THE SAMPLE MEAN AMONG THE MODERATE ORDER STATISTICS[1]

By Herbert T. David

Iowa State University

0. Summary. The first part of this paper (Section 2) is devoted to the derivation of the asymptotic distribution of the sample proportion above a normal sample mean. The second part of the paper, leaning on the first, treats the asymptotic joint distribution of runs of various lengths above the sample mean (Section 3). The approach here proves general enough to cover asymptotic run distributions arising when the dichotomy criterion is magnitude relative to a sample function other than the sample mean and the population is other than normal.

1. Introduction. This is one of two papers investigating the magnitude of the sample mean relative to the order statistics. Emphasis here is placed on the intermediate order statistics. Specifically, we study the asymptotic form of $\sum_{i=1}^{k} P(n, i)$, $(k \sim \frac{1}{2} n)$, where $P(n, i)$ is, as in [5], the probability that a normal sample mean will fall between the ith and $(i + 1)$st order statistics. This study leads in a natural way to investigating the asymptotic joint distribution of runs of various lengths above the sample mean, and the approach proves broad enough to cover run distributions where the dichotomy criterion is magnitude relative to sample functions other than the sample mean, and where the population is other than normal. This generality of the study arises from the fact that all such run distributions asymptotically are convolutions of the distributions of two vectors, one following the multivariate normal distribution derived by Mood in the first part of [8], the other following a singular (one-dimensional) distribution which is a function of the dichotomy criterion.

Previous work on the distribution theory of runs includes the (distribution-free) study of runs above a fixed sample quantile by Mood [8], and also by F. N. David [4], Mosteller [9], Stevens [12], Swed and Eisenhart [13], and Wald and Wolfowitz [14]. It includes as well the (distribution-free) study of runs above a fixed population quantile, notably by Mood [8], and by von Mises [7].

A possible application of the distribution theory of runs above the sample mean is to the testing of the homogeneity of a random sample, with seemingly good power against the two commonly most feared alternatives to homogeneity: one-sided outliers and trend. Indeed, positive (negative) outliers will lead to a preponderance of observations below (above) the sample mean, leading in turn to a dearth of runs above the sample mean, for lack of representation in one of the two run-producing classes. Again, trend will of course tend to depress the number of runs with respect to *any* dichotomy criterion.

It is true that a homogeneity test based on runs above the sample mean would

Received April 4, 1960; revised August 30, 1960.

[1] This paper is part of a dissertation submitted to the Department of Statistics of the University of Chicago in partial fulfillment of the requirements for the Ph.D. degree.

be most useful if applicable to the correlated residuals of a least-square model, and the requisite distribution theory should be developed. However, pending such development, the present theory is of some practical usefulness as it stands when applied to suitable subsets of degrees of freedom, say to the paired differences of a paired-comparison experiment. In a similar vein, it is applicable, in randomized complete block experiments involving substantial numbers of blocks, to the "replicates" obtained when the same treatment contrast is separately computed in each block.

I wish to thank L. J. Savage, who very generously gave of his knowledge and ideas in guiding this research. I wish to thank as well P. Billingsley and W. H. Kruskal, who were equally generous in guiding the composition of the final draft. I also am much indebted to M. G. Kendall, W. H. Kruskal, P. A. P. Moran, F. Mosteller, H. Ruben, D. L. Wallace, W. A. Wallis and W. J. Youden for their critical reading of earlier drafts of this paper, and for many bibliographical suggestions.

2. The asymptotic distribution of the sample proportion above a normal sample mean. The object of this section is to determine the asymptotic distribution of π_n, the proportion above the mean in a random sample of size n from a normal population. It will be shown that

$$\mathfrak{L}(n^{\frac{1}{2}}(\pi_n - \tfrac{1}{2})) \rightarrow N(0, \tfrac{1}{4} - 1/(2\pi)), \tag{2.1}$$

where $N(0, \frac{1}{4} - 1/(2\pi))$ represents a normal c.d.f. with zero mean and variance $(\frac{1}{4} - 1/(2\pi))$, the c.d.f. convergence of (2.1) being uniform in any finite interval and hence on the entire line by the continuity of the limit c.d.f.

The distribution of π_n clearly does not depend on the population mean and variance, which will for convenience be assumed equal to zero and one respectively.

