hat{\Phi}_N(x)) = \overline{\Phi}(x), \qquad E(\hat{\Psi}_N(y)) = \overline{\Psi}(y).$

Given an asymptotic model $\{H_n\}$, let $\{F_n\}$ and $\{G_n\}$ be the arrays of marginal d.f.'s. The union of the sets of discontinuity points of the individual F_n in the array will be denoted by $\{x_\nu : \nu = 1,2,\ldots\}$; relative to $\{G_n\}$ the set $\{y_\nu : \nu = 1,2,\ldots\}$ has a similar meaning. Let us write

(2.5) $$\begin{cases} \overline{p}_{\nu N} = \overline{p}_\nu = \overline{F}(x_\nu)-\overline{F}(x_\nu -), & \hat{p}_{\nu N} = \hat{F}_N(x_\nu)-\hat{F}_N(x_\nu -); \\ \overline{q}_{\nu N} = \overline{q}_\nu = \overline{G}(y_\nu)-\overline{G}(y_\nu -), & \hat{q}_{\nu N} = \hat{G}_N(y_\nu)-\hat{G}_N(y_\nu -), \end{cases}$$

for $\nu = 1,2,\ldots$.

The first set of assumptions concerns the functions involved in the definition of S_N in (1.6). <u>The functions</u> $K,L : (0,1) \to \mathbb{R}$ <u>are continuously differentiable on</u> $(0,1)$ <u>with</u>

(2.6) $|K^{(i)}| \le c\, r^{\alpha+i}$ <u>and</u> $|L^{(i)}| \le c\, r^{\beta+i}$ <u>for</u> $i = 0,1,$

<u>and some numbers</u> $\alpha,\beta \in [0,\frac{1}{2})$. <u>The functions</u> $\phi_N,\psi_N : \mathbb{R} \to \mathbb{R}$ <u>are measurable with</u>

(2.7) $|\phi_N| \le c, \quad |\psi_N| \le c \qquad \forall N \in \mathbb{N}.$

The second set of assumptions concerns the model. <u>The marginal</u>

<u>arrays</u> $\{F_n\}$ <u>and</u> $\{G_n\}$ <u>of the asymptotic model</u> $\{H_n\}$ <u>satisfy</u>

(2.8) $\Sigma_\nu |\hat{p}_{\nu N}-\bar{p}_\nu| = o_p(1)$ <u>and</u> $\Sigma_\nu |\hat{q}_{\nu N}-\bar{q}_\nu| = o_p(1)$, <u>as</u> $N \to \infty$.

<u>In relation to the statistics these marginal arrays also satisfy</u>

(2.9) $\begin{cases} \bar{\kappa} \to \kappa_0 \in \mathbb{R} \text{ and } \bar{\sigma}^2 \to \sigma_0^2 \in (0,\infty), \text{ as } N \to \infty; \\ \bar{\lambda} \to \lambda_0 \in \mathbb{R} \text{ and } \bar{\tau}^2 \to \tau_0^2 \in (0,\infty), \text{ as } N \to \infty, \end{cases}$

<u>for some numbers</u> $\kappa_0, \lambda_0 \in \mathbb{R}$ <u>and</u> $\sigma_0^2, \tau_0^2 \in (0,\infty)$.

Let us now turn to the basic decomposition for which we need the following r.v.'s (writing K' instead of $K^{(1)}$, etc.):

$$A_{0N} = N^{\frac{1}{2}}\int \phi_N K(\bar{\Phi})\, d[\hat{F}_N-\bar{F}],$$
$$A_{1N} = N^{\frac{1}{2}}\int \phi_N (\hat{\Phi}_N-\bar{\Phi}) K'(\bar{\Phi})\, d\bar{F},$$
$$B_{0N} = N^{\frac{1}{2}}\int \psi_N L(\bar{\Psi})\, d[\hat{G}_N-\bar{G}],$$
$$B_{1N} = N^{\frac{1}{2}}\int \psi_N (\hat{\Psi}_N-\bar{\Psi}) L'(\bar{\Psi})\, d\bar{G},$$
$$C_{0N} = N^{\frac{1}{2}}\int\int \phi_N K(\bar{\Phi}) . \psi_N L(\bar{\Psi})\, d[\hat{H}_N-\bar{H}],$$
$$C_{1N} = N^{\frac{1}{2}}\int\int \phi_N (\hat{\Phi}_N-\bar{\Phi}) K'(\bar{\Phi}) . \psi_N L(\bar{\Psi})\, d\bar{H},$$
$$C_{2N} = N^{\frac{1}{2}}\int\int \phi_N K(\bar{\Phi}) . \psi_N (\hat{\Psi}_N-\bar{\Psi}) L'(\bar{\Psi})\, d\bar{H}.$$

Under the conditions of Theorem 3.1 or Theorem 4.1 these r.v.'s are not only well defined but they even satisfy certain moment conditions; cf. in particular (2.4) and Lemma 5.1.

Now we shall write

(2.10) $N^{\frac{1}{2}}(T_N-\bar{\mu}+\bar{\kappa}\,\bar{\lambda}) = \Sigma_{i=0}^{2} C_{iN} - \bar{\lambda}\, \Sigma_{i=0}^{1} A_{iN} - \bar{\kappa}\, \Sigma_{i=0}^{1} B_{iN} + \rho_N,$

where ρ_N is the remainder term. Let us note that if (1.7) is satisfied, both the left and the right side of (2.10) reduce to simpler expressions, viz.

(2.11) $N^{\frac{1}{2}}(T_N-\bar{\mu}+\bar{\kappa}\,\bar{\lambda}) = N^{\frac{1}{2}}T_N = N^{\frac{1}{2}}(S_N-\hat{\kappa}_N\hat{\lambda}_N) =$
$= C_{0N}-\bar{\lambda}A_{0N}-\bar{\kappa}B_{0N}+\rho_N.$