RECURRENCE OF EXTREME OBSERVATIONS*

S. S. WILKS

(rec. 18 June 1958)

1. Introductory Remarks

Suppose a preliminary set of m independent observations are drawn from a population in which a random variable x has a continuous but unknown cumulative distribution function $F(x)$. Let y be the largest observation in this preliminary sample. Now suppose further observations are drawn one at a time from this population until an observation exceeding y is obtained. Let n be the number of further drawings required to achieve this objective. The problem is to determine the distribution function of the random variable n. More generally, suppose y is the r-th from the largest observation in the preliminary sample and let n denote the number of further trials required in order to obtain k observations which exceed y. What is the distribution function of n?

The distribution function of n and some of its properties are given in this paper. Furthermore, the asymptotic distribution of n/m for large values of m will be found to be of an extremely simple form. Certain further extensions will also be noted. The results presented are distribution-free in the sense that they do not depend on the functional form of $F(x)$.

2. The Simplest Recurrence Case

First, let us consider the simplest case. We draw a preliminary sample of m observations from a population having a continuous cumulative distribution function $F(x)$. Denote the largest observation in this preliminary sample by y, and let n denote the number of further observations required to obtain one which exceeds y. We shall show that the probability distribution on n is given by

$$(1) \qquad p(n) = \frac{m}{(m+n)(m+n-1)}, \quad n = 1, 2, 3, \cdots.$$

To establish (1) we observe that the random variable $F(y)$ which we may

* This problem arose in some research partially supported by the Office of Naval Research.

denote by F, has the probability element,

$$mF^{m-1}dF, \quad 0 \leqq F \leqq 1. \tag{2}$$

assuming, of course, that the m observations are independent.

For a given value of y, and hence of $F(y)$, the probability of having to make n additional trials in order to obtain an observation which exceeds y is

$$F^{n-1}(1-F), \quad n = 1, 2, 3, \cdots. \tag{3}$$

The joint distribution of F and n is therefore the product of expressions (1) and (2), namely

$$mF^{m+n-2}(1-F)dF. \tag{4}$$

(Note that F has a continuous distribution on the interval (0, 1) and n has a discrete distribution on the integers 1, 2, 3, $\cdots$). To obtain the probability distribution function of n, we simply take the marginal distribution of (4) with respect to n, i.e. we integrate (4) with respect to F over (0, 1). This yields $p(n)$ as given by (1).

It should be noted that the distribution of n is extremely spread out on the positive integers: Both its mean and variance are infinite.

The cumulative distribution function of n, say $G(n)$, as defined by $\sum_{i=1}^{n} p(i)$, is readily seen to be as follows

$$G(n) = \frac{n}{m+n} \tag{5}$$

Taking the ratio n/m, we see that

$$P\left(\frac{n}{m} \leqq z\right) = \frac{z}{1+z}, \quad z = \frac{1}{m}, \quad \frac{2}{m}, \cdots \tag{6}$$

and, of course,

$$\lim_{m\to\infty} P\left(\frac{n}{m} \leqq z\right) = \frac{z}{1+z}, \quad z > 0. \tag{7}$$

The density function of this limiting cumulative distribution is

$$f(z) = \frac{1}{(1+z)^2}, \quad z > 0. \tag{8}$$

The value of n, say n_β, for which $G(n) = \beta$ is given by

$$n_\beta = \frac{\beta}{1-\beta} m. \tag{9}$$

For instance, if $\beta = 0.95$, we have

$$n_{.95} = 19m$$

which means that if we take the largest observation in a preliminary sample of m observations we would have to be prepared to make up to $19m$ additional observations from the same population in order to have a probability of 0.95 of obtaining an x which exceeded the largest one in the preliminary sample. Similarly, by choosing $\beta = 0.05$ we find $n_{.05} = m/19$ which means that one cannot take more than $m/19$ further observations without having probability < 0.95 of having all x's less than y.

It should be noted that if y is the smallest x in the preliminary sample of size m and n is the number of subsequent trials required to find an x less than y, then the probability function of n is also given by (1).