Consider a random sample of size n from a normal population with zero mean and variance unity. Let z_p and ζ_p be, respectively, the sample and population quantiles of order p. Let π_n be the proportion of the sample exceeding the sample mean, and let $p(n) = \frac{1}{2} + tn^{-\frac{1}{2}}$, $q(n) = \frac{1}{2} - tn^{-\frac{1}{2}}$, t any real number. Then

$$\begin{aligned} \Pr\{n^{\frac{1}{2}}(\pi_n - \tfrac{1}{2}) \leqq t\} &= \Pr\{\pi_n \leqq p(n)\} \\ &= \Pr\{z_{q(n)} \leqq \bar{x}\} = \Pr\{z_{q(n)} - \bar{x} \leqq 0\}. \end{aligned} \tag{2.2}$$

The vector $v = (x_1 - \bar{x}, \cdots, x_n - \bar{x})$ is independent of $\bar{x}$, and $(z_{q(n)} - \bar{x})$ is a function of v. Hence $a_n = n^{\frac{1}{2}}[(z_{q(n)} - \bar{x}) - \zeta_{q(n)}]$ and $b_n = n^{\frac{1}{2}}\bar{x}$ are independent, so that

$$\gamma_n(s) = \alpha_n(s) \cdot \beta_n(s), \tag{2.3}$$

where α_n and β_n are, respectively, the characteristic functions of a_n and b_n, and γ_n is the characteristic function of $c_n = a_n + b_n = n^{\frac{1}{2}}(z_{q(n)} - \zeta_{q(n)})$.

A slight modification and specialization of Section 28.5 of [2] shows that the densities of the random variables $(2/\pi)^{\frac{1}{2}}c_n$ tends to the unit-normal density.

(The modification involves considering $(z_{q(n)} - \zeta_{q(n)})$ in place of $(z_q - \zeta_q)$, while the specialization is from a general density to the unit-normal density.)

Since the densities of $(2/\pi)^{\frac{1}{2}}c_n$ tend to the unit-normal density, the c.d.f.'s of $(2/\pi)^{\frac{1}{2}}c_n$ tend to the unit-normal c.d.f., by Scheffé's Theorem [11]. Hence, by the continuity theorem, the characteristic functions of $(2/\pi)^{\frac{1}{2}}c_n$ tend to $e^{-s^2/2}$, or

$$\lim_{n\to\infty} \gamma_n(s) = \exp(-\pi s^2/4). \tag{2.4}$$

Further, $\beta_n(s) = e^{-s^2/2}$, so that, taking limits in (2.3) and using (2.4),

$$\exp(-\pi s^2/4) = (\lim_{n\to\infty} \alpha_n(s))(\exp(-s^2/2)).$$

The continuity theorem thus implies that

$$\mathcal{L}(a_n) \to N(0, \pi/2 - 1), \tag{2.5}$$

where $N(0, \pi/2 - 1)$ represents a normal c.d.f. with zero mean and variance $\pi/2 - 1$.

Recapitulating from (2) (with A_n the c.d.f. of a_n), we have

$$\begin{aligned} \Pr\{n^{\frac{1}{2}}(\pi_n - \tfrac{1}{2}) \leqq t\} &= \Pr\{z_{q(n)} - \bar{x} \leqq 0\} \\ &= \Pr\{n^{\frac{1}{2}}[(z_{q(n)} - \bar{x}) - \zeta_{q(n)}] \leqq -n^{\frac{1}{2}}\zeta_{q(n)}\} = A_n(-n^{\frac{1}{2}}\zeta_{q(n)}), \end{aligned} \tag{2.6}$$

and, by the Taylor expansion about zero of the unit-normal c.d.f. evaluated at $\zeta_{q(n)}$,

$$\lim_{n\to\infty}(-n^{\frac{1}{2}}\zeta_{q(n)}) = t(2\pi)^{\frac{1}{2}}. \tag{2.7}$$

Hence, by the uniformity of the convergence of (2.5), and by (2.6) and (2.7), $\lim_{n\to\infty}[\Pr\{n^{\frac{1}{2}}(\pi_n - \frac{1}{2}) \leqq t\}] = [N(0, \pi/2 - 1)$ evaluated at $t(2\pi)^{\frac{1}{2}}]$, which implies (2.1).

