The Annals of Statistics
1982, Vol. 10, No. 4, 1306–1310

SOME NONDEGENERATE LIMIT LAWS FOR THE SELECTION DIFFERENTIAL

BY H. N. NAGARAJA

The Ohio State University

The difference between the average of the top k out of n order statistics and the population mean expressed in population standard deviation units is known as the selection differential. This paper obtains some nondegenerate limit laws for this quantity. The results are applied to the construction of tables used in testing for outliers.

1. Introduction. Let $X_{1:n} \leq X_{2:n} \leq \cdots X_{n:n}$ denote the order statistics of a random sample of size n from a distribution with distribution function (df) F, mean μ and variance σ^2. Suppose we select the k largest values in the sample. The difference between the average of the selected group and population mean expressed in standard deviation units is called the *selection differential*, and may be written as $D_{k,n} = k^{-1} \sum_{i=n-k+1}^{n} (X_{i:n} - \mu)/\sigma$. The selection differential has long been a familiar term to geneticists and breeders and is a useful measure of improvement due to selection in breeding problems (see, e.g., Burrows, 1972). It also serves as a good test statistic in testing for outliers from normal populations.

Throughout this paper we assume that μ and σ are known and without loss of generality take $\mu = 0$ and $\sigma = 1$ in our discussion. We obtain nondegenerate limit laws for $D_{k,n}$ in the following cases: (i) the "extreme case", where k is held fixed, and n becomes infinitely large; (ii) the "quantile case" where $k = [np]$, $0 < p < 1$ and $n \to \infty$. Here $[\cdot]$ is the greatest integer function. The extreme case is closely associated with extreme value theory, whereas in the quantile case, the asymptotic results on linear functions of order statistics are relevant. Finally, we compare the approximate percentage points for $D_{k,n}$ as given by these limit laws with those obtained by simulation in Barnett and Lewis (1978).

2. Extreme case. Suppose there exist constants a_n and $b_n > 0$ such that

$$P\{(X_{n:n} - a_n)/b_n \leq x\} = F^n(a_n + b_n x) \to G(x) \tag{1}$$

as $n \to \infty$, where G is a nondegenerate df. We then write $F \in D(G)$. Gnedenko (1943) has shown that G can be one of the three types of distributions denoted by Φ_α, Ψ_α and Λ. From Lamperti (1964) it follows that if (1) holds, then for each $k \geq 1$, the vector $((X_{n:n} - a_n)/b_n, (X_{n-1\cdot n} - a_n)/b_n, \cdots, (X_{n-k+1:n} - a_n)/b_n)$ has, in the limit, the joint distribution of $(T_1, \cdots, T_k)$, where a canonical representation for T_i's in terms of standard exponential random variables (rv's) is given by Hall (1978). Since $(D_{k,n} - a_n)/b_n = k^{-1} \sum_{i=1}^{k} (X_{n-i+1:n} - a_n)/b_n$ is a continuous function of these components, it is immediate that $(D_{k,n} - a_n)/b_n \to_{\mathscr{L}} (T_1 + \cdots + T_k)/k = D_k$, say. Using Hall's (1978) representation for the T_i's, one can write a representation for D_k. It turns out that only when $G = \Lambda$, can one write the closed form expression for the df of D_k (see Nagaraja, 1980). It is given by the following result.

THEOREM 1. *If $F \in D(\Lambda)$, then $(D_{k,n} - a_n)/b_n$ converges in law to a* rv *D_k whose* df *for $-\infty < x < \infty$ and $k \geq 2$ is given by*

$$F_k(x) = \frac{k^{k-1}}{(k-2)!} \sum_{j=0}^{k-1} \frac{e^{-xj}}{j!} \int_0^\infty \exp\{-\exp(u - x)\}\exp\{-u(k-j)\}u^{k-2}\, du. \tag{2}$$

Received February 1981; revised April 1982.

AMS 1980 *subject classifications.* Primary, 62G30, 62E20; secondary, 62F03.

Key words and phrases. Order statistics, selection differential, test for outliers.

Institute of Mathematical Statistics is collaborating with JSTOR to digitize, preserve, and extend access to *The Annals of Statistics.* www.jstor.org

Direct manipulation of the representation given by Hall proves the result. For details see Nagaraja (1980). Numerical computations show that this limit distribution is positively skewed and its density becomes more peaked as k increases.

3. Quantile case. Here we assume that $k = [np]$, $0 < p < 1$, and derive the asymptotic distribution of $D_{k,n}$, appropriately normalized, as $n \rightarrow \infty$. The limit distribution depends on whether the qth quantile of F is unique where $q = 1 - p$. When it is unique, the limit distribution can be obtained as a corollary to the limit law for linear functions of order statistics. Several authors have considered this problem, imposing different restrictions on F and the linear function. However, the most general result for $D_{k,n}$ with the least restrictions on F is obtained by appealing to Stigler (1973), since $D_{k,n}$ is a trimmed mean. He does not require a unique qth quantile and his result is given below.

