THE DISTRIBUTION OF THE NUMBER OF EXCEEDANCES[1]

By E. J. Gumbel and H. von Schelling

New York and Naval Medical Research Laboratory, New London, Connecticut

0. The problem. We study the probability that the mth observation in a sample of size n taken from an unknown distribution of a continuous variate will be exceeded x times in N future trials, and calculate the averages, the moments, and the cumulative probability function of the number of exceedances. This problem leads to the hypergeometric series. Our starting point is a special case of a distribution studied by Wilks [3] who considered several order statistics whereas we consider only one. His tolerance limits are special cases of our cumulative probability function. Thus the present paper is, at the same time, a specialization and a generalization of the work done by Wilks.

1. Distribution. From a continuous variate ξ an alternative is constructed by choosing the mth among n observations $\xi_m (m = 1, 2, \cdots, n)$. The rank m is counted from the top, which means that $m = 1$ ($m = n$) stands for the largest (smallest) observation. The observation ξ_m is thus the mth largest value. We ask: In how many cases x will the past mth observation be equalled or exceeded in N future trials taken from the same population? For the sake of simplicity, x is called the number of exceedances.

If the initial probability $F(\xi_m) = F_m$ for a value less than ξ_m is known, the alternative probability for exceeding ξ_m is $1 - F_m$, and Bernoulli's theorem gives the probability

$$w_1(F_m, N, x) = \binom{N}{x} (1 - F_m)^x F_m^{N-x} \tag{1.1}$$

that x among N future trials will exceed ξ_m. However, as a rule the probability F_m is unknown. The only data known are the n past observations. To eliminate the probability F_m, we introduce the distribution $v(F_m)$ of the frequency F_m of the mth largest among n values

$$v(n, m, F_m)\, dF_m = \binom{n}{m} m F_m^{n-m} (1 - F_m)^{m-1}\, dF_m , \tag{1.2}$$

consider F_m as a variate, and integrate (1.1) over all values of this variate. Thus F_m is replaced by a function of n and m.

The convolution of (1.1) and (1.2) leads to *the distribution* $w(n, m, N, x)$ *of*

[1] Opinions or conclusions contained in this paper are those of the authors. They are not to be construed as necessarily reflecting the views or endorsement of the Navy Department.

the number of exceedances over the *mth* largest among n observations in N future trials

$$w(n, m, N, x) = \frac{\binom{n}{m} m \binom{N}{x}}{(N+n)\binom{N+n-1}{m+x-1}}. \tag{1.3}$$

This probability depends upon the parameters n, m, and N, but not upon the unknown probability F_m. Therefore it is distribution-free. If we are interested in the dependence of $w(n, m, N, x)$ on x only we simply write $w(x)$. The conditions for the positive integers m and x, and for the probability $w(x)$ are

$$1 \leqq m \leqq n; \qquad 0 \leqq x \leqq N; \qquad \sum_0^N w(x) = 1. \tag{1.3'}$$

The distribution (1.3) possesses the following *symmetry*

$$w(n, m, N, x) = w(n, n - m + 1, N, N - x) \tag{1.4}$$

which reads: *The probability that the past mth value from above will be exceeded x times in N new trials is equal to the probability that the past mth value from below will be exceeded $N - x$ times.*

The nN probabilities $w(n, m, N, x)$ are linked by several recurrence formulas which follow easily from the usual combinatorial rules. For fixed m, the probability for $x + 1$ is obtained from the probability for x by

$$\begin{aligned} w(n, m, N, x+1) &= w(n, m, N, x) \frac{(N-x)(m+x)}{(N+n-m-x)(x+1)} \\ &= w(n, n-m+1, N, N-x). \end{aligned} \tag{1.5}$$

In the same way, the probabilities $w(n, m, N, x + 1)$, $w(n, m + 1, N, x)$ and $w(n, x, N, m)$ are easily obtained from the probabilities $w(n, m, N, x)$. The distribution (1.3) has many aspects since, besides the number of exceedances x, also the rank m and the number of future trials N may be considered as variates.

For $m = 1$ and $m = n$, the distribution of the number of exceedances over the largest value diminishes with x, and the distribution of the number of exceedances over the smallest value increases with x. For $x = 0$, and $m = 1$, we obtain from (1.3)

$$w(n, 1, N, 0) = \frac{n}{N+n} = w(n, n, N, N). \tag{1.6}$$

For $x = 0$, $m = n$, the probability that the smallest observation will never be exceeded, equal to the probability that the largest value will always be exceeded, is very small, even for moderate sample sizes.

If n is odd, then $m = (n + 1)/2$ corresponds to the median of the initial variable ξ, and the symmetry relation (1.4) becomes

$$w(n, (n+1)/2, N, x) = w(n, (n+1)/2, N, N - x). \tag{1.7}$$

It is equally probable that the median of the n past observations is surpassed x or $N - x$ times in N future trials.

2. The two asymptotic distributions. If both n and N are large, m may increase with n such that the quotient m/n remains constant, and the mth values remain near the median. Or, m remains constant such that $m << n$, and the mth values are extremes.