3. The asymptotic distribution of runs above a normal sample mean.

LEMMA 1. (Chernoff) *Let Y_n andY be r-dimensional random vectors such that $\mathcal{L}(Y_n) \to \mathcal{L}(Y)$, i.e., the* c.d.f. *$F_n$ of Y_n tends point-wise to the* c.d.f. *F of Y at every continuity point of F. Let ϕ be a continuous function from r-space to s-space. Then $\mathcal{L}(\phi(Y_n)) \to \mathcal{L}(\phi(Y))$.*

Lemma 1 is Theorem 4 of [10]. Essentially the same result may be derived by appealing first to Theorem 2.1 of [1], and then to Theorem 1 of [3].

Let π_n be a random variable the range of which is a finite discrete subset V_n of the closed unit interval. Suppose that for each $p \varepsilon V_n$, $R_{p,n}$ is a k-dimensional random vector. Then $\bar{R}_n \equiv R_{\pi_n,n}$ is also a k-dimensional random vector. Let $\pi_n^* = n^{\frac{1}{2}}(\pi_n - p_0)$, where $0 < p_0 < 1$, and suppose that

$$\mathcal{L}(\pi_n^*) \to \mathcal{L}(\pi). \tag{3.1}$$

Let $X_{p,n} = n^{-\frac{1}{2}}B(p)(R_{p,n} - na(p))$, where $B(p)$ and $a(p)$ are, respectively, $k \times k$ matrix-valued and k-dimensional vector-valued functions on the closed unit interval. Suppose that $\mathcal{L}(X_{p,n}) \to \mathcal{L}(X)$ uniformly in p near p_0 in the following sense. If β is a continuity point of the c.d.f. of X, then there is some positive ϵ (perhaps depending on β) such that

$$\lim_{n\to\infty} \sup_{|p-p_0|<\epsilon,\, p\,\varepsilon\, V_n} |F_{p,n}(\beta) - F(\beta)| = 0, \tag{3.2}$$

where $F_{p,n}$ and F are, respectively, the c.d.f.'s of $X_{p,n}$ and X.

THEOREM 1. *Suppose that* (3.1) *and* (3.2) *hold, and that, if* $p \,\varepsilon\, V_n$, *then* $R_{p,n}$ *and* π_n *are independent. Suppose further that* $B(p)$ *is component-wise continuous and non-singular, and that the components* $a_i(p)$ *of* $a(p)$ *have derivatives* $\dot{a}_i(p_0)$ *at* p_0. *Then*

$$\mathcal{L}(n^{-\frac{1}{2}}(\bar{R}_n - na(p_0))) \to \mathcal{L}(B^{-1}(p_0)X + \dot{a}(p_0)\pi), \tag{3.3}$$

where X *and* π *are to be taken independent.*

PROOF. Let

$$\bar{X}_n = X_{\pi_n,n} = n^{-\frac{1}{2}}B(\pi_n)(\bar{R}_n - na(\pi_n)).$$

If $p \,\varepsilon\, V_n$, then, by the independence assumption,

$$\Pr\{\bar{X}_n \leqq \beta, \pi_n = p\} = \Pr\{X_{p,n} \leqq \beta, \pi_n = p\} = \Pr\{X_{p,n} \leqq \beta\}\cdot\Pr\{\pi_n = p\}.$$

Hence, for t a continuity point of the c.d.f. of π,

$$\Pr\{\bar{X}_n \leqq \beta, \pi_n^* \leqq t\} = \sum_{\substack{p \leqq p_0 + tn^{-\frac{1}{2}} \\ p\,\varepsilon\, V_n}} \Pr\{X_{p,n} \leqq \beta\}\cdot P\{\pi_n = p\}.$$

It follows from (3.2) that, if β is a continuity point of the c.d.f. of X, the difference between the right-hand side of this expression and that of

$$\Pr\{X \leqq \beta\}\cdot\Pr\{\pi_n^* \leqq t\} = \sum_{\substack{p \leqq p_0 + tn^{-\frac{1}{2}} \\ p\,\varepsilon\, V_n}} \Pr\{X \leqq \beta\}\cdot\Pr\{\pi_n = p\}$$

tends to zero as n tends to infinity, so that, working with the corresponding left-hand sides,

$$\lim_{n\to\infty} |\Pr\{\bar{X}_n \leqq \beta, \pi_n^* \leqq t\} - \Pr\{X \leqq \beta\}\cdot\Pr\{\pi_n^* \leqq t\}| = 0. \tag{3.4}$$