THEOREM 2 (Stigler, 1973). *Let $a = \sup\{x: F(x) \leq q\}$ and $A = a - \inf\{x: F(x) \geq q\}$. Then as $n \rightarrow \infty$ (with $\mu = 0$, $\sigma = 1$),*

$$\sqrt{k}\,(D_{k,n} - \mu_p) \rightarrow_{\mathscr{L}} Y_1 + (a - \mu_p) Y_2 - A \max(0, Y_2) \tag{3}$$

where Y_1 is $N(0, \sigma_p^2)$, Y_2 is $N(0, q)$ and Y_1 and Y_2 are independent. Here μ_p and σ_p^2 are the mean and variance of the distribution obtained by truncating F below at a.

If $A = 0$, $a = \xi_q$, the unique quantile, (3) can be written as

$$\sqrt{k}\,(D_{k,n} - \mu_p) \rightarrow_{\mathscr{L}} N(0, \sigma_p^2 + q(\mu_p - \xi_q)^2). \tag{4}$$

We now investigate the case when k is not exactly $[np]$, but is fairly close. To be precise, when $\sqrt{n}\,(p - k/n)$ converges to a finite constant, we find the asymptotic distribution of $D_{k,n}$.

THEOREM 3. *If $\sqrt{n}\,(p - k/n) \rightarrow c$, a finite constant, then*

$$\sqrt{k}\,(D_{k,n} - \mu_p) \rightarrow_{\mathscr{L}} N(c(\mu_p - \xi_q)/\sqrt{p}, \sigma_p^2 + q\,(\mu_p - \xi_q)^2) \tag{5}$$

if ξ_q is the unique qth quantile.

PROOF. Without loss of generality we take $k \neq [np]$ in the proof and let $S_{k,n} = \sum_{i=n-k+1}^{n} X_{i:n} \equiv kD_{k,n}$. Then it can be seen that

$$\min(X_{n-k:n}, X_{n-[np]:n}) \leq \frac{S_{k,n} - S_{[np],n}}{k - [np]} \leq \max(X_{n-k:n}, X_{n-[np]:n}).$$

Note that $X_{n-k:n} \rightarrow_{\mathscr{P}} \xi_q$ and $X_{n-[np]:n} \rightarrow_{\mathscr{P}} \xi_q$ (Smirnov, 1952, page 9). Hence $(S_{k,n} - S_{[np],n})/(k - [np]) \rightarrow_{\mathscr{P}} \xi_q$ and consequently as $n \rightarrow \infty$ $(S_{k,n} - S_{[np],n})/\sqrt{k} \rightarrow_{\mathscr{P}} (-c)\xi_q/\sqrt{p}$. Now

$$\sqrt{k}\,(D_{k,n} - \mu_p) = \{(S_{k,n} - S_{[np],n})/\sqrt{k}\} + \{[np]/\sqrt{k}\}\{(S_{[np],n}/[np]) - \mu_p\}$$
$$+ \{\mu_p([np] - k)/\sqrt{k}\},$$

where the first term converges to $-\xi_q c/\sqrt{p}$ in probability, the second term converges in law to $N(0, \sigma_p^2 + q(\mu_p - \xi_q)^2)$ from (4) and the last term tends to $c\mu_p/\sqrt{p}$. Hence we obtain (5).

4. Application to testing for outliers. Let $X_1, X_2, \cdots, X_n$ be independent rv's, with $X_i \sim N(\mu_i, \sigma^2)$, $i = 1, 2, \cdots, n$. Consider the problem of testing the hypothesis H: $\mu_1 = \mu_2 = \cdots \mu_n = \mu$ against the alternative A: k of the μ_i's are equal to $\mu + \delta(\delta > 0)$ and the remainder are equal to μ. In this outlier testing problem $kD_{k,n} \equiv (S_{k,n} - k\mu)/\sigma$ is used as a test statistic when μ and σ are known. In fact, when μ and σ are estimated by $\bar{X}$ and S, Barnett and Lewis (1978, pages 95-96) point out that the test which rejects H for large values of $(S_{k,n} - k\bar{X})$ has some desirable properties.

