63

LIMITING FORMS OF THE FREQUENCY DISTRIBUTION OF THE LARGEST OR SMALLEST MEMBER OF A SAMPLE

Author's Note (CMS 15.179a) *

* Reproduced from "Contributions to Mathematical Statistics" (1950) by permission of John Wiley & Sons, Inc.

The distribution of the largest and smallest observation in a sample of given size offers considerable numerical difficulties. It is here shown that the limiting forms are few and comparatively simple, although with a normal distribution they are approached exceedingly slowly.

 Proceedings of the Cambridge Philosophical Society, 24: 180-190, (1928).

Reproduced from the Proceedings of the Cambridge Philosophical Society, v. 24: 180-190 (1928)

Limiting forms of the frequency distribution of the largest or smallest member of a sample. By R. A. FISHER, Sc.D., Gonville and Caius College, and L. H. C. TIPPETT, M.Sc.

[*Received* 24 November, *read* 5 December, 1927.]

1. *Introductory.*

In a previous paper on the subject of the distribution of the largest member of a sample from a normal population, one of the authors has given constants involving the first four moments for samples up to 1000. In this paper, possible limiting forms of such distributions in general are discussed. It will appear that a particular group of distributions provides the limiting distributions in all cases, and that the case derived from the normal curve is peculiar for the extreme slowness with which the limiting form is approached.

2. *The possible limiting forms deduced from the functional relation which they must satisfy.*

Since the extreme member of a sample of mn may be regarded as the extreme member of a sample of n of the extreme members of samples of m, and since, if a limiting form exist, both of these distributions will tend to the limiting form as m is increased indefinitely, it follows that the limiting distribution must be such that the extreme member of a sample of n from such a distribution has itself a similar distribution.

If P is the probability of an observation being less than x, the probability that the greatest of a sample of n is less than x is P^n, consequently in the limiting distributions we have the functional equation

$$P^n(x) = P(a_n x + b_n);$$

the solutions of this functional equation will give all the possible limiting forms.

If a is not equal to unity, then

$$x = ax + b,$$

when $$x = \frac{b}{1-a},$$

and at this point $$P^n = P,$$

$$P = 0 \text{ or } 1,$$

consequently the solutions fall into three classes:

I.	$a = 1$,	$P^n(x) = P(x + b_n)$,
II.	$P = 0$ when $x = 0$,	$P^n(x) = P(a_n x)$,
III.	$P = 1$ when $x = 0$,	$P^n(x) = P(a_n x)$.

I. If $$P^n(x) = P(x + b_n),$$

then $$n \log P(x) = \log P(x + b_n),$$

and $$\log n + \log(-\log P(x)) = \log(-\log P(x + b_n));$$

therefore the expression $\log(-\log P(x)) - \dfrac{x \log n}{b_n}$ is constant, or periodic, with period b_n.

Now for all values of m and n

$$b_{mn} = b_m + b_n,$$

and if b_n is an analytic function of n, a supposition which excludes the periodic solution,

$$nb'_{mn} = b'_m, \qquad mb'_{mn} = b'_n,$$

whence $$mb'_m = nb'_n,$$

or $$b'_n = \frac{c}{n},$$

and $$b_n = c \log n + d,$$ where c and d are constants.

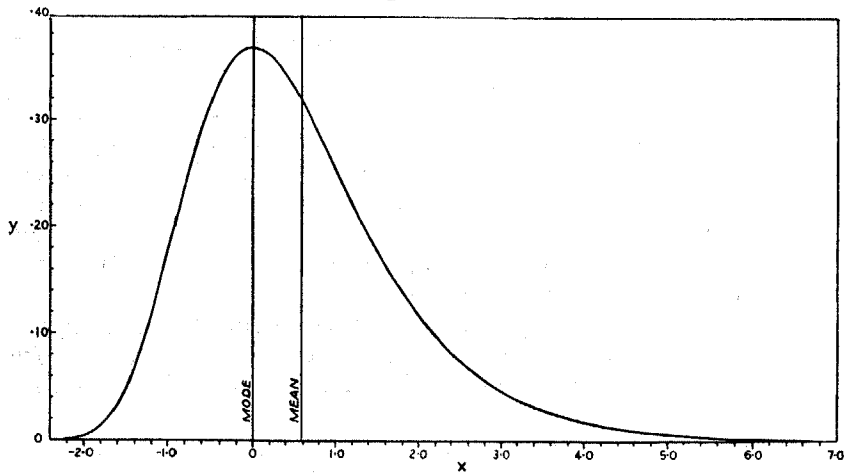

Fig. 1. Distribution $p = e^{-e^{-x}}$, or $dp = e^{-x-e^{-x}}dx$ represented by the curve $y = e^{-x-e^{-x}}$.