In the first case, let $n = N = 2k - 1$, where k is large. Then $m = k$ is the rank of the median of the initial distribution. As shown in (1.7), the distribution of the number of exceedances over the initial median is symmetrical. To obtain the asymptotic distribution we reduce x by writing

$$x = k + z\sqrt{k} \tag{2.1}$$

where z remains in a finite interval. The same reduction may be applied to mth values in the neighborhood of the initial median. The distribution of the number of exceedances over the initial median is, from (1.3) and (2.1),

$$w(2k-1, k, 2k-1, x) = \text{const}\, \frac{\binom{2k-1}{k+z\sqrt{k}}}{\binom{4k-3}{2k+z\sqrt{k}-1}}.$$

Consider only the factors involving the variate z, then the right side becomes, by Stirling's formula,

$$\frac{(2k+z\sqrt{k}-1)!(2k-z\sqrt{k}-2)!}{(k+z\sqrt{k})!(k-z\sqrt{k}-1)!} \sim \frac{(2k+z\sqrt{k})^{2k+z\sqrt{k}}(2k-z\sqrt{k})^{2k-z\sqrt{k}}e^{-z\sqrt{k}+z\sqrt{k}}}{(k+z\sqrt{k})^{k+z\sqrt{k}}(k-z\sqrt{k})^{k-z\sqrt{k}}e^{-z\sqrt{k}+z\sqrt{k}}}.$$

Combination of the factors with the same powers leads to

$$\frac{(4k^2-kz^2)^{2k}}{(k^2-kz^2)^k}\left(\frac{(2k+z\sqrt{k})(k-z\sqrt{k})}{(2k-z\sqrt{k})(k+z\sqrt{k})}\right)^{z\sqrt{k}} \sim \frac{\left(1-\frac{z^2}{4k}\right)^{2k}}{\left(1-\frac{z^2}{k}\right)^k}\left(\frac{\left(1+\frac{z}{2\sqrt{k}}\right)\left(1-\frac{z}{\sqrt{k}}\right)}{\left(1-\frac{z}{2\sqrt{k}}\right)\left(1+\frac{z}{\sqrt{k}}\right)}\right)^{z\sqrt{k}}.$$

Since k and $\sqrt{k}$ are large, and z is small, all factors lead to exponential functions whence

$$\exp\left[-\frac{z^2}{2}+z^2+\frac{z^2}{2}+\frac{z^2}{2}-z^2-z^2\right] = \exp\left[-\frac{z^2}{2}\right]$$

and finally,

$$\lim_{k=\infty} w(2k-1, k, 2k-1, x) = \text{const}\, e^{-z^2/2}. \tag{2.2}$$

The number of exceedances over the initial median, $m = k$, in a large sample of size $2k - 1$ in $2k - 1$ future trials is normally distributed with mean, median, mode, and variance equal to k. Therefore the probabilities (2.2) may be called *the distribution of normal exceedances.*

In the second case where N and n are large, and m and x are small, a distribution analogous to the Poisson distribution will be obtained. To indicate that N and n are large, they are written $\underline{N}$ and $\underline{n}$. The probability

$$w(\underline{n}, m, \underline{N}, x) = \frac{(x + m - 1)!\,\underline{n}!\,\underline{N}!\,(\underline{N} + \underline{n} - x - m)!}{(m - 1)!\,x!\,(\underline{n} - m)!\,(\underline{N} - x)!\,(\underline{N} + \underline{n})!}$$

obtained from (1.3) becomes, by use of the Stirling formula,

$$w(\underline{n}, m, \underline{N}, x) = \binom{x + m - 1}{x} \frac{\underline{n}^m \underline{N}^x}{(\underline{N} + \underline{n})^{m+x}} = w(\underline{n}, \underline{n} - m + 1, \underline{N}, \underline{N} - x). \tag{2.3}$$

If $\underline{n} = \underline{N}$, the preceding formula becomes

$$w(\underline{n}, m, \underline{n}, x) = \binom{x + m - 1}{x} (\tfrac{1}{2})^{m+x} = w(\underline{n}, \underline{n} - m + 1, \underline{n}, \underline{n} - x). \tag{2.4}$$

This probability that the mth largest (or smallest) value will be exceeded x times (or $\underline{n} - x$ times) in $\underline{n}$ future trials is independent of $\underline{n}$. Since m is small compared to $\underline{n}$, the probabilities (2.4) may be called *the distribution of rare exceedances.*

For $x = 0$, we obtain the probability

$$w(\underline{n}, m, \underline{n}, 0) = (\tfrac{1}{2})^m = w(\underline{n}, \underline{n} - m + 1, \underline{n}, \underline{n})$$

that the largest (or smallest) mth extreme value is never (or always) exceeded. For $m = 1$, and $\underline{n} = \underline{N}$, the probability

$$w(\underline{n}, 1, \underline{n}, x) = (\tfrac{1}{2})^{x+1} = w(\underline{n}, \underline{n}, \underline{n}, \underline{n} - x) \tag{2.5}$$

that the largest (or smallest) value is exceeded x times (or $\underline{n} - x$ times) is a geometric series.

