ON A TEST FOR HOMOGENEITY AND EXTREME VALUES

By D. A. Darling

University of Michigan

Summary. Let $x_1, x_2, \cdots, x_n$ be positive, identically distributed, independent random variables. It is of some statistical interest to study the distribution of $z_n = (x_1 + x_2 + \cdots + x_n)/\max(x_1, x_2, \cdots, x_n)$. In this paper we give its characteristic function and in a few cases its distribution. A limiting distribution of fairly wide applicability is given in the last section.

1. Introduction. Suppose $x_1, x_2, \cdots, x_n$ are positive, independent observations coming from populations with cumulative distribution functions $F(\sigma_1 x)$, $F(\sigma_2 x), \cdots, F(\sigma_n x)$ respectively, where $\sigma_i > 0$ and $F(x)$ is some known continuous cdf with $F(0) = 0$. It is of some practical importance to devise a test for the hypothesis H that $\sigma_1 = \sigma_2 = \cdots = \sigma_n$. For instance, in the analysis of variance it is desirable to know if a set of n independent sample variances are "homogeneous," or if to the contrary they come from normal populations with different variances.

Let $t_n = x_1 + x_2 + \cdots + x_n$ and $m_n = \max(x_1, x_2, \cdots, x_n)$, and put $u_n = m_n/t_n$. A possible test consists of rejecting H when u_n exceeds a certain value. In 1929 R. A. Fisher [1] used this test in determining whether the largest amplitude in a harmonic analysis was "significantly" large, H being the hypothesis that each x_i was independently distributed as $\sigma^2\chi^2$ with 2 degrees of freedom. In 1941 Cochran [2] considered an extension to k degrees of freedom. Other writers have taken the ratio of an extreme variance to an independent estimate of the variance as a test for homogeneity and outlying observations (cf. Nair [3] and Finney [4]). This test is also used implicitly in quality control work to test the equality of scale parameters; here $x_1, x_2, \cdots, x_n$ are each independently distributed as the range of a sample of size k.

A test of the nature described above is supposed to have good power against alternatives which consist of the occurrence of just one anomalously large value of σ, whereas for other alternatives (in the case of testing variances) the well known test of Bartlett (which is essentially the likelihood ratio test) is perhaps superior. This test might also be employed to determine the equality of location parameters in a set of cdf's, and generally to test the significance of extreme values. Since, under H, the distribution of u_n is independent of the common population variances (or is "Studentized" in the terminology of Hartley [5]), it is superior, for many purposes, to tests based on other order statistics (cf. Pearson and Chandrasekhar [6]).

The distribution of u_n when H is true is not, in general, known. In this note we consider the distribution of $z_n = 1/u_n = t_n/m_n$. Of course, the test of H using z_n is identical to the test using u_n, and if the critical rejection region using u_n is $u_n > \alpha$, then the critical region using z_n is $z_n < 1/\alpha$. We give the characteristic function of z_n and show how, in some cases, it can be used to determine the rejection region.

2. The distribution of Z_n. Let x_1, x_2, $\cdots$, x_n, t_n and m_n be the variables described in the preceding section, and let $z_n = t_n/m_n$. Suppose the variables x_i have the common density $\phi(x)$, $x > 0$. Then for the characteristic function of z_n we have the following formula, proven in [7]:

$$\xi_n(t) = E(e^{itz_n}) = ne^{it} \int_0^\infty \left(\beta \int_0^1 e^{it\alpha} \phi(\alpha\beta)\, d\alpha\right)^{n-1} \phi(\beta)\, d\beta. \tag{2.1}$$

In certain cases it is possible to invert this Fourier transform, and so obtain the distribution of z_n, and in any case we can find all moments of z_n in terms of quadratures. For example, the mean is

$$\mu = E(z_n) = \frac{1}{i} \xi_n'(0) = 1 + n(n-1) \int_0^\infty (F(\beta))^{n-2} \left\{\int_0^1 \phi(\beta u)\, du - \phi(\beta)\right\} \phi(\beta)\, d\beta,$$

where $F(\beta)$ is the cdf, that is, $F(\beta) = \int_0^\beta \phi(x)\, dx$.

We consider a few examples.

(a) Suppose the variables x_i are uniformly distributed over (0, 1)—or for that matter uniformly distributed over any finite interval $(0, a)$, since the distribution of z_n is independent of a. Then (2.1) becomes

$$\xi_n(t) = e^{it} \left(\frac{e^{it} - 1}{it}\right)^{n-1}$$

and we have the rather surprising result the distribution of z_n is the same as the distribution of $1 + y_1 + y_2 + \cdots + y_{n-1}$, where y_1, y_2, $\cdots$ y_{n-1} are independent and uniformly distributed over (0, 1). This fact may also be readily deduced from certain properties of order statistics of uniformly distributed variables which have been studied recently by Malmquist [8].

