The Annals of Mathematical Statistics
1968, Vol. 39, No. 6, 2114–2117

THE DISTRIBUTION OF GALTON'S STATISTIC[1]

BY SHULAMITH GROSS AND PAUL W. HOLLAND

Harvard University

0. Summary. Let $X_{(1)} < \cdots < X_{(n)}$ and $Y_{(1)} < \cdots < Y_{(n)}$ be the order statistics of two independent random samples from the absolutely continuous distribution functions $F(x)$ and $G(y)$, respectively. Let T_n be the proportion of pairs, $(X_{(i)}, Y_{(i)})$, for which $X_{(i)} \geqq Y_{(i)}$. Tests of the equality of F and G based on T_n are among the oldest nonparametric procedures in the literature, going back at least to Galton's analysis of Darwin's data [3]. Hodges [5] showed the null distribution of nT_n to be uniform over $0, 1, \cdots, n$. Bickel and Hodges [1] treated the asymptotic distribution of the Lehmann estimate based on the one-sample version of T_n. In this note we use very elementary methods to derive expressions for the distribution and moments of T_n from which conditions for the consistency of tests based on T_n follow immediately. More generally we can show that (unnormalized) T_n always has an asymptotic distribution for any pair (F, G). This distribution is degenerate at zero if Y happens to be stochastically larger than X. We give informative expressions for the first two moments of this asymptotic distribution. Our technique is to express the distribution of T_n in terms of integrals of certain multinomial probabilities.

1. Results. Define $I(x, y) = 1$ if $x \geqq y$ and zero otherwise. Then $T_n = n^{-1} \sum_{k=1}^{n} I(X_{(k)}, Y_{(k)})$. Let f and g be the densities of F and G. We assume F^{-1} exists and define $h(u) = G(F^{-1}(u))$ for $0 \leqq u \leqq 1$. Then $h(u)$ is increasing on $[0, 1]$. For any pair $1 \leqq j \leqq n$ and any $0 \leqq u_1 < \cdots < u_n \leqq 1$ let $(M_1, \cdots, M_{j+1})$ and $(N_1, \cdots, N_{j+1})$ be independent multinomial random vectors with parameters $(n - j; u_1, u_2 - u_1, \cdots, u_j - u_{j-1}, 1 - u_j)$ and $(n; h(u_1), h(u_2) - h(u_1), \cdots, h(u_j) - h(u_{j-1}), 1 - h(u_j))$ respectively. Finally we define

$$(1) \quad p_{n,j}(u_1, \cdots, u_j) = P\{\textstyle\sum_{i=1}^{k} (N_i - M_i) \geqq k;\ k = 1, \cdots, j\} \qquad \text{for } 1 \leqq j \leqq n.$$

LEMMA 1. *For every* $1 \leqq m \leqq n$,

$$P\{nT = m\} = \sum_{j=m}^{n} (-1)^{j-m}\binom{j}{m} n!((n-j)!)^{-1} \int \cdots \int_{0 \leqq u_1 < \cdots < u_j \leqq 1} p_{n,j}(u_1, \cdots, u_j)\,du_1, \cdots, du_j .$$

PROOF. For every $1 \leqq m \leqq n$ we have

$$(2) \quad P\{nT_n = m\} = \sum_{j=m}^{n} (-1)^{j-m}\binom{j}{m} \cdot \sum_{1 \leqq i_1 < \cdots < i_j \leqq n} P\{X_{(i_1)} \geqq Y_{(i_1)}, \cdots, X_{(i_j)} \geqq Y_{(i_j)}\}.$$

Received 6 October 1967.

[1] This work was facilitated by a grant from the National Science Foundation (GS-341).