To obtain the moments of the distribution of rare exceedances (2.4) we construct its generating function

$$G_x(t) = (\tfrac{1}{2})^m \sum_{x=0}^{\infty} \binom{x + m - 1}{m - 1} \left(\frac{e^t}{2}\right)^x.$$

From the well known expression for the negative binomial follows

$$G_x(t) = (\tfrac{1}{2})^m \left(1 - \frac{e^t}{2}\right)^{-m}, \tag{2.6}$$

whence, by the usual procedure

$$\bar{x} = m. \tag{2.7}$$

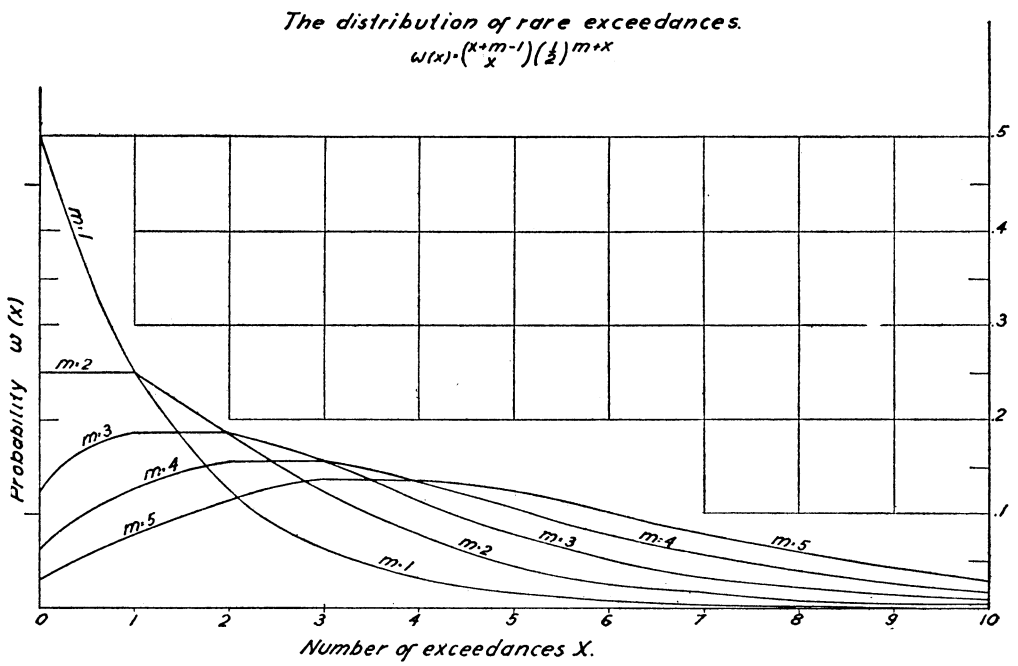
The mean number of exceedances over the mth value from above in the dis-

tribution of rare exceedances is m itself. The second derivative of (2.6) for $t = 0$ leads to the variance

$$\sigma^2 = 2m \tag{2.8}$$

which is the double of the variance in the Poisson distribution. This difference is easily explained: If we apply the Poisson law to the exceedances, we have to know the mean number of exceedances. In our case we only know one observed number of exceedances. Consequently the variance must be larger than in the Poisson case.

GRAPH 1



The variance for the distribution (2.2) of the normal exceedances was $(N + 1)/2$, whereas the variance (2.8) for the distribution of rare exceedances, $2m$, is much smaller since m is small compared to N. This interesting relation will be generalized in paragraph 3.

For m increasing, the distributions (2.4) spread as shown in graph 1. The distributions have two modes

$$\tilde{x}_1 = m - 2; \tilde{x}_2 = m - 1 \tag{2.9}$$

except for $m = 1$, where the probability diminishes with x. The distributions (2.4) are similar to the Poisson distribution for integer m. However, for this distribution the modes are $m - 1$ and m.

The similarity between the two distributions may also be seen from their behavior for large m. In this case, the Poisson distribution for the standardized variate $y = (x - m)/\sigma$ converges toward a normal distribution. The same holds for the distribution of rare exceedances. For the proof consider the standardized variate

$$y = (x - m)/\sqrt{2m}. \tag{2.10}$$

Its moment generating function $G_y(t)$ becomes, from (2.6),

$$G_y(t) = \left(2e^{t/\sqrt{2m}} - e^{2t/\sqrt{2m}}\right)^{-m}.$$

The usual development leads to the second member

$$\left(2 + \frac{2t}{\sqrt{2m}} + \frac{2t^2}{4m} - 1 - \frac{2t}{\sqrt{2m}} - \frac{4t^2}{4m} + O(m^{-3/2})\right)^{-m}$$

$$= \left(1 - \frac{t^2}{2m} + O(m^{-3/2})\right)^{-m}.$$

If we neglect the factors $O(m^{-3/2})$, we finally obtain

$$G_y(t) = e^{t^2/2} \tag{2.11}$$

which is the normal generating function. Thus the distribution of rare exceedances converges toward normalcy in the same way as the Poisson distribution.

3. Moments. We return to the general distribution (1.3). For the calculation of the moments, the hypergeometric series $F(\alpha, \beta, \gamma, 1)$ defined by

$$F(\alpha, \beta, \gamma, 1) = 1 + \frac{\alpha}{1}\frac{\beta}{\gamma} + \frac{\alpha(\alpha+1)}{1\cdot 2}\frac{\beta(\beta+1)}{\gamma(\gamma+1)} + \cdots \tag{3.1}$$

is used. The $x+$ 1st member of this series is

$$f(x) = \frac{\alpha(\alpha+1)\cdots(\alpha+x-1)}{x!}\,\frac{\beta(\beta+1)\cdots(\beta+x-1)}{\gamma(\gamma+1)\cdots(\gamma+x-1)}. \tag{3.2}$$

