AD-A052 368 STATE UNIV OF NEW YORK AT BINGHAMTON DEPT OF MATHEMAT--ETC F/G 12/1
THE FUNCTIONAL RELATIONSHIP BETWEEN DISTRIBUTION FUNCTIONS WITH--ETC(U)
MAR 78 D L HANSON N00014-76-C-0791

UNCLASSIFIED NL

| OF |

AD A052368

END
DATE FILMED
5 -78
DDC

AD A052368

AD No. ___
DDC FILE COPY

State Univ. of New York at Binghamton
Mathematical Sciences Dept.

FINAL REPORT

The Functional Relationship between Distribution Functions with Applications

(ONR contract N00014-76-C-0791)
(Identifying number NR 042-355)

David L. Hanson
Principal Invistigator

March 1978

5 p.

New 410629

Let F and G be two distribution functions. Under certain circumstances there exists a function $h:(0,1) \to (0,1)$ such that $G = hF$. Steck and Zimmer originally proposed the "model" $G = hF$ for possible use in accelerated life testing. In that case G and F can be taken to be, respectively, the distribution functions of an "object" under normal conditions and under conditions causing accelerated aging. Presumably one would have a fairly good idea of F from testing; an estimate of h would then provide an estimate of G. It is also possible that h might be determined from theoretical considerations in which case an estimate of F would provide an estimate of G. Finally, it is possible that estimates of h could provide insights into the aging process itself.

The model $G = hF$ is "obviously" not restricted to use in accelerated life testing but is potentially applicable wherever two distributions are "naturally" related to one another. It could, for example, be used to investigate human response times under different sets of conditions (e.g., different accelerations), in learning theory, to investigate the effects of collaboration on the ability to perform tasks or on test score distributions, etc.

Suppose F and G are continuous and that they are strictly increasing on the sets $F^{-1}((0,1))$ and $G^{-1}((0,1))$, respectively. Then a function h such that $G = hF$ exists if and only if $G^{-1}((0,1)) \subset F^{-1}((0,1))$, and in this case $h(y) = G[F^{-1}(y)]$ for every y in $(0,1)$.

An estimator suggested by Steck and Zimmer for estimating h is obtained in the following way. Let $X_1,\cdots,X_m,Y_1,\cdots,Y_n$ be an independent collection of random variables such that the X_i's have distribution function F and the Y_j's have distribution function G. Order the combined sample using some "rule" to determine order in case of ties. Consider the random walk from (0,0) to (1,1) which, after i+k steps, is at the point (i/m, k/n) if i of the first i+k combined order statistics are X's and the other k of them are Y's. The j-th step of this random walk goes to the right an amount 1/m (up an amount 1/n) if the j-th combined order statistic is an X (a Y). The estimator of Zimmer and Steck, $\hat{h}$, is the right continuous function obtained from the "path" of this random walk and extended so that $\hat{h}(x) = 0$ if $x < 0$ and $\hat{h}(x) = 1$ if $x > 1$.

If h exists and F is continuous, then in making probability computations about $\hat{h}$ it suffices to assume that the X_i's are uniformly distributed on (0,1) and that G = h.

Steck and Zimmer suggested using confidence regions for h of the form

$$S_h = S_h(m,n,\varepsilon,\delta) = \{\hat{h}(x-\varepsilon)-\delta \leq h(x) \leq \hat{h}(x+\varepsilon)+\delta \quad \text{for all} \quad x\}$$
$$= \{h(x-\varepsilon)-\delta \leq \hat{h}(x) \leq h(x+\varepsilon)+\delta \quad \text{for all} \quad x\}.$$

Unfortunately, the probability of such a region depends on h and, even if one knows h is not easily computed. Let $\alpha_F = P(D_m \leq \varepsilon) \geq P\{\sup_x |F_m(x)-F(x)| \leq \varepsilon\}$ and $\alpha_G = P(D_n \leq \delta) \geq P\{\sup_x |G_n(x)-G(x)| \leq \delta\}$ be the usual coverage probabilities of the standard Kolmogorov-Smirnov confidence regions. Let P_h denote a probability computed under the assumption that G = hF is true. Steck and Zimmer showed that $P_h(S_h) \geq \alpha_F \cdot \alpha_G$ for all h. They also pointed out that this bound is too conservative to be of any practical use. They conjectured that

(A) $P_h(S_h) \geq 1 - (1 - \alpha_F)(1 - \alpha_G)$ for all h, and

(B) if $m^{1/2}\varepsilon \approx n^{1/2}\delta < 1$ then $P_h(S_h)$ is minimized when $h(x) = x$ for all x in (0,1).

We have counterexamples to both of these conjectures. The counterexample to conjecture (B) is also an example illustrating the "discreteness effect" whereby a small change in h can produce a fairly sizeable change in the coverage probability by excluding a large number of sample paths in the second representation for S_h. We have had no success to date in producing usable lower bounds for $P_h(S_h)$. We know quite a bit more about what won't work. Other estimators (besides $\hat{h}$) and other confidence regions (besides S_h) have been proposed and investigated, mainly in an effort to "remove corners" or smoothen things out; these investigations have been of a preliminary nature; no theoretical results have been obtained and only preliminary simulation work has been done.

Asymptotic results have been obtained. Under suitable regularity conditions it has been shown that

$$\left| P\{S_h(m,n,\varepsilon/\sqrt{m},\delta/\sqrt{n})\} - P\{-1 \leq \left[\frac{B_1(h(t))}{n^{1/2}} + h'(t)\frac{B_2(t)}{m^{1/2}}\right] \Big/ [\delta/\sqrt{n} + h'(t)\varepsilon/\sqrt{m}] \leq 1\} \right|$$

converges to zero uniformly as $n,m \to \infty$. B_1 and B_2 are independent Brownian bridges; B_1 has undergone a time shift. No work has been done on rates of convergence and no progress has been made on providing a uniform bound on the probabilities given in terms of the two Brownian bridges. Note that we can replace m and n by their ratio.

Under suitable regularity conditions, h is concave if and only if $G'(t)/F'(t)$ is non-decreasing. Thinking of G' and F' as the fatality rates in the G and F populations, respectively, this says that earlier (in time) the higher fatality rate is in the F population, later the higher fatality rate is in the G population, and as time progresses the fatality rate shifts from the F population to the G population. This seems to be a reasonable assumption in some situations. The estimator which goes through the points (0,0) and (1,1), which is concave, and which is (in a sense) the best least squares fit to Steck's estimator subject to the conditions just mentioned, would seem to be a logical choice for estimating concave h's. Computing this estimator for specific data is a quadratic programming problem. Several methods of computation were available but none seemed suitable for a simulation study in which many estimators must be computed. A class of estimators has been produced each of which is a "cross" between Hildreth's procedure and "optimization" holding certain coordinates fixed on "the boundary." It has been proved that each procedure in this class produces the desired estimator.