AN INEQUALITY CONCERNING TESTS OF FIT OF THE KOLMOGOROV-SMIRNOV TYPE

By Georges Vandewiele and Marc Noé

Free University of Brussels and M.B.L.E. Research Laboratory, Brussels

1. Introduction. Let $F_n(x)$ be the sumpolygon (empirical distribution function) of a sample of size n from a continuous distribution function $F(x)$. Let $K(x)$, $G_1(x)$, $G_2(x)$, $H_1(x)$, $H_2(x)$ be functions of x, such that for all x,

$$G_1(x) \geqq G_2(x); \qquad H_2(x) \geqq H_1(x).$$

The object of the present paper is to prove the following inequalities

$$\begin{aligned}(1)\qquad & P[\inf_x (F_n - K) \geqq 0 \mid \inf_x (G_1 - F_n) \geqq 0,\quad \inf_x (F_n - H_2) \geqq 0] \\ & \geqq P[\inf_x (F_n - K) \geqq 0 \mid \inf_x (G_2 - F_n) \geqq 0,\quad \inf_x (F_n - H_1) \geqq 0],\end{aligned}$$

$$\begin{aligned}(2)\qquad & P[\inf_x (K - F_n) \geqq 0 \mid \inf_x (G_1 - F_n) \geqq 0,\quad \inf_x (F_n - H_2) \geqq 0] \\ & \leqq P[\inf_x (K - F_n) \geqq 0 \mid \inf_x (G_2 - F_n) \geqq 0,\quad \inf_x (F_n - H_1) \geqq 0],\end{aligned}$$

where all probabilities are supposed to exist. Since these inequalities are symmetrical, it suffices to prove one of them.

These inequalities provide an approximation for the distribution of two-sided statistics of the Kolmogorov-Smirnov type. Such a distribution is written

$$P\{\sup_x n^{\frac{1}{2}}|F_n(x) - F(x)|\psi[F(x)] \leqq \lambda\}$$

or more generally

$$(3)\qquad P_n = P[\inf_x (G_2 - F_n) \geqq 0,\quad \inf_x (F_n - H_2) \geqq 0].$$

In order to approximate P_n, take $H_1(x)$ and $H_2(x)$ in (1) smaller than zero for all x and replace $K(x)$ in (1) by $H_2(x)$; similarly take $G_1(x)$ and $G_2(x)$ in (2) larger than 1 for all x and replace $K(x)$ in (2) by $G_1(x)$. One then easily obtains the upper bound

$$\begin{aligned}(4)\qquad P_n \leqq P_n'\, & P[\inf_x (G_2 - F_n) \geqq 0]\cdot P[\inf_x (F_n - H_2) \geqq 0] \\ & \cdot\{P[\inf_x (G_1 - F_n) \geqq 0]\cdot P[\inf_x (F_n - H_1) \geqq 0]\}^{-1}\end{aligned}$$

where $P_n' = P[\inf_x (G_1 - F_n) \geqq 0, \inf_x (F_n - H_1) \geqq 0]$. If now G_1 and H_1 are chosen close to G_2 resp. H_2, but such that P_n' is more easily calculable than P_n, then (4) provides an interesting approximation of (3). A lower bound can be found in a similar way.

Wald and Wolfowitz [3] and [4] have given the following two bounds for P_n

$$\begin{aligned}(5)\qquad & P_n \leqq P[\inf_x (G_2 - F_n) \geqq 0]\cdot P[\inf_x (F_n - H_2) \geqq 0], \\ & P_n \geqq P[\inf_x (G_2 - F_n) \geqq 0] + P[\inf_x (F_n - H_2) \geqq 0] - 1.\end{aligned}$$

Received 27 December 1966; revised 23 February 1967.

However they did not prove (5) but only conjectured it. As a matter of fact (5) constitutes the particular form of (4) corresponding to $G_1(x) \geqq 1$ and $H_1(x) \leqq 0$ for all x. These bounds constitute good approximations for large values of P_n, which are the more interesting for purposes of testing. Furthermore they have the advantage of reducing the two-sided case to the one-sided case. On the other hand any other choice of $G_1(x)$ and $H_1(x)$ in (4) provides a closer upper bound than (5).