Putting $n = 1$, it appears that

$$d = 0.$$

Hence $$\log(-\log P_x) = \frac{x}{c} + \text{constant},$$

or, the limiting form is that of $-\log(-\log P_x) = x$, for c must be negative since x is assumed to increase with P. The distribution of the greatest of a sample of n from this distribution is

$$-\log(-\log P_x) = x - \log n,$$

the distribution being merely shifted, without change of size or form, through a distance $\log n$. The curve is shown in Fig. 1.

II and III. If $$P^n(x) = P(a_n x),$$

$$a_{mn} = a_m a_n,$$

and, if a is an analytic function of n,

$$ma'_{mn} = a_m a'_n, \qquad na'_{nm} = a'_m a_n,$$

whence $$\frac{a'_n}{a_n} = \frac{-1}{kn},$$

of which the solution is

$$\log a_n = \frac{-1}{k}\log n, \qquad a_n = n^{-1/k},$$

since $a = 1$ when $n = 1$.

Now $\log(-\log P(x))$ is increased by $\log n$ when $\log x$ is increased by $\log a_n$, so that, excluding as before the periodic solution,

$$\log(-\log P(x)) - \frac{\log n \log x}{\log a_n}$$

must be constant. This gives

$$\log(-\log P(x)) = -k(\log x + c)$$

or $$-\log P(x) = (Ax)^{-k}.$$

If $P = 0$ when $x = 0$, k will be positive (II).

The form of the curve is then that of

$$P = e^{-x^{-k}},$$

$$dP = \frac{k}{x^{k+1}} e^{-x^{-k}} dx, \text{ where } k \text{ is positive.}$$

If $P = 1$ when $x = 0$, k will be negative and all possible values of x will be negative; in this case (III) the form of the curve is given by

$$-\log P = (-x)^k, \text{ where } k \text{ is positive,}$$

$$P = e^{-(-x)^k},$$

$$dP = k(-x)^{k-1} e^{-(-x)^k} dx.$$

The only possible limiting curves are therefore:

I. $$dP = e^{-x - e^{-x}} dx,$$

in which the effect of selecting the greatest value of a sample of n is merely to shift the curve, without affecting its scale, through a distance $\log n$.

II. $$dP = \frac{k}{x^{k+1}} e^{-x^{-k}} dx,$$

in which the effect of selection is to increase the scale of the curve by the factor $n^{1/k}$, maintaining the terminus $x = 0$ unchanged.

III. $$dP = k(-x)^{k-1} e^{-(-x)^k} dx,$$

in which the effect of selection is to decrease the scale of the curve by the factor $n^{-1/k}$, while maintaining the terminus $x = 0$ unchanged. In this case alone will the selected curve increase materially in accuracy as selection is increased; the weight of an observation, from curves of constant form, will be inversely proportional to the square of the scale, and will be proportional to $n^{2/k}$. The accuracy of the extreme observation will therefore increase more rapidly than that of, for example, the mean, if k is less than 2.

3. *The limiting form appropriate to any particular frequency distribution.*

If in any frequency distribution p is the probability of an observation being less than x, and if as $p \to 1$ the quantity

$$(1-p)\,x^k$$

tends to a finite limit, a^k, then it is evident that $P = p^n$ will have the form

$$P = e^{-na^k x^{-k}}$$

in the limit for large samples of n.

Since, for any two values of P other than 0 and 1, the values of x as n tends to infinity tend to the finite ratio of the values of

$$(-\log P)^{-1/k}$$

the limiting form of the distribution will be the same if

$$1 - p = x^{-k}\phi(x),$$

where the range of $\log \phi$, for any finite range of $\log x$, tends to zero as x tends to infinity.