On the other hand, the $x+$ 1st member of the distribution $w(x)$ may be written, from (1.3), after changing the signs,

$$w(x) = \frac{\binom{n}{m}}{\binom{N+n}{m}}\,\frac{m(m+1)\cdots(m+x-1)}{x!}\cdot \tag{3.3}$$

$$\frac{(-N)(-N+1)\cdots(-N+x-1)}{(m-n-N)(m-n-N+1)\cdots(m-n-N+x-1)}.$$

This is the general member (3.2) of the hypergeometric series, if we write

$$\alpha = m;\ \beta = -N;\ \gamma = m - n - N. \tag{3.4}$$

Therefore the probability $w(n, m, N, x)$ is the $x + 1$st member in the development of

$$\frac{n!(N + n - m)!}{(N + n)!(n - m)!} F(m, -N, m - n - N, 1).$$

Since the sum of the probabilities $w(x)$ must be unity, we obtain

$$F(m, -N, m - n - N, 1) = \frac{(N + n)!}{n!} \frac{(n - m)!}{(N + n - m)!}. \tag{3.5}$$

This relation will be used for the calculation of the factorial moments $\bar{x}_{[k]}$ of order k which are, from (3.3.),

$$\bar{x}_{[k]} = \frac{n!(N + n - m)!}{(n - m)!(N + n)!} \sum_{x=k}^{N} \frac{N(N - 1) \cdots (N - x + 1)m(m + 1) \cdots (m + x - 1)}{(x - k)!(N + n - m)(N + n - m - 1) \cdots (N + n - m - x + 1)}. \tag{3.6}$$

The first member in the sum is

$$\varphi(1) = \frac{N(N - 1) \cdots (N - k + 1)m(m + 1) \cdots (m + k + 1)}{0!(N + n - m)(N + n - m - 1) \cdots (N + n - m - k + 1)}. \tag{3.7}$$

The second member is

$$\varphi(2) = \varphi(1) \frac{(N - k)(m + k)}{1!(N + n - m - k)}.$$

Generally, each successive member is obtained from the preceding one by the same rules as the successive members of the hypergeometric series (3.1). Consequently, from (3.6),

$$\bar{x}_{[k]} = \frac{n!(N + n - m)!}{(n - m)!(N + n)!} \varphi(1) \left(1 + \frac{(N - k)(m + k)}{1!(N + n - m - k)} + \cdots\right). \tag{3.8}$$

The sum in the brackets is the hypergeometric series

$$F(m + k, -(N - k), (m - n - N + k), 1).$$

If we replace, in (3.5), m by $m + k$, N by $N - k$, n by $n + k$,we obtain for the sum in (3.8)

$$F(m + k, -(N - k), m - n - N + k, 1) = \frac{(N + n)!(n - m)!}{(n + k)!(N + n - m - k)!}. \tag{3.9}$$

Introduction of (3.9) and (3.7) into (3.8) leads to the factorial moments

$$\bar{x}_{[k]} = \frac{m(m + 1) \cdots (m + k - 1)N(N - 1) \cdots (N - k + 1)}{(n + 1)(n + 2) \cdots (n + k)} \tag{3.10}$$

and to the recurrent relation

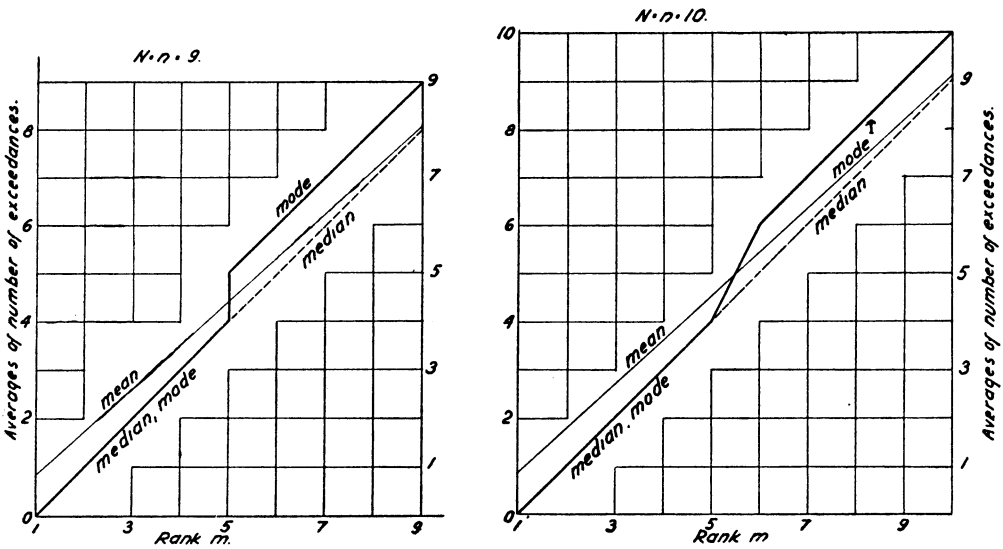
$$\bar{x}_{[k]} = \frac{(m + k - 1)(N - k + 1)}{n + k} \bar{x}_{[k-1]} . \tag{3.10'}$$

If n and N are both of the same order of magnitude, and large compared to k, the expression (3.10) simplifies to

$$\bar{x}_{[k]} = m(m + 1) \cdots (m + k - 1). \tag{3.10''}$$

GRAPH 2

Averages of numbers of exceedances.