As Professor W. Hoeffding pointed out to the authors, (5) also is a consequence of a theorem of Lehmann [2]. Moreover he observed that the same theorem implies an analogous inequality for the distribution of the two-sample statistics of the Kolmogorov-Smirnov type. More details will be given in Section 4. Sections 2 and 3 are devoted to the proof of (1).

2. Lemma.

LEMMA. *Let X be a many-dimensional random variable and S and T two measurable subsets of the range space of X. Let Y be a one-dimensional random variable. If for all y such that $a \leqq y \leqq b$*

$$P[X \varepsilon S \mid (X \varepsilon T) \cap (Y = y)]$$

is a non-increasing function of y, then

$$P[X \varepsilon S \mid (X \varepsilon T) \cap (y_0 \leqq Y \leqq y_1)]$$

is a non-increasing function of y_0 and y_1 when $a \leqq y_0 \leqq y_1 \leqq b$.

PROOF. Let (X', Y') have the conditional distribution of (X, Y) under the condition $X \varepsilon T$ and let

$$h(y) = P[X \varepsilon S \mid (X \varepsilon T) \cap (Y = y)] = P[X' \varepsilon S \mid Y' = y].$$

Then

$$\begin{aligned} &P[X \varepsilon S \mid (X \varepsilon T) \cap (y_0 \leqq Y \leqq y_1)] \\ &\qquad = P[X' \varepsilon S \mid y_0 \leqq Y' \leqq y_1] \\ &\qquad = \textstyle\int_{y_0 \leqq y \leqq y_1} h(y) \cdot dF_{Y'}(y) / \int_{y_0 \leqq y \leqq y_1} dF_{Y'}(y) \end{aligned} \tag{6}$$

where $F_{Y'}$ denotes the distribution function of Y'. Under the assumption that $h(y)$ is non-increasing in $[a, b]$, it has to be shown that (6) is a non-increasing function of y_0 and y_1 for $a \leqq y_0 \leqq y_1 \leqq b$, which is obviously true.

This proof has been suggested by Professor W. Hoeffding who pointed out that this lemma is closely related to Lemma 4 of Lehmann [2].

3. Proof of the inequality (1). Without loss of generality it will be supposed that $F(x) \equiv x$ for $0 \leqq x \leqq 1$. Let $X_1 \leqq X_2 \leqq \cdots \leqq X_n$ be the order statistics of a size n sample from $F(x)$. The requirement that

$$\inf_x [F_n(x) - K(x)] \geqq 0$$

is equivalent to the requirement that every X_j should not be larger than some well defined number k_j, i.e. $X_j \leqq k_j$, $j = 1, \cdots, n$, with $0 \leqq k_1 \leqq k_2 \leqq \cdots \leqq$

$k_n \leqq 1$. Similarly $\inf_x [F_n(x) - H(x)] \geqq 0$ is equivalent to $X_j \leqq h_j$, $j = 1, \cdots, n$, with $0 \leqq h_1 \leqq h_2 \leqq \cdots \leqq h_n \leqq 1$. On the other hand $\inf_x [G(x) - F_n(x)] \geqq 0$ is equivalent to $X_j \geqq g_j$, $j = 1, \cdots, n$, with $0 \leqq g_1 \leqq g_2 \leqq \cdots \leqq g_n \leqq 1$. Proving the inequality (1) then is equivalent to proving the following theorem.

THEOREM. *If* $g_j \leqq h_j$ *for all* j, *then*

$$P[\textstyle\bigcap_{j=1}^{n} (X_j \leqq k_j) \mid \bigcap_{j=1}^{n} (g_j \leqq X_j \leqq h_j)]$$

is a non-increasing function of $g_1, \cdots, g_n, h_1, \cdots, h_n$.