The scale of the distribution for the greatest of n, measured by $an^{1/k}$, will in such cases approach the limit

$$(\phi n)^{1/k},$$

where the argument of ϕ is given by the equation

$$x^k = n\phi(x).$$

Equally, for any frequency distribution for which

$$(1-p)\,e^{x/c}$$

tends to a finite limit A as p tends to unity, the limiting forms of the distribution of the largest of a sample of n will be given by

$$P = e^{-nAe^{-x/c}}.$$

Since, in this case, for any two values of P other than 0 and 1, the difference of the two values of x/c tends to a constant value, the limiting form of distribution will be the same if

$$1-p=e^{-x/c}\,\phi\,(x),$$

when the range of $\log\phi$ in any finite range of x/c tends to zero as x tends to infinity. Thus, if c is constant, $\phi\,(x)$ may contain factors such as x^t. The location of the distribution, given by

$$\frac{x}{c}=\log\,(nA),$$

will then, as the limiting form is approached, change as $c\log\,(n\phi)$, in which the argument of ϕ is given by the equation

$$x=c\log\,(n\phi\,(x)).$$

The case in which c is constant does not exhaust the applications of this limiting form, for whatever function $1-p$ may be of x, if we write

$$\frac{1}{c}=-\frac{d}{dx}\log\,(1-p),$$

then, if the range of $\log\,(1-p)+x/c$ from $x=\xi$ to $x=\xi+ct$, tends to zero, as x tends to infinity, for all real values of t, then will the same limiting form be valid.

For example, let

$$1-p=e^{-x^r},$$

then

$$c=\frac{1}{r\xi^{r-1}},$$

and

$$\frac{ct}{\xi}=\frac{t}{r\xi^r},$$

which tends to zero, if r is positive, for all values of t.

But

$$\log\,(1-p)+x/c=rx\xi^{r-1}-x^r$$

$$=(r-1)\,\xi^r-\frac{r\,(r-1)}{2}\,\xi^{r-2}c^2t^2+\text{smaller terms};$$

the range will therefore tend to zero, for

$$\frac{r\,(r-1)}{2}\,\xi^{r-2}c^2t^2=\frac{r-1}{2r}\,\frac{t^2}{\xi^r},$$

which tends to zero, for all values of t, if r is positive.

The parameter c, which measures the scale of the distribution, will increase if $r<1$, and decrease if $r>1$, while the location of the mode as the limiting form is approached is given in general by

$$P=e^{-1},$$

or

$$n(1-p)=1.$$

Again, for the normal curve with unit standard deviation

$$1-p=\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}x^{-1}(1-X),$$

where X tends to zero as x tends to infinity,

$$\frac{1}{c}=\xi+\frac{1}{\xi}-X,$$

$$\log(1-p)+x/c=-\tfrac{1}{2}x^2-\log x+x\xi+1-\tfrac{1}{2}\log(2\pi)-X$$

$$=+\tfrac{1}{2}\xi^2-\log\xi+1-\tfrac{1}{2}\log(2\pi)-X,$$

where X vanishes as $x\to\infty$, at all values of x from ξ to $\xi+ct$.

For sufficiently large samples of n from a normal curve, the distribution of the largest of the sample will be centred about a mode m given by

$$e^{\frac{1}{2}m^2}m\sqrt{2\pi}=n,$$

with scale given by

$$c=\frac{m}{m^2+1}.$$

4. *The approach of the distribution of the greatest of a normal sample to its final form.*

The final form for the largest of a normal sample has been shown to be given by

$$P=e^{-e^{-x/c}},$$

where c diminishes to zero as the sample increases, in such a way that to the degree of approximation required in very large samples

$$c=\frac{m}{m^2+1}$$

and

$$e^{+\frac{1}{2}m^2}m\sqrt{2\pi}=n.$$

Since for any finite value of m, however large, c will still be diminishing as n increases, the case has an analogy at any stage with the distribution derived from

$$p=e^{-(-x)^k},$$

in which also the scale diminishes as n increases. This analogy

may be utilised by equating the rate of change of the scale with increasing n in the two cases.

Now, for $$P = e^{-n(-x)^k},$$

we have $$dP = kn(-x)^{k-1} e^{-n(-x)^k} dx,$$

so that the logarithm of the ordinate at any point is

$$(k-1)\log(-x) - n(-x)^k + \text{constant},$$

giving as equation for the mode, m,

$$(-x)^k = \frac{k-1}{nk},$$

whence $$\frac{d\log(-x)}{d(\log n)} = -\frac{1}{k}.$$

But for the normal curve

$$\frac{d\log c}{d\log n} = \frac{d\log c}{dm}\frac{dm}{d\log n} = -\frac{m^2-1}{(m^2+1)^2}.$$

Hence the distribution in which

$$\frac{m^2-1}{(m^2+1)^2} = \frac{1}{k} = h$$

should provide a penultimate form of approximation, which will duly tend to the ultimate form as h tends to zero.