Now for $1 \leqq i_1 < i_2 < \cdots < i_j \leqq n$

$$\begin{aligned} &P\{X_{(i_1)} \geqq Y_{(i_1)}, \cdots, X_{(i_j)} \geqq Y_{(i_j)}\} \\ &= \int \cdots \int_{-\infty < y_1 < \cdots < y_j < \infty} P\{Y_{(i_1)} \leqq y_1, \cdots, Y_{(i_j)} \leqq y_j\} \\ &\qquad \cdot f_{i_1 \cdots i_j}(y_1 \cdots y_j) dy_1 \cdots dy_j, \end{aligned} \tag{3}$$

where $f_{i_1 \cdots i_j}(y_1, \cdots, y_j)$ is the joint density of $X_{(i_1)}, \cdots, X_{(i_j)}$ and is given by

$$\begin{aligned} &f_{i_1 \cdots i_j}(y_1, \cdots, y_j) \\ &= n! f(y_1) \cdots f(y_j)[(i_1 - 1)!(i_2 - i_1 - 1)! \cdots (n - i_j)!]^{-1} \\ &\qquad \cdot [F(y_1)]^{i_1 - 1}[F(y_2) - F(y_1)]^{i_2 - i_1 - 1} \cdots [1 - F(y_j)]^{n - i_j} \end{aligned} \tag{4}$$

for $y_1 < y_2 < \cdots < y_j$, and zero otherwise. Another easy calculation shows that for $1 \leqq i_1 < i_2 < \cdots < i_j \leqq n$ and $y_1 < y_2 < \cdots < y_n$

$$\begin{aligned} &P\{Y_{(i_1)} \leqq y_1, \cdots, Y_{(i_j)} \leqq y_j\} \\ &= \sum_{l_j = i_j}^{n} \cdots \sum_{l_2 = i_2}^{l_3} \sum_{l_1 = i_1}^{l_2} \binom{n}{l_1, l_2 - l_1, \cdots, n - l_j} \\ &\qquad \cdot [G(y_1)]^{l_1}[G(y_2) - G(y_1)]^{l_2 - l_1} \\ &\qquad\qquad \cdots [G(y_j) - G(y_{j-1})]^{l_j - l_{j-1}}[1 - G(y_j)]^{n - l_j}. \end{aligned} \tag{5}$$

Now putting (2), (3), (4), and (5) together and making the transformation $u_i = F(y_i)$, one obtains

$$\begin{aligned} P\{nT_n = m\} &= \sum_{j=m}^{n} (-1)^{j-m} \binom{j}{m} \\ &\qquad \int \cdots \int_{0 < u_1 < \cdots < u_j < 1} p_{n,j}^{*}(u_1, \cdots, u_j) du_1 \cdots du_j \end{aligned} \tag{6}$$

where

$$\begin{aligned} &p_{n,j}^{*}(u_1, \cdots, u_j) \\ &= \sum_{1 \leqq i_1 < \cdots < i_j \leqq n} \sum_{l_j = i_j}^{n} \\ &\qquad \cdots \sum_{l_1 = i_1}^{l_2} n![(i_1 - 1)! \cdots (n - i_j)!]^{-1} u_1^{\,i_1 - 1}(u_2 - u_1)^{i_2 - i_1 - 1} \\ &\qquad \cdots (1 - u_1)^{n - i_j} \binom{n}{l_1, l_2 - l_1, \cdots, n - l_j} [h(u_1)]^{l_1}[h(u_2) - h(u_1)]^{l_2 - l_1} \\ &\qquad \cdots [1 - h(u_j)]^{n - l_j}. \end{aligned} \tag{7}$$

If, in (7) we replace i_k by $i_k - k$, then $p_{n,j}^{*}(u_1, \cdots, u_j)$ becomes

$$\begin{aligned} &n![(n - j)!]^{-1} \sum_{i_1 = 0}^{i_2} \sum_{i_2 = i_1}^{i_3} \cdots \sum_{i_j = i_{j-1}}^{n-j} \sum_{l_1 = i_1 + 1}^{l_2} \sum_{l_2 = i_2 + 2}^{l_3} \cdots \sum_{l_j = i_j + j}^{n} \\ &\qquad \cdot \binom{n-j}{i_1, i_2 - i_1, \cdots, n - j - i_j} u_1^{\,i_1}(u_2 - u_1)^{i_2 - i_1} \cdots (1 - u_j)^{n - j - i_j} \\ &\qquad \cdot \binom{n}{l_1, l_2 - l_1, \cdots, n - l_j} [h(u_1)]^{l_1}[h(u_2) - h(u_1)]^{l_2 - l_1} \cdots [1 - h(u_j)]^{n - l_j}. \end{aligned} \tag{8}$$