(b) Let the variables have the density

$$\phi(x) = \frac{1}{\Gamma\left(\frac{k}{2}\right)} e^{-x} x^{\frac{1}{2}k-1}, \tag{2.2}$$

which is the density of $2\chi^2$ with k degrees of freedom. For variables which have the density $\sigma^2\chi^2$ the characteristic function of z_n is given by (2.1) with $\phi(x)$ as in (2.2). If k is an even integer we can find the distribution function of z_n explicitly. Let $k = 2r + 2$ where r is an integer $\geqq 0$. Then (2.2) becomes

$$\phi(x) = \frac{1}{r!} e^{-x} x^r,$$

and if we work with the Laplace transform $\xi_n(it)$ rather than $\xi_n(t)$, (2.1) becomes

$$\xi_n(it) = E(e^{-tz_n}) = ne^{-t} \int_0^\infty \psi^{n-1}(\beta) \frac{1}{r!} e^{-\beta} \beta^r \, d\beta, \tag{2.3}$$

where

$$\begin{aligned} \psi(\beta) &= \frac{\beta}{r!} \int_0^1 (\alpha\beta)^r e^{-(\beta+t)\alpha} \, d\alpha, \\ &= \left(\frac{\beta}{\beta + t}\right)^{r+1} (1 - e^{-(\beta+t)} P_r(\beta + t)), \end{aligned}$$

in which $P_r(x)$ is the polynomial $P_r(x) = \sum_{j=0}^r x^j/j!$. Now $\xi_n(it)/t$ is the Laplace transform of the cumulative distribution function of z_n, and (2.3) can be rewritten

$$\frac{\xi_n(it)}{t} = \frac{n}{r!} \int_1^\infty \left(1 - \frac{1}{u}\right)^{(n-1)(r+1)} (u - 1)^r (1 - e^{-ut} P_r(ut))^{n-1} t^r e^{-ut} \, du.$$

Letting

$$f(u) = \begin{cases} \left(1 - \dfrac{1}{u}\right)^{(n-1)(r+1)} (u - 1)^r, & u > 1, \\ 0, & u \leqq 1, \end{cases}$$

and $f^{(k)}(u) = 0$ for all $k \geqq 0$ and $u \leqq 1$ we have

$$\frac{\xi_n(it)}{t} = \frac{1}{r!} \sum_{j=1}^n \binom{n}{j} (-1)^{j+1} \int_j^\infty f\left(\frac{u}{j}\right) \left[P_r\left(\frac{ut}{j}\right)\right]^{j-1} e^{-ut} t^r \, du,$$

and putting $(P_r(x))^i = \sum_{\nu=0}^{ir} a_\nu^{(i)} x^\nu$,

$$\frac{\xi_n(it)}{t} = \frac{1}{r!} \sum_{j=1}^n \binom{n}{j} (-1)^{j-1} \int_0^\infty f\left(\frac{u}{j}\right) \left\{ \sum_{\nu=0}^{(j-1)r} a_\nu^{(j-1)} \left(\frac{u}{j}\right)^\nu t^{r+\nu} \right\} e^{-ut} \, du,$$

and this expression can be now directly inverted termwise to give

$$Pr\,\{z_n \leqq x\} = \frac{1}{r!} \sum_{j=1}^n \binom{n}{j} (-1)^{j-1} \left\{ \sum_{\nu=0}^{r(j-1)} a_\nu^{(j-1)} \frac{d^{r+\nu}}{dx^{r+\nu}} f\left(\frac{x}{j}\right) \left(\frac{x}{j}\right)^\nu \right\},$$

or finally

$$\begin{aligned} Pr\,\{z_n \leqq x\} &= \frac{1}{r!} \sum_{1 \leqq j < x} \binom{n}{j} (-1)^{j-1} \sum_{\nu=0}^{r(j-1)} a_\nu^{(j-1)} \\ &\quad \cdot \frac{d^{r+\nu}}{dx^{r+\nu}} \left(1 - \frac{j}{x}\right)^{(n-1)(r+1)} \left(\frac{x}{j} - 1\right)^r \left(\frac{x}{j}\right)^\nu, \qquad 1 \leqq x \leqq n. \end{aligned} \tag{2.4}$$

If we put $r = 0$, and calculate the distribution of $u_n = 1/z_n$, (2.4) becomes

$$Pr\,\{u_n > x\} = \sum_{1 \leqq j < (1/x)} \binom{n}{j} (-1)^{j-1} (1 - jx)^{n-1}, \quad \frac{1}{n} \leqq x \leqq 1,$$

which is the result of Fisher [1]. Cochran [2] showed how to express the terms of u_n, corresponding to (2.4), for an even number of degrees of freedom as multiple integrals of Beta functions, and on the basis of certain approximations gave a table of approximate percentage points for u_n. This table, as well as a subsequent table by Eisenhart, Hastay and Wallis [9], appears to be subject to certain indeterminate errors (cf. [9] ch. 15).