5. *The moments of the ultimate and penultimate forms.*

The moments of the ultimate form

$$dP = e^{-x-e^{-x}} dx$$

may be found most directly from the generating function of the semi-invariants

$$K = \log M,$$

where $$M = \int_{-\infty}^{\infty} e^{tz} dP.$$

For, writing z for e^{-x},

$$M = \int_0^{\infty} z^{-t} e^{-z} dz = (-t)!$$

and $$K = \log M = \gamma t + \frac{\pi^2}{6}\frac{t^2}{2!} + \frac{t^3}{3!}\left(-\frac{d^3}{dz^3}\log x!\right)_{x=0} + \dots,$$

whence it follows that the distance of the mean from the mode is

$$\mu_1' = \gamma = \cdot 577215665,$$

the variance is $$\mu_2 = \frac{\pi^2}{6} = 1\cdot 64493407,$$

the third moment is

$$\mu_3 = 2\left\{1 + \frac{1}{2^3} + \frac{1}{3^3} + \dots\right\} = 2\cdot 40411381,$$

while the fourth moment is given by

$$\mu_4 - 3\mu_2^2 = 6\left\{1 + \frac{1}{2^4} + \frac{1}{3^4} + \ldots\right\} = \frac{\pi^4}{15} = 6{\cdot}4939394.$$

Consequently, for sufficiently large samples we shall have

$$\text{Mean} - \text{Mode} = \gamma c = \frac{\gamma m}{m^2 + 1},$$

$$\text{Variance} = \frac{\pi^2}{6} c^2,$$

$$\beta_1 = 1{\cdot}2985676,$$

$$\beta_2 = 5{\cdot}4.$$

For the penultimate form

$$dP = k(-x)^{k-1} e^{-(-x)^k} dx,$$

writing $-x = \frac{1}{t^k} = t^{\frac{1}{k}}$,

we have $dP = -e^{-t} dt$,

and $\mu_r' = (-)^r (-x)^r dP = (-)^r t^{\frac{r}{k}} dP = (-)^r (kr)!$;

also the mode is given by

$$-x = (1-h)^h.$$

Hence we have as penultimate formulae

$$\text{Mean} - \text{Mode} = \frac{c}{h}\{(1-h)^h - h!\},$$

$$\text{Variance} = \frac{c^2}{h^2}\{(2h)! - (h!)^2\},$$

together with β_1 and β_2 expressed in terms of h only.

The extreme slowness with which the ultimate form is approached is well shown by the fact that even for enormous samples the penultimate form is still materially different in its β coefficients. The following tables show, for different values of h, the corresponding values of m and n, and, in parallel columns, the distance of the mean from the mode, the variance and the β coefficients. It will be observed that even for samples of nearly a billion the penultimate form is still considerably different from the ultimate form. The appropriateness of the penultimate form for samples of 1000 downwards can be tested from the results given in a previous paper*, using for m the value of x for which $p = 1/n$, and the corresponding values of c and h.

* *Biometrika*, XVII, pp. 364–387 (1925).

It is apparent that the penultimate form effectively bridges the great gap between samples of 1000 or less and the ultimate

TABLE A.

h	m	n	Mean − Mode		Standard Deviation		β_1		β_2	
			Ultimate	Pen-ultimate	Ultimate	Pen-ultimate	Ultimate	Pen-ultimate	Ultimate	Pen-ultimate
0	∞	∞	0	0	0	0	1·2986	1·2986	5·400	5·400
0·02	6·8493	$264 \cdot 10^9$	·0825	·0769	·1833	·1787	1·2986	1·0503	5·400	4·878
0·04	4·6699	$637 \cdot 10^3$	·1182	·1020	·2626	·2499	1·2986	·8436	5·400	4·451
0·06	3·6528	7228	·1470	·1169	·3266	·3039	1·2986	·6709	5·400	4·100
0·08	3·0000	677	·1732	·1261	·3848	·3504	1·2986	·5267	5·400	3·810

TABLE B.