It now follows from (3.1) and (3.4) that

$$\lim_{n\to\infty} |\Pr\{\bar{X}_n \leqq \beta, \pi_n^* \leqq t\} - \Pr\{X \leqq \beta\}\cdot\Pr\{\pi \leqq t\}| = 0, \tag{3.5}$$

i.e., since (β, t) now is a continuity point of the joint c.d.f. of X and π,

$$\mathcal{L}(\bar{X}_n, \pi_n^*) \to \mathcal{L}(X, \pi), \tag{3.6}$$

where X and π are to be taken independent. Now, by the definition of $\bar{X}_n$, we have the vector equation

$$B^{-1}(\pi_n)\bar{X}_n = n^{-\frac{1}{2}}(\bar{R}_n - na(p_0)) - n^{-\frac{1}{2}}(na(\pi_n) - na(p_0)),$$

where, by the lemma on page 777 of [6], the last term differs from $-\dot{a}(p_0)\pi_n^*$ by a term ϵ_n with $\operatorname{plim}_n \epsilon_n = 0$. Hence

$$n^{-\frac{1}{2}}(\bar{R}_n - na(p_0)) = B^{-1}(\pi_n)\bar{X}_n + \dot{a}(p_0)\pi_n^* + \epsilon_n. \tag{3.7}$$

But by (3.6), the fact that $\operatorname{plim}_n \pi_n = p_0$, and Lemma 1, (with ϕ taking the $(k+1)$-dimensional vector $((x_1, \cdots, x_k), x_{k+1})$ into the $2k$-dimensional vector $(B^{-1}(p_0)(x_1, \cdots, x_k), \dot{a}(p_0)x_{k+1}))$,

$$(3.8) \qquad \mathcal{L}(B^{-1}(\pi_n)\bar{X}_n, \dot{a}(p_0)\pi_n^*) \rightarrow \mathcal{L}(B^{-1}(p_0)X, \dot{a}(p_0)\pi)$$

and (3.8) yields (3.3) by a second application of Lemma 1 (with ϕ taking the $2k$-dimensional vector $(x_1, \cdots, x_{2k})$ into the k-dimensional vector

$$(x_1 + x_{k+1}, \cdots, x_k + x_{2k})),$$

since $\operatorname{plim}_n \epsilon_n = 0$.

This concludes the proof of Theorem 1.

In the special case when π is normal with mean zero and variance v, and X is normal with zero means and unit covariance matrix, Theorem 1 yields for the limiting distribution of $n^{-\frac{1}{2}}(\bar{R}_n - na(p_0))$ a multivariate normal distribution with zero means and covariance matrix

$$\begin{aligned} E[(B^{-1}(p_0)X + \dot{a}(p_0)\pi)(B^{-1}(p_0)X + \dot{a}(p_0)\pi)'] \\ = E[B^{-1}(p_0)XX'B^{-1}(p_0)' + \dot{a}(p_0)\dot{a}(p_0)'\pi^2] \\ = B^{-1}(p_0)B^{-1}(p_0)' + (v)(\dot{a}(p_0)\dot{a}(p_0)'). \end{aligned}$$

The computation of the asymptotic distribution of runs above a normal sample mean illustrates this special case, with the following values for p_0, v, $B(p)$ and $a(p)$:

$$\begin{aligned} p_0 &= \tfrac{1}{2} \text{ (see relation (2.1))}, \\ v &= \tfrac{1}{4} - \tfrac{1}{2}\pi \text{ (see relation (2.1))}, \\ a_i(p) &= p^i q^2 \text{ for } i < k \text{ (see (5.1) of [8])}, \\ &= p^k q \text{ for } i = k \text{ (see (5.1) of [8])}, \\ B'(p)B(p) &= \|\sigma^2(p)\|^{-1}, \text{ where the matrix } \sigma^2(p) \text{ is as given in (5.2) of [8].} \end{aligned}$$

4. Extensions. The main assumption underlying Theorem 1 of Section 3 is that, for each p in V_n, $R_{p,n}$ and π_n are independent. This assumption clearly is satisfied in the application to runs above a normal sample mean given in Section 3; it is satisfied as well in many other applications, some of which are indicated below. However, the assumption will not always hold, as would be the case, for example, if π_n were, as in the application of Section 3, the proportion of the sample below the sample mean, but $R_{p,n}$ were identically equal to $R_{\pi_n,n}$ for all p.