PROOF. The theorem holds if, for some j, $g_j > k_j$; in the sequel we suppose $g_j \leqq k_j$ for all j.

The theorem is evidently true for sample size 1. For proving it in general we will proceed by complete induction, supposing it is true for sample sizes $1, 2, \cdots, n - 1$.

Let i be any integer in $[1, n]$. We introduce the following vectors;

$$\begin{aligned}
X &= (X_1, \cdots, X_{i-1}, X_{i+1}, \cdots, X_n);\\
X' &= (X_1, \cdots, X_{i-1});\\
X'' &= (X_{i+1}, \cdots, X_n);\\
k &= (k_1, \cdots, k_{i-1}, k_{i+1}, \cdots, k_n);\\
k' &= (k_1, \cdots, k_{i-1});\\
k'' &= (k_{i+1}, \cdots, k_n);\\
c' &= (c, \cdots, c) \qquad (i - 1 \text{ components});\\
c'' &= (c, \cdots, c) \qquad (n - i \text{ components}).
\end{aligned}$$

In a similar way we define g, g', g'', h, h' and h''. An inequality of two vectors will mean the simultaneous inequality of all corresponding components, e.g.: $X' \leqq h'$ means

$$(X_1 \leqq h_1) \cap (X_2 \leqq h_2) \cap \cdots \cap (X_{i-1} \leqq h_{i-1}).$$

Let $g_{i-1} \leqq c \leqq h_{i+1}$ (it is understood that $g_0 = 0$ and $h_{n+1} = 1$) and

$$Q = P[X \leqq k \mid (g \leqq X \leqq h) \cap (X_i = c)].$$

If $i \neq 1$ and $i \neq n$,

$$\begin{aligned}
Q &= P[(X' \leqq k') \cap (X'' \leqq k'') \mid (g' \leqq X' \leqq h') \cap (g'' \leqq X'' \leqq h'') \cap (X_i = c)]\\
&= Q' \cdot Q'',
\end{aligned}$$

with

$$\begin{aligned}
Q' &= P[X' \leqq k' \mid (g' \leqq X' \leqq h') \cap (g'' \leqq X'' \leqq h'') \cap (X_i = c)],\\
Q'' &= P[X'' \leqq k'' \mid (X' \leqq k') \cap (g' \leqq X' \leqq h') \cap (g'' \leqq X'' \leqq h'') \cap (X_i = c)].
\end{aligned}$$

Since $g_j \leqq k_j$ for all j, Q'' is defined. Since X' and X'' are independent under the hypothesis $X_i = c$ (see e.g. Hajos and Rényi [1]), one has

$$Q' = P[X' \leqq k' \mid (g' \leqq X' \leqq h') \cap (X_i = c)],$$

$$Q'' = P[X'' \leqq k'' \mid (g'' \leqq X'' \leqq h'') \cap (X_i = c)].$$

Let $Y' = (Y_1, Y_2, \cdots, Y_{i-1})$ be the order statistic of a size $(i - 1)$ sample from a rectangular distribution within [0, 1]. Under the hypothesis $Y' \leqq c'$, Y' is distributed as the order statistic from a rectangular distribution within [0, c]. The same may be said of X' under the hypothesis $X_i = c$ (see e.g. Hajos and Rényi [1]). We thus have

$$Q' = P(Y' \leqq k' \mid (g' \leqq Y' \leqq h') \cap (Y' \leqq c')].$$

Similarly let Y'' be the order statistic of a size $(n - i)$ sample from a rectangular distribution within [0, 1]. We then obtain

$$Q'' = P[Y'' \leqq k'' \mid (g'' \leqq Y'' \leqq h'') \cap (Y'' \geqq c'')].$$

According to the present theorem applied to sample sizes smaller than n, Q' and Q'' are non-increasing functions of c when $g_{i-1} \leqq c \leqq h_{i+1}$. If $i = 1$ or $i = n$, we arrive easily at the same conclusion. Thus the same holds for Q.