For $k = 1$ we obtain the mean number of exceedances $\bar{x}_m$ over the mth largest value in N future trials

$$\bar{x}_m = N \frac{m}{n + 1} . \tag{3.11}$$

This expression is identical with the classical formula $\bar{x} = N(1 - F_m)$ in the Bernoulli distribution (1.1), since the mean of $1 - F_m$ obtained from (1.2) is $m/(n + 1)$. In both distributions the means need not be integers. The mean number of exceedances over the smallest value is n times the mean number of exceedances over the largest value. If $N = n + 1$, we have $\bar{x}_m = m$, and the same holds if n and N are large. If n is odd, and $m = (n + 1)/2$, the mean number of exceedances over the median of n observations is $N/2$. The means $\bar{x}_m$ are traced against m in Graph 2 for $n = N = 9$, and $n = N = 10$.

The mean number ${}_m\bar{x}$ of exceedances over the mth value from below is related to $\bar{x}_m$ by

$$\bar{x}_m + {}_m\bar{x} = N. \tag{3.12}$$

The variances $\sigma_m^{2'}$ and ${}_m\sigma^2$ of the number of exceedances over the mth values from above and below become, from (3.10),

$$\overline{x^2} - \bar{x}^2 = \frac{mN}{n+1}\left(1 + \frac{(m+1)(N-1)}{n+2} - \frac{mN}{n+1}\right).$$

The choice of a common denominator leads, after trivial calculations, to

$$\sigma_m^2 = \frac{mN(n-m+1)(N+n+1)}{(n+1)^2(n+2)} = {}_m\sigma^2. \tag{3.13}$$

The variances increase with N and diminish strongly with increasing n. *The variance is maximum* for $m = (n+1)/2$, i.e. *for the median observation* where it becomes

$$\sigma^2_{(n+1)/2} = \frac{N}{4}\frac{(N+n+1)}{(n+2)}. \tag{3.13'}$$

The variances of the number of exceedances over the largest and the smallest value are

$$\sigma_1^2 = \frac{nN(N+n+1)}{(n+1)^2(n+2)} = {}_1\sigma^2. \tag{3.13''}$$

The quotient of the variances of the median and of the extremes is

$$\frac{\sigma^2_{(n+1)/2}}{\sigma_1^2} = \frac{(n+1)^2}{4n} = \frac{\sigma^2_{(n+1)/2}}{{}_1\sigma^2}. \tag{3.14}$$

Consequently the variance of the median is about $n/4$ times larger than the variance of the extremes. In other words, *the extremes are more reliable* than the median, and this quality increases with the sample size. This is a generalization of the relation obtained in paragraph 2. Such a behavior seems singular. However, it also holds for the uniform distribution, and for the distribution (1.2) of the frequencies [1].

In Bernoulli's case, the variance σ_B^2 is, after replacing $1 - F_m$ by $m/(n+1)$,

$$\sigma_B^2 = N\frac{m}{(n+1)}\frac{(n-m+1)}{(n+1)},$$

whence, from (3.13),

$$\sigma_m^2 = \sigma_B^2\frac{N+n+1}{n+2} > \sigma_B^2.$$

The variance in our case is larger than in Bernoulli's case, since we do not assume the knowledge of the probability F_m which is required for the Bernoulli distribu-

tion. For $N = n + 3$, the variance becomes twice the variance of the Bernoulli distribution. This is a generalization of formula (2.8).

4. The mode and the median. We ask for the most probable number $\tilde{x}$ of exceedances over the previous mth largest among n observations in N future trials. If a mode exists, it must be an integer. Since the distribution $w(x)$ decreases (or increases) with x for $m = 1$ (or $m = n$) we only consider

$$2 \leqq m \leqq n - 1. \tag{4.1}$$

The mode is obtained from the inequalities

$$w(n, m, N, x - 1) \leqq w(n, m, N, x) \geqq w(n, m, N, x + 1) \tag{4.2}$$

which lead, from (1.5) to

$$(m - 1)\frac{N + 1}{n - 1} - 1 \leqq \tilde{x} \leqq (m - 1)\frac{N + 1}{n - 1}. \tag{4.3}$$

The length of the interval is unity, as for the Bernoulli distribution.

There are several cases where *two* modes exist.

a) Let the number of future trials N be such that

$$N = k(n - 1) - 1 \tag{4.4}$$

where k is a positive integer. Then the modes are, from (4.3)

$$\tilde{x}_{(1)} = k(m - 1) - 1; \tilde{x}_{(2)} = k(m - 1). \tag{4.5}$$

b) The modes (4.5) also hold if n and N are large compared to unity, and if $N = k'n$, where k' is again an integer.

c) If n is odd, the median of the initial variate has the rank $m = (n + 1)/2$. If, at the same time, N is odd, there are two modes, namely

$$\tilde{x}_{(1)} = (N - 1)/2; \tilde{x}_{(2)} = (N + 1)/2. \tag{4.6}$$

In the case $N = n$, the two modes $\tilde{x}_{(1)} = m - 1$, and $\tilde{x}_{(2)} = m$ differ by unity from the modes valid in the two previous cases.