TABLE 1
Percentile points of the df F_k *of Theorem* 1.

k	2	3	4
$\xi_{k,.95}$	1.800	1.154	0.715
$\xi_{k,.99}$	2.813	1.933	1.364

TABLE 2
Values of the norming constants for selected n.

n	a_n	a_n^*	b_n	b_n^*
30	1.8882	1.9146	.3834	.5223
50	2.1009	2.1118	.3575	.4735
100	2.3663	2.3753	.3295	.4210
500	2.9075	2.9080	.2836	.3439
1000	3.1165	3.1153	.2690	.3210

Because of the above motivation, considerable attention has been given to the percentage points of $D_{k,n}$. Since the distribution of $D_{k,n}$ is free from μ and σ, without loss of generality we take $\mu = 0$, $\sigma = 1$. We compare approximations to 95th and 99th percentiles of $D_{k,n}$ under H obtained by using the asymptotic theory assuming (a) k fixed, (b) $k = [np]$, $0 < p < 1$, and the ones computed from Table IXg of Barnett and Lewis (1978). That table gives percentage points for $kD_{k,n}$ and is obtained by simulation.

(a) When k is fixed since the standard normal df, $\Phi \in D(\Lambda)$, the 100sth percentile point, $\xi_{k,s}$, of the limiting distribution of $(D_{k,n} - a_n)/b_n$ can be obtained from (2). It is given in Table 1 for $s = .95$ and .99.

Now the problem is to use "good" choices of a_n and b_n. Often these are given as (see, e.g., Galambos, 1978, page 65)

$$a_n = (2 \log n)^{1/2} - [\{\log(4\pi \log n)\}/\{2(2 \log n)^{1/2}\}], \tag{6}$$

$$b_n = (2 \log n)^{-1/2}.$$

It is also known that any other sequence a_n' and b_n' such that $b_n/b_n' \to 1$ and $(a_n - a_n')/b_n \to 0$ as $n \to \infty$ would serve asymptotically. But Hall (1979) has shown that the best rate of convergence of $\sup_{-\infty<x<\infty} |\Phi^n(a_n + b_n x) - \Lambda(x)|$ is achieved when a_n and b_n are chosen such that

$$2\pi a_n^2 \exp(a_n^2) = n^2 \text{ and } b_n = 1/a_n. \tag{7}$$

Let a_n^* and b_n^* be the solutions of (7). Table 2 illustrates the differences in a_n and b_n as given by (6) and a_n^*, b_n^*.

The approximate percentage points of $D_{k,n}$ are then given by $a_n + b_n\xi_{k,s}$ and $a_n^* + b_n^*\xi_{k,s}$ for the two choices of constants, and are denoted by Ext(a_n, b_n) and Ext(a_n^*, b_n^*), respectively.

(b) For given n and k we can take $p = (k/n)$ and use (4) of the quantile case set-up. For a normal parent, Burrows (1975) has tabulated the limiting variance $\sigma_D^2 \equiv \sigma_p^2 + q(\mu_p - \xi_q)^2$ for various values of p. Also $\mu_p = \phi(\xi_q)/p$, where ϕ is the standard normal density. Burrows (1972) has also obtained an approximation to $E(D_{k,n})$ which converges to μ_p at the rate of $1/n$. Hence we can use his approximation $\hat{\mu}_p = \mu_p - [(n - k)/\{2\mu_p k(n + 1)\}]$ instead of μ_p in (4). These give another pair of approximations to the percentage points of $D_{k,n}$, namely $\mu_p + z_\alpha\sigma_D/\sqrt{k}$ and $\hat{\mu}_p + z_\alpha\sigma_D/\sqrt{k}$ where z_α is the upper 100α percentile point of Φ. These are denoted by Qnt(μ_p) and Qnt$(\hat{\mu}_p)$ respectively.

We compare the above four approximations to the simulated 95th and 99th percentile points of $D_{k,n}$ obtained from Table IXg of Barnett and Lewis (1978). These simulated points are labeled Sim (B&L). Table 3 gives these five approximations for $k = 2, 3, 4$ and $n = 20, 30, 40, 50, 100$.

DISCUSSION. The empirical evidence expressed in Table 3 shows that Ext(a_n, b_n) does much better than Ext(a_n^*, b_n^*) for all n, k and the percentages considered in the sense that it is much closer to Sim (B&L) than the latter. Even though a_n^* and b_n^* are supposed to