Inspection of the summation in (8) reveals it to be $p_{n,j}(u_1, \cdots, u_j)$. □

Using Lemma 1 to compute the expectation of $T_n^{\,k}$ we easily obtain

COROLLARY 1. *If* $k \geqq 1$ *then*

$$(9)\quad E(T_n^{\,k}) = \sum_{j=1}^{\min[k,n]} \Delta^j(0^k)n![n^k(n-j)!]^{-1} \int\cdots\int_{0\leqq u_1<\cdots<u_j\leqq 1} p_{n,j}(u_1\cdots u_j)\,du_1\cdots du_j\,.$$

The numbers, $\Delta^j(0^k)$ are shorthand for $\Delta^j(X^k)\mid_{X=0}$ where Δ is the difference operator. Two useful properties of these numbers needed in the proofs of Corollaries 1 and 2 are: $\Delta^j(0^k) = 0$ if $j > k$ and $\Delta^k(0^k) = k!$ (see [6] page 36–50).

To show that the moments of T_n all converge we apply dominated convergence to the integrals in (9). To show $\lim_{n\to\infty} p_{n,j}(u_1, \cdots, u_j) = p_j(u_1, \cdots, u_j)$ exists we first define these subsets of [0, 1].

$$(10)\quad S_+ = \{u\colon h(u) > u\},\ S_- = \{u\colon h(u) < u\},\ S_0 = \{u\colon h(u) = u\}.$$

Observe that the vector $n^{-\frac{1}{2}}((N_1 - M_1) - n(h(u_1) - u_1), \cdots, \sum_{i=1}^{j}(N_i - M_i) - n(h(u_j) - u_j))$ has a limiting multivariate normal distribution with zero mean vector and covariance matrix, $\sum(u_1, \cdots, u_j)$, whose elements are given by

$$(11)\qquad \sigma_{l,m} = h(u_m)(1 - h(u_l)) + u_m(1 - u_l), \qquad 1 \leqq l \leqq m \leqq j.$$

If $u_i \varepsilon S_0$ for all $i = 1, \cdots, j$, $\sigma_{l,m}$ reduces to

$$(12)\qquad \sigma_{l,m} = 2u_m(1 - u_l).$$

There are several possibilities for $p_j(u_1, \cdots, u_j)$. (i) If $u_i \varepsilon S_-$ for some $i = 1, \cdots, j$ then $p_j(u_1, \cdots, u_j) = 0$. (ii) If $u_i \varepsilon S_+$ for all $i = 1, \cdots, j$, then $p_j(u_1, \cdots, u_j) = 1$. (iii) If $u_{i_1}, \cdots, u_{i_l} \varepsilon S_0$ while the remaining u's are in S_+, then $p_j(u_1, \cdots, u_j)$ is the probability content of the positive orthant in l-dimensional space given by the multivariate normal distribution with mean vector zero and covariances $\sigma_{i_\alpha i_\beta} = 2u_{i_\alpha}(1 - u_{i_\beta})$; $i_\alpha \leqq i_\beta$. This shows that $p_j(u_1, \cdots, u_j)$ exists and we obtain

COROLLARY 2. *If* $k \geqq 1$ *then*

$$(13)\quad \lim_{n\to\infty} E(T_n^{\,k}) = k!\int\cdots\int_{0\leqq u_1<\cdots<u_k\leqq 1} p_k(u_1, \cdots, u_k)\,du_1\cdots du_k\,.$$

Since T_n is bounded and all of its moments converge, T_n has an asymptotic distribution for any choice of F and G and the limiting moments are the moments of this limiting distribution. For $k > 2$, (13) does not readily simplify. For $k = 1, 2$, it may be applied to give interesting results. Let $\lambda(S)$ denote the Lebesgue measure of a set $S \subseteq [0, 1]$ and $I_S(x)$ denote the indicator function of S. Recall the fact (see Cramér [2], page 290, for example) that the probability content of the first quadrant of a central bivariate normal distribution with correlation ρ is given by $4^{-1} + (2\pi)^{-1}\sin^{-1}(\rho)$. We have

COROLLARY 3.