In the case $n = N$, and $m \neq (n + 1)/2$, only one mode exists. To find its location, consider first the case that $n = N$ is even, and $m \leqq n/2$. Then the upper limit in (4.3) is

$$[m - 1] + \frac{2}{n - 1}(m - 1) \leqq [m - 1] + 1 - \frac{1}{n - 1} < [m].$$

Since the interval has unit length, the mode is $\tilde{x} = m - 1$. If $m > (n + 1)/2$, the lower limit is

$$[m - 2] + \frac{2}{n - 1}(m - 1) > [m - 1].$$

The case that $n = N$ is odd is treated in the same way, and leads to the follow-

ing result: The most probable numbers of exceedances over the mth value in $N = n$ future trials are

$$\tilde{x} = m - 1 \text{ for } m \leqq n/2;\ \tilde{x} = m \text{ for } m > (n/2) + 1, \quad \text{if } n = N \text{ is even,}$$

(4.7)

$$\tilde{x} = m - 1 \text{ for } m \leqq (n+1)/2;\ \tilde{x} = m \text{ for } m \geqq (n+1)/2, \quad \text{if } n = N \text{ is odd.}$$

We now consider the median. If the probabilities $w(x)$ are summed up from $x = 0$ onward, there may exist an integer $\check{x}_m$ such that the probability for at most $\check{x}_m$ exceedances is $\frac{1}{2}$. This is the median number of exceedances. Such a number need not exist. Assume, for example, $N < n$, then the probability $w(n, 1, N, 0)$ alone (see (1.6)) surpasses $\frac{1}{2}$, and the number of exceedances over the largest and the smallest value do not possess a median. If the median $\check{x}_m$ exists, it follows from the symmetry (1.4) that $N - \check{x}_m - 1$ is the median of the number of exceedances over the mth value from below. The relation

$$\check{x}_m + {}_m\check{x} = N - 1 \tag{4.8}$$

differs from the corresponding relation (3.12) for the mean. In some special cases, the median can be obtained immediately. For $x = 0, m = 1, n = N$, formula (1.6) leads to

$$w(n, 1, n, 0) = \tfrac{1}{2} = w(n, n, n, n).$$

The probability that the largest (or smallest) of n past observations will never (or always) be exceeded in n future trials is equal to $\frac{1}{2}$. If n and N are odd, and $m = (n+1)/2$, the summation of equation (1.7) yields, with the help of (1.3′),

$$\sum_0^x w(z) = \sum_{N-x}^{N} w(z) = 1 - \sum_{x+1}^{N} w(z).$$

Now the median number of exceedances $\check{x}$ is such that the two sums on the right sides are equal to $\frac{1}{2}$. Consequently the median number of exceedances in this case is $m - 1$.

We claim that

$$\check{x}_m = m - 1 \tag{4.9}$$

for all m, provided that $n = N$. For the proof, consider the probability $W(n, m, N, x)$ that the mth largest value is exceeded at most x times in N future trials. This is the sum of the first $x + 1$ members $w(x)$. Let $F_v(\alpha, \beta, \gamma, 1)$ be the sum of the first v members of the hypergeometric series (3.1). Then the substitutions (3.4) and $v = x + 1$ lead to

$$W(n, m, N, x) = \frac{\binom{n}{m}}{\binom{N+n}{m}} F_{x+1}(m, -N, m - n - N, 1). \tag{4.10}$$

For the sums of the hypergeometric series $F(\alpha, \beta, \gamma, 1)$ the following recurrence formula [2] is used.

$$\frac{(\gamma-\beta-\alpha)(\gamma-\beta-\alpha+1)\cdots(\gamma-\beta-1)}{(\gamma-\alpha)(\gamma-\alpha+1)\cdots(\gamma-1)}F_v(\alpha,\beta,\gamma,1)$$

$$= 1 - \frac{\beta(\beta+1)\cdots(\beta+v-1)}{(\gamma-\alpha)(\gamma-\alpha+1)\cdots(\gamma-\alpha+v-1)} \tag{4.11}$$

$$F_\alpha(v, \gamma-\beta-\alpha, \gamma-\alpha+v, 1).$$

The substitutions used in (3.4), and $v = x + 1$ lead to

$$\frac{(-n)(-n+1)\cdots(-n+m-1)}{(-n-N)(-n-N+1)\cdots(-n-N+m-1)}F_{x+1}(m,-N,m-n-N,1)$$

$$= 1 - \frac{(-N)(-N+1)\cdots(-N+x)}{(-n-N)(-n-N+1)\cdots(-n-N+x)}$$

$$F_m(x+1, -n, -n-N+x+1, 1).$$

This equation may be written from (4.10)

$$W(n, m, N, x) = 1 - \frac{\binom{N}{x+1}}{\binom{N+n}{x+1}} F_m(x+1, -n, -n-N+x+1, 1). \tag{4.12}$$

For $x = m - 1$, and $N = n$, the equation becomes

$$W(n,m,n,m-1) = 1 - \frac{\binom{n}{m}}{\binom{2n}{m}} F_m(m, -n, -2n+m, 1).$$

From (4.10) it follows that the second factor on the right side is equal to the left side

$$W(n, m, n, m-1) = \frac{\binom{n}{m}}{\binom{2n}{m}} F_m(m, -n, -2n+m, 1).$$

Consequently

$$W(n, m, n, m-1) = \tfrac{1}{2}. \tag{4.13}$$

If $n = N$, the median number $\breve{x}_m$ of exceedances over the mth largest value is $m - 1$, as stated previously. The means, modes, and medians obtained from the exact formulae (3.11), (4.7) and (4.9) are traced in graph (2) for $n = N = 9$, and $n = N = 10$.