3. Recurrence of r-th Largest Observation in Sample

In this case let y be the r-th largest in the preliminary sample of m observations and let n be the number of additional observations required to obtain an observation which exceeds y. The probability function of n is given by

$$(10) \qquad p(n) = \frac{\binom{m-1}{r-1}}{\binom{m+n-1}{r}} \left(\frac{m}{m+n}\right), \quad n = 1, 2, 3, \cdots.$$

The argument for (10) is similar to that for (1). For the probability element of $F(y)$ is

$$(11) \qquad \frac{m!}{(r-1)!\,(m-r)!} F^{m-r}(1-F)^{r-1}dF,$$

and the probability of having to make n further observations to obtain one which exceeds y is given by (3). The joint distribution of F and n, is the product of the expressions in (10) and (3), that is

$$(12) \qquad \frac{m!}{(r-1)!(m-r)!} F^{m+n-r-1}(1-F)^{r}dF.$$

To find the probability function $p(n)$ we merely integrate (12) with respect to F from 0 to 1, remembering that for positive integers p and q

$$\int_0^1 x^p(1-x)^q dx = \frac{p!q!}{(p+q+1)!}.$$

This gives

$$(13) \qquad p(n) = \frac{m!r!(m+n-r-1)!}{(r-1)!(m-r)!(m+n)!}, \quad n = 1, 2, 3, \cdots$$

which reduces to (10).

The mean of the distribution (10) is found by multiplying expression (12) by n, summing with respect to n from 0 to ∞, and then integrating with respect to F from 0 to 1. This gives

$$\mathscr{E}(n) = \frac{m}{r-1}, \tag{14}$$

which, of course, is finite only if $r = 2, 3, \cdots, m$.

The variance of the distribution $\sigma^2(n)$ can be similarly found by evaluating $\mathscr{E}[n(n-1)]$ and using the fact that $\sigma^2(n) = \mathscr{E}[n(n-1)] + \mathscr{E}(n) - [\mathscr{E}(n)]^2$. This yields

$$\sigma^2(n) = \frac{mr(m-r+1)}{(r-1)^2(r-2)}, \tag{15}$$

which is finite only if $r = 3, 4, \cdots, m$.

The cumulative distribution function of n, say $G(n)$, defined by $\sum_{i=1}^{n} p(i)$, is found by summing the expression (12) for $n = 1, 2, \cdots, n$, and integrating with respect to F from 0 to 1. This gives

$$G(n) = 1 - \frac{m(m-1)\cdots(m-r+1)}{(m+n)(m+n-1)\cdots(m+n-r+1)} \tag{16}$$

Considering the ratio n/m, we see that

$$P\left(\frac{n}{m} \leqq z\right) = 1 - \frac{m(m-1)\cdots(m-r+1)}{(m+mz)(m+mz-1)\cdots(m+mz-r+1)} \tag{17}$$

from which we obtain

$$\lim_{m\to\infty} P\left(\frac{n}{m} \leqq z\right) = 1 - \frac{1}{(1+z)^r}. \tag{18}$$

Hence, for large m we have

$$P\left(\frac{n}{m} \leqq z\right) \simeq 1 - \frac{1}{(1+z)^r}, \tag{19}$$

the probability density function of this limiting distribution being

$$f(z) = \frac{r}{(1+z)^{r+1}}, \quad z > 0. \tag{20}$$

From (20) we find

$$\begin{aligned} \mathscr{E}(z) &= \frac{1}{r-1}, & r > 1 \\ \sigma^2(z) &= \frac{r}{(r-1)^2(r-2)}, & r > 2. \end{aligned} \tag{21}$$

Suppose y_1 and y_2 are the smallest and largest x in the preliminary sample, and let n be the number of subsequent trials required to obtain an x outside the interval $[y_1, y_2]$. It can be shown by argument similar to that given above that the probability function of n is given by (10) with $r = 2$, i.e.

$$(22) \quad p(n) = \frac{2m(m-1)}{(m+n)(m+n-1)(m+n-2)}, \quad n = 1, 2, 3, \cdots.$$

The mean of this distribution as we see from (14) for $r = 2$, is

$$(23) \qquad \mathscr{E}(n) = m$$

while the variance is infinite.

The cumulative distribution of n in this case is given by (16) with $r = 2$, i.e.

$$(24) \qquad G(n) = 1 - \frac{m(m-1)}{(m+n)(m+n-1)}.$$

The value of n, say n_β, for which

$$G(n) = \beta$$

is given by solving

$$1 - \frac{m(m-1)}{(m+n)(m+n-1)} = \beta$$

which gives

$$n_\beta \cong (m - \tfrac{1}{2})\left(\frac{1}{\sqrt{1-\beta}} - 1\right) + O\left(\frac{1}{m}\right).$$

For instance, if $\beta = 0.95$ we have

$$n_{.95} \cong (m - \tfrac{1}{2})(\sqrt{20} - 1) = 3.47\ (m - \tfrac{1}{2}).$$

Thus, if we take the interval formed by the smallest and largest x in a preliminary sample of m observations, we must be prepared to make up to approximately $3.47m$ further observations in order to obtain an x outside this interval with probability 0.95.