(a) $\lim_{n\to\infty} E(T_n) = \frac{1}{2}\lambda(S_0) + \lambda(S_+)$,

(b) $\lim_{n\to\infty} \mathrm{Var}\,(T_n)$

$$= \pi^{-1}\int_0^1\int_0^v \sin^{-1}(uv^{-1}(1 - v)(1 - u)^{-1})^{\frac{1}{2}} I_{S_0}(u)I_{S_0}(v)\,du\,dv.$$

Applying the last corollary we obtain

THEOREM 1. *T_n has a degenerate distribution if and only if* $\lambda(S_0) = 0$. *If* $\lambda(S_0) = 0$, *then T_n converges in probability to* $\lambda(S_+)$.

We observe that if $G(x) < F(x)$ for all x, then $\lambda(S_0) = \lambda(S_+) = 0$ so that rejection for small values of T_n yields a consistent test of the hypothesis $F = G$. By applying the fact that if $F = G$, T_n has an asymptotic uniform distribution on [0, 1] we obtain a simple evaluation of the definite integral

$$\int_0^1 \int_0^v \sin^{-1}(uv^{-1}(1 - v)(1 - u)^{-1})^{\frac{1}{2}} du\, dv = \tfrac{1}{12}\pi.$$

2. Remarks. The assumption of *absolute* continuity may be reduced to simple continuity by the following device, essentially given in [4]. (a) Find a distribution function H such that $H \gg F$ and $H \gg G$ (use $H = \frac{1}{2}(F + G)$ for example); (b) If T is a measurable transformation of the line into itself and H^*, F^*, and G^* are the induced distributions then $H^* \gg F^*$ and G^*; (c) Use H itself as T. Then if F and G are continuous so is H and therefore $H^*(x) \equiv x$ a.e. in [0, 1]. Hence F^* and G^* both possess densities with respect to Lebesgue measure in [0, 1]; (d) Since $H \gg F$ and G, is continuous and monotone (though not necessarily strictly monotone) T_n is unchanged under the transformation of the X's and Y's by H, with probability one.

Strict monotonicity of F was only used in the transformation of the integrals in (6). If $F^{-1}(y)$ is defined as $\inf \{x: F(x) \geqq y\}$ for $0 \leqq y \leqq 1$, this restriction may be removed. Because $h(u) = G(F^{-1}(u)) = G^*(F^{*-1}(u))$ it is unnecessary to actually transform the problem from (F, G) to (F^*, G^*) to use the results of this paper.

It may be possible to use the fact that $p_{n,j}(u_1, \cdots, u_j)$ is the probability of a large deviation when $\lambda(S_-) = 1$, and Lemma 1 to obtain more useful expressions for the distribution of T_n under the hypothesis of stochastic ordering. We have no results in this direction yet.

Finally, we would like to thank Dr. F. Mosteller for helpful discussions of this research.

REFERENCES

[1] BICKEL, P. J. and HODGES, J. L., JR. (1967). The asymptotic theory of Galton's test and a related simple estimate of location. *Ann. Math. Statist.* **38** 73–89.

[2] CRAMÉR, H. (1946). *Mathematical Methods of Statistics*. Princeton Univ. Press.

[3] DARWIN, C. (1878). *The Effect of Cross- and Self-fertilization in the Vegetable Kingdom* (2nd edition). John Murray, London.

[4] GOVINDARAJULU, Z., LECAM, L. and RAGHAVACHARI, M. (1967). Generalizations of theorems of Chernoff and Savage on the asymptotic normality of test statistics. *Proc. Fifth Berkeley Symp. Math. Statist. Prob.* **1** 609–638. Univ. of California Press.

[5] HODGES, J. L., JR. (1955). Galton's rank order test. *Biometrika* **42** 261–262.

[6] MILNE-THOMPSON, L. M. (1951). *The Calculus of Finite Differences*. Macmillan, New York.