TABLE 3

n	95% points					99% points				
	Ext (a_n, b_n)	Ext (a_n^*, b_n^*)	Qnt (μ_p)	Qnt $(\hat{\mu}_p)$	Sim (B&L)	Ext (a_n, b_n)	Ext (a_n^*, b_n^*)	Qnt (μ_p)	Qnt $(\hat{\mu}_p)$	Sim (B&L)
$k = 2$										
20	2.44	2.78	2.46	2.34	2.37	2.86	3.36	2.76	2.64	2.72
30	2.58	2.85	2.61	2.50	2.51	2.97	3.38	2.89	2.78	2.84
40	2.67	2.92	2.70	2.58	2.62	3.05	3.42	2.97	2.85	2.93
50	2.74	2.97	2.78	2.67	2.68	3.10	3.45	3.04	2.93	3.02
100	2.96	3.13	3.00	2.90	2.92	3.29	3.56	3.23	3.13	3.20
$k = 3$										
20	2.18	2.41	2.17	2.08	2.10	2.50	2.85	2.43	2.34	2.39
30	2.33	2.52	2.33	2.25	2.26	2.63	2.92	2.57	2.49	2.54
40	2.43	2.60	2.44	2.36	2.38	2.72	2.98	2.67	2.59	2.63
50	2.51	2.66	2.52	2.44	2.45	2.79	3.03	2.74	2.66	2.72
100	2.75	2.86	2.76	2.69	2.70	3.00	3.19	2.96	2.89	2.94
$k = 4$										
20	2.00	2.15	1.96	1.89	1.90	2.26	2.53	2.20	2.13	2.16
30	2.16	2.29	2.14	2.08	2.08	2.41	2.63	2.36	2.30	2.32
40	2.27	2.38	2.26	2.20	2.21	2.51	2.70	2.46	2.40	2.43
50	2.36	2.46	2.34	2.28	2.28	2.59	2.76	2.54	2.48	2.53
100	2.60	2.68	2.60	2.54	2.55	2.82	2.95	2.78	2.72	2.78

Five approximations to the percentage points of $D_{k,n}$ for the normal parent population.

make the convergence of the df of $X_{n:n}$ faster in the sense of the supremum over the entire real line, Ext(a_n^*, b_n^*) does not perform as well as Ext(a_n, b_n) at the 95th and 99th percentile points of $D_{k,n}$. Further, the quantile approach seems to yield better approximation than the extreme approach. At 95 percent level Qnt$(\hat{\mu}_p)$ comes closest to Sim (B&L), but is always less than the latter. Also, at this level Ext(a_n, b_n) and Qnt(μ_p) approach each other as n increases for $k \geq 3$ even though both are off from Sim (B&L). But at the 99 percent level Qnt(μ_p) does very well indeed, the better with increased k for given n.

5. Concluding remarks. In this paper we considered two cases where k is fixed and $k = [np]$, $0 < p < 1$, while obtaining the limit distribution of $D_{k,n}$. One may also consider the asymptotically extreme case where $k \to \infty$ but $k = o(n)$. We are as yet unable to obtain the limit distribution for an arbitrary F under this setup. For the standard exponential parent, however, we can show that $\sqrt{k}\,\{D_{k,n} - \log(n/k)\} \to_{\mathscr{L}} N(0, 2)$. Our proof uses the independence of the spacings and the representation for $X_{n-k:n}$ in terms of standard exponential rv's (see Renyi, 1953, page 194).

Acknowledgments. This note is based on a portion of my Ph.D. dissertation written at Iowa State University which was supported by a grant from the U.S. Army Research Office. I am indebted to my thesis advisor, Dr. H. A. David, for his guidance and to the referee and an Associate Editor for their helpful comments.

REFERENCES

BARNETT, V. and LEWIS, T. (1978). *Outliers in Statistical Data.* Wiley, Chichester.

BURROWS, P. M. (1972). Expected selection differentials for directional selection. *Biometrics* **28** 1091–1100.

BURROWS, P. M. (1975). Variances of selection differentials in normal samples. *Biometrics* **31** 125–133.

GALAMBOS, J. (1978). *The Asymptotic Theory of Extreme Order Statistics.* Wiley, New York.

GNEDENKO, B. (1943). Sur la distribution limite du terme maximum d'une série aléatoire. *Ann. Math.* **44** 423–453.

HALL, P. (1978). Representations and limit theorems for extreme value distributions. *J. Appl. Probab.* **15** 639–644.

HALL, P. (1979). On the rate of convergence of normal extremes. *J. Appl. Probab.* **16** 433–439.

LAMPERTI, J. (1964). On extreme order statistics. *Ann. Math. Statist.* **35** 1726–1737.

NAGARAJA, H. N. (1980). Contributions to the theory of the selection differential and to order statistics. Unpublished Ph.D. dissertation, Iowa State University.

RENYI, A. (1953). On the theory of order statistics. *Acta Math. Acad. Sci. Hung.* **4** 191–231.

SMIRNOV, N. V. (1952). Limit distributions for the terms of a variational series. *Amer. Math. Soc. Transl. Ser.* **1** No. 67.

STIGLER, S. M. (1973). The asymptotic distribution of the trimmed mean. *Ann. Statist.* **1** 472–477.

DEPARTMENT OF STATISTICS
THE OHIO STATE UNIVERSITY
1958 NEIL AVENUE
COLUMBUS, OHIO 43210