5. Probabilities of at least one exceedance. If we sum up the probabilities $w(x)$ from zero up to a certain x (or from a certain x up to N), we obtain the probabilities $W(x)$ (or $P(x)$) for at most (or at least) x exceedances over the mth past value in N future trials

$$W(x) = \sum_{z=0}^{x} w(z); \qquad P_{(x)} = \sum_{z=x}^{N} w(z) \tag{5.1}$$

where

$$W(x) + P(x-1) = 1; \; W(x-1) + P(x) = 1.$$

The boundary conditions are

$$W(0) = w(0); \; W(N) = 1; \; P(0) = 1; \; P(N) = w(N).$$

From the symmetry (1.4) it follows that the probability for the mth value from above to be exceeded at most x times is equal to the probability for the mth value from below to be exceeded at least $N - x$ times.

From (5.1) and (1.3) it follows for $m = 1$ (and $m = n$) that the probabilities for the largest (or smallest) among n observations to be exceeded at most once in n future trials converges toward 3/4 (or zero), respectively. If n is large, the probability that the largest value will be exceeded at most x times in n future trials is, by virtue of (2.5),

$$W(\underline{n}, 1, \underline{n}, x) = 1 - (\tfrac{1}{2})^x = P(\underline{n}, \underline{n}, \underline{n}, \underline{n} - x) \tag{5.2}$$

independent of n.

Consider now the probability that the mth largest value will be exceeded *at least once* in N future trials

$$P(n, m, N, 1) = 1 - \frac{n!}{(n-m)!}\frac{(N+n-m)!}{(N+n)!} \tag{5.3}$$

$$= W(n, n-m+1, N, N-1).$$

If N and n are large, and m is small, this expression becomes

$$P(\underline{n}, m, \underline{N}, 1) = 1 - \left(\frac{n}{n+N}\right)^m = W(\underline{n}, \underline{n} - m + 1, \underline{N}, \underline{N} - 1)$$

For $m = 1$ and $n = N$, the probability is $\frac{1}{2}$, independent of the size of n.

The least number of exceedances over the smallest value for given probabilities P, called the *tolerance limit*, has been derived by S. S. Wilks [3]. A related problem is the following: How many trials N have to be made in order that there is a given probability α for the mth largest value to be exceeded at least once? By virtue of (5.3) we obtain N from

$$\frac{n!(N+n-m)!}{(n-m)!(N+n)!} = 1 - \alpha. \tag{5.4}$$

For the largest value $m = 1$, this equation leads to

$$\frac{N}{n} = \frac{1}{1-\alpha} - 1. \tag{5.5}$$

Of course, N/n increases with α. If n is large, and m remains small, equation (5.4) leads, in first approximation, to

$$\frac{N}{n} = (1-\alpha)^{-1/m} - 1. \tag{5.6}$$

The quotients N/n as function of α are traced in graph (3). The quotient is plotted vertically against $1/(1-\alpha)$ plotted horizontally, both in logarithmic scales. The abscissa shows the probability α. The curve for $m = 1$ is exact. The corresponding curves for the penultimate and the two preceding values ($m = 2, 3, 4$) are obtained from the approximation (5.6). The graph reads in the following way: The probability that the largest, or second, or third, or fourth value from above are exceeded at least once in $100n$, or $9n$, or $3.6n$, or $2.2n$ future trials is $\alpha = .99$. Inversely, in $4n$ future trials the probability that the largest, or the second, or the third, or fourth extreme value is exceeded at least once is $\alpha = 0.80$, or 0.96, or 0.992, or 0.9984, respectively.

In a similar way we calculate the probabilities that the largest (and penultimate) among n observations is exceeded at least twice in N future trials. Let α_2 be this probability. Then we have for the largest value

$$\begin{aligned} 1-\alpha_2 &= w(n, 1, N, 0) + w(n, 1, N, 1) \\ &= \frac{n}{n+N}\left(1 + \frac{N}{n+N-1}\right). \end{aligned}$$