Typical additional applications of Theorem 1 are similar to that of Section 3. Specifically, let π_n be the proportion of the sample in one of two classes determined by magnitude with respect to some sample function, and suppose that there exist constants p_0 and v such that

$$(4.1) \qquad \mathcal{L}(n^{\frac{1}{2}}(\pi_n - p_0)v^{-\frac{1}{2}}) \rightarrow N(0, 1).$$

Then the joint limit distribution of runs of various lengths, of items in the given class, will be multivariate normal, with mean vector

$$np^iq^2 \quad \text{for} \quad i < k, \qquad np^kq \quad \text{for} \quad i = k, \tag{4.2A}$$

and covariance matrix

$$(n)(\|\sigma^2(p_0)\| + (v)(\dot{a}(p_0)\dot{a}(p_0)')). \tag{4.2B}$$

The normality of the limit distribution of (4.1) is not of course essential to the convolution character, but must be included if one insists on the normality of the resultant joint distribution. Note also that the convergence of $n^{\frac{1}{2}}(\pi_n - p_0)v^{-\frac{1}{2}}$ to zero leads to the asymptotic covariance matrix $\|\sigma^2(p_0)\|$.

As an application of (4.2B), consider the asymptotic distribution of runs above the population quantile ζ_π, for random samples drawn from an arbitrary population. Here $p_0 = \pi$ and $v = \pi(1 - \pi)$. Hence the asymptotic variance of runs of length i, of items larger than ζ_π, equals

$$\sigma_{ii}(\pi) + (\pi)(1 - \pi)[(d/d\pi)\pi^i(1 - \pi)^2]^2, \tag{4.3}$$

where $\sigma_{ii}(\pi)$ is as given in the second line of Mood's equation (5.2). Expression (4.3), as it should, agrees with the first line of Mood's (8.2). Another application of (4.2B) would be to the asymptotic distribution of the number of runs above a sample mid-quartile $(z_p + z_{1-p})/2$, in random samples from a symmetric population.

A final observation is suggested by the form of (4.2B). Consider any integer solution in i of the equation $(d/dp_0)p_0^i(1 - p_0)^2 = 0$. This integer, call it I, will have the property that runs of length I will have the same marginal asymptotic distribution for all dichotomy criteria such that π_n has asymptotic mean p_0. Thus, for example, since $(d/dp)p^2(1 - p)^2 |_{\frac{1}{2}} = 0$, the marginal asymptotic distribution of runs of length 2 is the same, whether the dichotomy criterion is magnitude relative to the sample median, the population median, the sample mean (normal population), or a sample mid-quartile (symmetric population).

REFERENCES

[1] BILLINGSLEY, PATRICK (1956). The invariance principle for dependent random variables. *Trans. Amer. Math. Soc.* **83** 250–268.

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

[3] CRAMÉR, H. and WOLD, H. (1936). Some theorems on distribution functions. *J. London Math. Soc.* **11** 290–295.

[4] DAVID, F. N. (1947). A smooth test for goodness of fit. *Biometrika* **34** 299–310.

[5] DAVID, H. T. (1962). The sample mean among the extreme order statistics. *Ann. Math. Statist.* **33**.

[6] HOEFFDING, WASSILY and ROBBINS, HERBERT (1948). The central limit theorem for dependent random variables. *Duke Math. J.* **15** 773–780.

[7] v. MISES, R. (1921). Das problem der iterationen. *Z. Angew. Math. Mech.* **1** 298–307.

[8] MOOD, A. M. (1940). The distribution theory of runs. *Ann. Math. Statist.* **11** 367–392.

[9] MOSTELLER, FREDERICK (1941). Note on an application of runs to quality control charts. *Ann. Math. Statist.* **12** 228–232.

[10] Pratt, John W. On order relations and convergence in distribution. (Dittoed notes based on lectures by Herman Chernoff)

[11] Scheffé, Henry (1947). A useful convergence theorem for probability distributions *Ann. Math. Statist.* **18** 434–437.

[12] Stevens, W. L. (1939). Distribution of groups in a sequence of alternatives. *Ann. Eugenics* **9** 10–17.

[13] Swed, F. S. and Eisenhart, C. (1943). Tables for testing randomness of grouping in a sequence of alternatives. *Ann. Math. Statist.* **14** 66–87.

[14] Wald, A. and Wolfowitz, J. (1940). On a test whether two samples are from the same population. *Ann. Math. Statist.* **11** 147–162.