3. A limiting distribution. In example (b) above it is possible to find a limiting distribution for z_n when $n \to \infty$ which may be of use when n is large. (The limiting distribution for $u_n = 1/z_n$ given by Cochran [2] appears to be erroneous owing to an oversight concerning the dependence of a set of random variables.) We remove the restriction that r be a nonnegative integer and merely require $r > -1$. Then

$$\phi(\beta) = \frac{1}{\Gamma(r+1)} e^{-\beta}\beta^r$$

and $\xi_n(t)$ is given by (2.1). Now define $f(\beta) = 1 - F(\beta) = \int_\beta^\infty \phi(x)\,dx$ and

$$\begin{aligned}\psi(\beta) &= 1 - f(\beta) + \beta\int_0^1 (e^{it\alpha} - 1)\phi(\alpha\beta)\,d\alpha\\ &= 1 - f(\beta) + g(\beta),\end{aligned}$$

so that $\xi_n(t) = -ne^{it}\int_0^\infty \psi^{n-1}(\beta)\,df(\beta)$. It is necessary to study the behavior of $g(\beta)$ for large β, and it is simple to show that

$$\begin{aligned}g(\beta) &= \beta\int_0^1 (e^{it\alpha} - 1)\phi(\alpha\beta)\,d\alpha\\ &= \frac{it(r+1)}{\beta} + \frac{(it)^2}{\beta^2}\frac{(r+1)(r+2)}{2} + o\left(\frac{1}{\beta^2}\right), \qquad \beta \to \infty\end{aligned}$$

for bounded $|t|$.

We next need to get an asymptotic solution to the equation $nf(\beta) = v$ for β when $n \to \infty$; that is, we need to solve

$$v = \frac{n}{\Gamma(r+1)}\int_\beta^\infty e^{-x}x^r\,dx$$

for β when $n \to \infty$. An asymptotic development similar to that above shows that we have

$$v = \frac{n}{\Gamma(r+1)} e^{-\beta}\beta^r\left(1 + \frac{r}{\beta} + o\left(\frac{1}{\beta}\right)\right)$$

as an equation to solve for β. After some calculation we find

$$\beta = \log n + r\log\log n + r^2\frac{\log\log n}{\log n} - \log(v\Gamma(r+1)) + O\left(\frac{1}{\log n}\right),$$

and hence

$$\frac{1}{\beta} = \frac{1}{\log n} - r\,\frac{\log\log n}{\log^2 n} + \frac{\log\,(v\Gamma(r+1))}{\log^2 n} + o\left(\frac{1}{\log^2 n}\right),$$

so that

$$\psi(\beta) = 1 - f(\beta) + g(\beta) \;= 1 - \frac{v}{n} + it(r+1)$$

$$\left\{\frac{1}{\log n} - \frac{r\log\log n}{\log^2 n} + \frac{\log\,(v\Gamma(r+1))}{\log^2 n} + o\left(\frac{1}{\log^2 n}\right)\right\}.$$

Now define numbers c_n and d_n as follows:

$$c_n = (r+1)\,\frac{n}{\log^2 n},$$

$$d_n = \log n - r\log\log n + \log\Gamma(r+1),$$

and it is simple to verify that

$$\xi_n(t/c_n)e^{-itd_n} = e^{it/c_n}\int_0^n\left(1 - \frac{v}{n} + \frac{it\log v}{n} + o\left(\frac{1}{n}\right)\right)^{n-1} dv,$$

and that when $n \to \infty$ the limit may be taken under the integration sign to give

$$\xi_n(t/c_n)e^{-itd_n} \to \int_0^\infty e^{-v+it\log v}\,dv = \Gamma(it+1).$$

But the expression on the right is simply the characteristic function for the random variable $\log w$, where w has the exponential density e^{-x}, and hence

$$\lim_{n\to\infty} Pr\{z_n/c_n - d_n < x\} = Pr\{\log w < x\} = 1 - e^{-e^x}, \qquad -\infty < x < \infty,$$

by the continuity theorem for characteristic functions. It does not follow, of course, that the constants c_n and d_n are the "best" in the sense that they give the "closest" approximation to the limiting distribution function when n is finite.