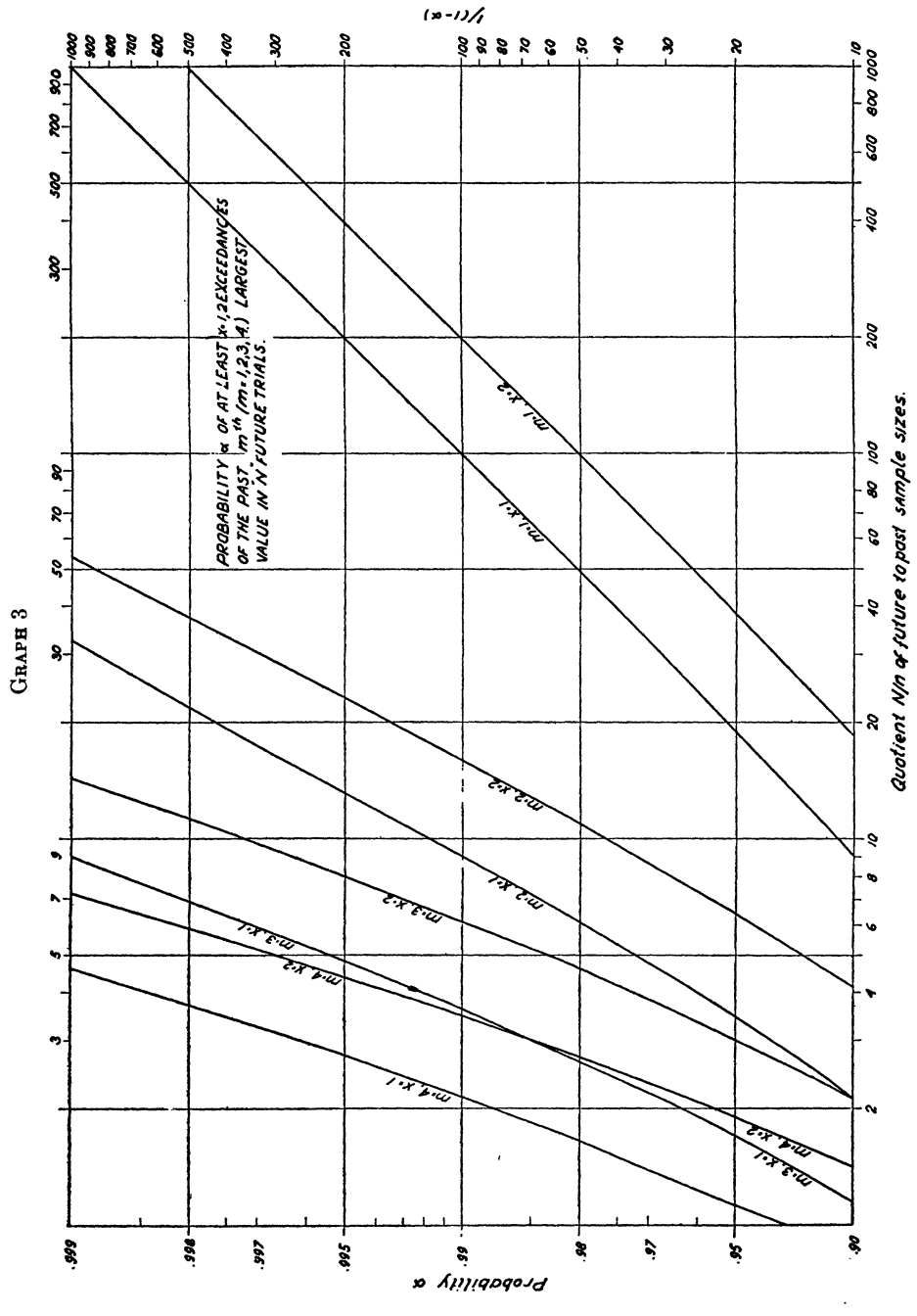
For n sufficiently large, the expression simplifies to

$$\frac{1}{1-\alpha_2} = \frac{\left(\frac{N}{n}+1\right)^2}{\frac{2N}{n}+1}. \tag{5.7}$$

The probability α_2 as function of N/n is also traced in Graph (3) and designated by $m = 1, x = 2$. Finally, for $m = 2$ the probability α_2 for the penultimate value to be exceeded at least twice is obtained for large n by

$$\frac{1}{1-\alpha_2} = \frac{\left(\frac{N}{n}+1\right)^3}{\frac{3N}{n}+1}. \tag{5.8}$$

This probability α_2 is also traced in Graph 3 and designated by $m = 2, x = 2$. If we fix the probabilities α_2, the graph shows the number of future trials corresponding to 1 and 2 exceedances over the largest, the penultimate, and the two preceding observations.


PROBABILITY α OF AT LEAST x=1,2 EXCEEDANCES OF THE PAST m^{th} $(m=1,2,3,4)$ LARGEST VALUE IN N FUTURE TRIALS.
m=1 x=2
m=1 x=1
m=2 x=2
m=2 x=1
m=3 x=2
m=3 x=1
m=4 x=2
m=4 x=1
Probability α
Quotient N/n of future to past sample sizes.

GRAPH 3

6. Applications. In 50% of all cases, the largest (or smallest) of n past observations will not (or always) be exceeded in $N = n$ future trials. The mean number of exceedances is the mean in the Bernoulli distribution. *The variance is largest for the median, and smallest for the extremes*, and this superiority of the extremes increases with the sample size.

If the previous, and the future sample sizes both are large and equal, the distribution of the number of exceedances over the median observation is normal with mean and variance of the order $n/2$, whereas the distribution of the exceedances over the mth extremes (the law of rare exceedances), similar to the Poisson distribution, has the mean m, and the variance $2m$, m being small compared to the sample size. Elementary calculations lead to the setting of sample sizes N corresponding to given probabilities for 1 or 2 exceedances over the past largest and penultimate observation.

These methods may be of interest for forecasting floods if, instead of the size of the flood, we are interested only in the frequency. The same procedure may also be applied to other meteorological phenomena such as droughts, the extreme temperatures (the killing frost), the largest precipitations, etc., and permits to forecast the number of cases surpassing a given severity within the next N years.

REFERENCES

[1] E. J. GUMBEL, "Simple tests for given hypotheses," *Biometrika*, Vol. 32 (1942).

[2] H. VON SCHELLING, "A formula for the partial sums of some hypergeometric series," *Annals of Math. Stat.*, Vol. 20 (1949), No. 1.

[3] S. S. WILKS, "Statistical predictions with special reference to the problem of tolerance limits," *Annals of Math. Stat.*, Vol. 13 (1942).