Let

$$R = P[X \leqq k \mid (g \leqq X \leqq h) \cap (g_i \leqq X_i \leqq h_i)].$$

According to the lemma, R is a non-increasing function of g_i and h_i when $g_{i-1} \leqq g_i \leqq h_i \leqq h_{i+1}$. Let now

$$S = P[X_i \leqq k_i \mid (X \leqq k) \cap (g \leqq X \leqq h) \cap (g_i \leqq X_i \leqq h_i)],$$

$$T = P[(X \leqq k) \cap (X_i \leqq k_i) \mid (g \leqq X \leqq h) \cap (g_i \leqq X_i \leqq h_i)].$$

The probability S evidently is a non-increasing function of g_i and h_i. Since $T = R \cdot S$, T has the same property. As i is any integer in [1, n], the theorem is proved.

4. Two corollaries of a theorem of Lehmann (Hoeffding).

Two real-valued functions r and s of n arguments are called by Lehmann discordant for the ith coordinate if, considered as functions of the ith coordinate (with all other coordinates held fixed), they are monotone in opposite directions, i.e. either r non-decreasing and s non-increasing or the inverse.

Let $(T_1, U_1), \cdots, (T_n, U_n)$ be independent pairs of random variables such that

$$P(T_i \leqq t, U_i \leqq u) \geqq P(T_i \leqq t) \cdot P(U_i \leqq u) \tag{7}$$

for all u and t, and let

$$T = r(T_1, \cdots, T_n), \qquad U = s(U_1, \cdots, U_n).$$

Lehmann proves that, if r and s are discordant for all i, then

$$P(T \geqq t, U \geqq u) \leqq P(T \geqq t) \cdot P(U \geqq u)$$

for all u and t.

We now show that this theorem implies inequality (5). Let $T_1, \cdots, T_n$ be n independent random variables and let $F_n(x)$ be their sumpolygon. Let $U_i = T_i$. Define

$$T = r(T_1, \cdots, T_n) = \inf_x [F_n(x) - H(x)],$$
$$U = s(U_1, \cdots, U_n) = \inf_x [G(x) - F_n(x)],$$

where $G(x)$ and $H(x)$ are arbitrary functions of x. Then condition (7) is fulfilled and r and s are discordant for all i. This proves inequality (5) even in the case of discontinuous variables.

The same theorem of Lehmann implies an analogous inequality for the two-sample case. Let $T_1, \cdots, T_k$ be independent random variables with sumpolygon $F_1(x)$ and let $T_{k+1}, \cdots, T_n$ be independent random variables with sumpolygon $F_2(x)$. Let $U_i = T_i$ and define

$$D(x) = F_1(x) - F_2(x),$$
$$T = r(T_1, \cdots, T_n) = \inf_x [D(x) - H(x)],$$
$$U = s(U_1, \cdots, U_n) = \inf_x [G(x) - D(x)],$$

where $G(x)$ and $H(x)$ are arbitrary functions of x. Again condition (7) is fulfilled and r and s are discordant for all i. In particular when $G(x) \equiv -H(x) = a > 0$, one has

$$P[\sup_x |D(x)| \leqq a] \leqq \{P[\sup_x D(x) \leqq a]\}^2.$$

Acknowledgments. We wish to thank Professor V. Belevitch and Professor J. Teghem for encouragement and reading the manuscript. We also thank Professor W. Hoeffding to whom we are indebted for his remarks.

REFERENCES

[1] HAJOS, G. and RÉNYI, A. (1954). Elementary proofs of some basic facts concerning order statistics. *Acta Math. Acad. Sci. Hungary* **5** 1–6.

[2] LEHMANN, E. L. (1966). Some concepts of dependence. *Ann. Math. Statist.* **37** 1137–1153.

[3] WALD, A. and WOLFOWITZ, J. (1939). Confidence limits for continuous distribution functions. *Ann. Math. Statist.* **10** 105–118.

[4] WALD, A. and WOLFOWITZ, J. (1941). Note on confidence limits for continuous distribution functions. *Ann. Math. Statist.* **12** 118–119.