Finally, then

$$\lim_{n\to\infty} Pr\left\{z_n < (r+1)\,\frac{n}{\log n} - r(r+1)\,\frac{n\log\log n}{\log^2 n} + (r+1)\log\Gamma(r+1)\right.$$

$$\left.\cdot\,\frac{n}{\log^2 n} + (r+1)\,\frac{n}{\log^2 n}\,x\right\} = 1 - e^{-e^x}, \qquad -\infty < x < \infty.$$

4. A general limiting expression. Following the analysis of Section 3, it is possible to get a general limiting distribution for z_n for quite a broad class of distributions $\phi(x)$. There are two essentially distinct cases.

(a) If x_i is bounded (i.e., if $\phi(x) = 0$ for $x > x_0$) then $m_n = \max\,(x_1, x_2, \cdots, x_n)$

tends with probability 1 to some sure number M and hence z_n has the same asymptotic distribution as $1/M\ (x_1 + x_2 + \cdots + x_n)$—that is, the Gaussian distribution.

(b) If $\mu = E(x_i)$ exists and x_i is unbounded from above we define, as before $\psi(\beta) = 1 - f(\beta) + g(\beta)$, where

$$g(\beta) = \beta \int_0^1 (e^{it\alpha} - 1)\phi(\alpha\beta)\, d\alpha,$$

so that (2.1) becomes $\xi_n(t) = -ne^{it} \int_0^\infty (\psi(\beta))^{n-1}\, df(\beta)$, $f(\beta)$ being $1 - F(\beta) = \int_\beta^\infty \phi(x)\, dx$ as in Section 3. An asymptotic development for $g(\beta)$ is

$$g(\beta) = \frac{it\mu}{\beta} + o\left(\frac{1}{\beta}\right), \qquad \beta \to \infty,$$

and as before we must solve $nf(\beta) = v$ for $1/\beta$, asymptotically for $n \to \infty$.

Let us suppose that there are constants a_n and b_n such that if $nf(\beta) = v$, v bounded,

(1) $\dfrac{1}{\beta} = a_n + h(v)b_n + f_n(v), \qquad n \to \infty,$

(2) $nb_n \to \infty, \qquad \dfrac{a_n}{nb_n} \to 0, \qquad n \to \infty,$

(3) $f_n(v)/b_n \to 0$ uniformly for v in any closed interval $0 < \epsilon \leqq v \leqq M < \infty$.

Then $h(v)$ is a monotone increasing function, and we consider the normalized variable $(z_n - \mu na_n)/\mu nb_n$. For its characteristic function we have

$$\xi_n(t/nb_n)e^{-it(a_n/b_n)} = \int_0^n \left(1 - \frac{v}{n} + \frac{it\,h(v)}{n} + o\left(\frac{1}{n}\right)\right)^{n-1} dv \to \int_0^\infty e^{-v+ith(v)}\, dv.$$

As a consequence the normalized variable has the limiting distribution of $h(w)$, where w is distributed with an exponential density e^{-x}. Hence

$$Pr\left\{\frac{z_n}{nb_n} - \frac{a_n}{b_n} < x\right\} \to Pr\{h(w) < x\} = 1 - e^{-h^{-1}(x)},$$

or

$$Pr\ \{z_n < \mu na_n + \mu nb_n x\} \to 1 - e^{-h^{-1}(x)}.$$

REFERENCES

[1] R. A. Fisher, "Tests of significance in harmonic analysis," *Proc. Roy. Soc. Edinburgh Sect. A*, Vol. 125 (1929), pp. 54–59.

[2] W. G. Cochran, "The distribution of the largest of a set of estimated variances as a fraction of their total," *Annals of Eugenics*, Vol. 11 (1941), pp. 47–52.

[3] K. R. Nair, "The studentized form of the extreme mean square test in the analysis of variance," *Biometrika*, Vol. 35 (1948), pp. 16–31.

[4] D. J. Finney, "The joint distribution of variance ratios based on common error mean square," *Annals of Eugenics*, Vol. 11 (1941), pp. 136–140.

[5] H. O. Hartley, "Studentization and large sample theory," *Jour. Roy. Stat. Soc. Suppl.*, Vol. 5 (1938), pp. 80–90.

[6] E. S. Pearson and C. Chandrasekar, "The efficiency of statistical tools and a criterion for rejection of outlying observations," *Biometrica*, Vol. 28 (1936), pp. 308–320.

[7] D. A. Darling, "The influence of the maximum term in the addition of independent random variables," *Trans. Am. Math. Soc.* (to be published).

[8] S. Malmquist, "On a property of order statistics from a rectangular distribution," *Skandinavisk Aktuarietidskrift*, Vol. 33 (1950), pp. 214–222.

[9] C. Eisenhart, M. W. Hastay and W. A. Wallis, "*Selected Techniques* of *Statistical Analysis*," McGraw-Hill Book Co., 1947.