If $\beta = 0.05$ we have $n_{.05} \cong m/38$ which means that one cannot take more than $m/38$ further observations without lowering the probability below 0.95 of having all observations fall in $[y_1, y_2]$.

4. The General Case

As before, suppose y is the r-th largest x in the preliminary sample and let n be the number of subsequent observations required to obtain k observations which exceed y. It can be shown by straightforward extension of the argument in the preceding section that the probability function of n is given by

$$(25) \quad p(n) = \frac{\binom{n-1}{k-1}\binom{m-1}{r-1}}{\binom{m+n-1}{k+r-1}}\left(\frac{m}{m+n}\right), \quad n = k,\ k+1,\ k+2, \cdots.$$

For the mean of this distribution we have

$$(26) \qquad \mathscr{E}(n) = \frac{mk}{r-1}, \quad r > 1.$$

Writing $p(n)$ in the form

$$(27) \quad \frac{m!}{(k-1)!(r-1)!(m-r)!}\left[\frac{d^{k-1}t^{n-1}}{dt^{k-1}}\right]_{t=1}\left\{\frac{\Gamma(m+n-k-r+1)\Gamma(k+r)}{\Gamma(m+n+1)}\right\},$$

and noting that the expression in { } can be written as

$$\int_0^1 u^{m+n-k-r}(1-u)^{k+r-1}du,$$

we can write the cumulative distribution function of n as

$$\begin{aligned}
G(n) &= 1 - \sum_{i=n+1}^{\infty} p(i) \\
(28) \qquad &= 1 - \left[\frac{m!}{(k-1)!(r-1)!(m-r)!}\right] \\
&\qquad \cdot \int_0^1 \frac{d^{k-1}}{dt^{k-1}}\,[t^n(1-tu)^{-1}]_{t=1}\, u^{m+n-k-r-1}(1-u)^{k+r-1}du \\
&= 1 - \sum_{j=0}^{k-1}\binom{k+r-j-2}{r-1}\Phi(j,\ k,\ r,\ m,\ n)
\end{aligned}$$

where

$$(29)\ \Phi(j,\ \kappa,\ r,\ m,\ n) = \frac{(m)(m-1)\cdots(m-r+1)(n)(n-1)\cdots(n-k+j+2)}{(m+n-2)(m+n-3)\cdots(m+n-k-r+j)}.$$

If we put $n = mz$ we find the following limiting cumulative distribution function of n/m to be

$$(30) \quad \lim_{m\to\infty} P\left(\frac{n}{m} \leq z\right) = 1 - \frac{z^{k-1}}{(1+z)^{k+r-1}}\sum_{j=0}^{k-1}\binom{k+r-i-2}{r-1}\left(\frac{1+z}{z}\right)^{i}.$$

For $k = 1$, we obtain, of course, (18) as a special case of (30). The probability density function of the limiting distribution given by (30) is

$$(31) \quad f(z) = \frac{z^{k-2}}{(1+z)^{k+r}}\sum_{j=0}^{k-1}\binom{k+r-j-2}{r-1}(rz-k+j+1)\left(\frac{1+z}{z}\right)^{j}$$

for

$$z > 0.$$

The mean and variance of z are found to be

$$\mathscr{E}(z) = \frac{k}{r-1}, \quad \sigma^2(z) = \frac{k(k+r-1)}{(r-1)^2(r-2)} \tag{32}$$

and are finite for $r > 1$ and $r > 2$, respectively.

If we take any interval of form

$$(x_{(s)}, \; x_{(m-r+s)}) \tag{33}$$

$s = 0, \; 1, \cdots, r+1$, where $x_{(0)} = -\infty$, $x_{(m+1)} = +\infty$ and where $x_{(1)} < x_{(2)} < \cdots < x_{(m)}$ are the order statistics of the preliminary sample of size m, and if we draw subsequent observations from the population until we obtain k observations falling outside the interval (33), it can be shown by essentially the same argument as that already used that the cumulative distribution function of n, the number of subsequent observations required to accomplish this objective, is given by (28). The limiting cumulative distribution function of n/m as $m \to \infty$, is, of course, given by (30), while the limiting density function is given by (31).

Princeton University