h	n	m	Mean − Mode		Standard Deviation		β_1		β_2	
			Pen-ultimate	Actual	Pen-ultimate	Actual	Pen-ultimate	Actual	Pen-ultimate	Actual
·0768	1000	3·0902	·1249	·1262	·3433	·3514	·548	·618	3·852	4·088
·0845	500	2·8782	·1276	·1287	·3604	·3704	·498	·570	3·751	4·003
·0967	200	2·5758	·1309	·1314	·3874	·4009	·425	·495	3·607	3·875
·1073	100	2·3263	·1334	·1323	·4124	·4294	·368	·429	3·493	3·765
·1154	60	2·1281	·1355	·1318	·4340	·4545	·328	·376	3·414	3·677

form for very large samples. The distance from mode to mean is given correctly for samples somewhere between 100 and 200 and

is underestimated by an amount which seems to attain a maximum of about 1 °/₀ for samples of about 1000, whereas the value for the ultimate form is over 7 °/₀ in error for samples of nearly a billion. The standard deviation is given by the penultimate form with a negative error of about 2½ °/₀ at 1000 and only about 4½ °/₀ at 60, while the ultimate form is nearly 10 °/₀ out at 1000, and just under 3 °/₀ at a billion. In both comparisons the largest deviations occur in the β coefficients. The latter are consistently too low in the penultimate form for samples of 1000 and less, and probably do not attain a close approximation until the sample number is nearly a million, while an equally good approximation to the ultimate values $\beta_1 = 1{\cdot}299$ and $\beta_2 = 5{\cdot}4$ would only be attained by such incredibly large samples as are represented by values of about ·004 for h (c. 10^{55}). The changes in β_1 and β_2 with varying h, together with the actual values for samples up to 1000, are shown in Figs. 2 and 3.

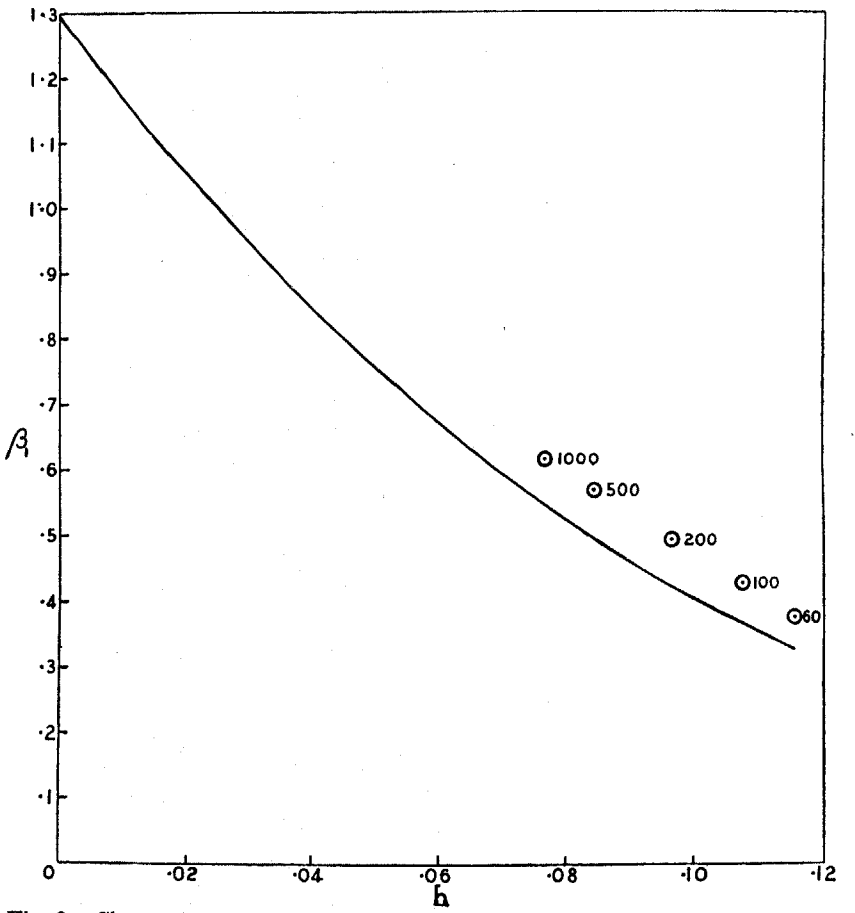

Fig. 2. Change in β_1 with sample size as indicated by the penultimate formula, with actual values for samples up to 1000.

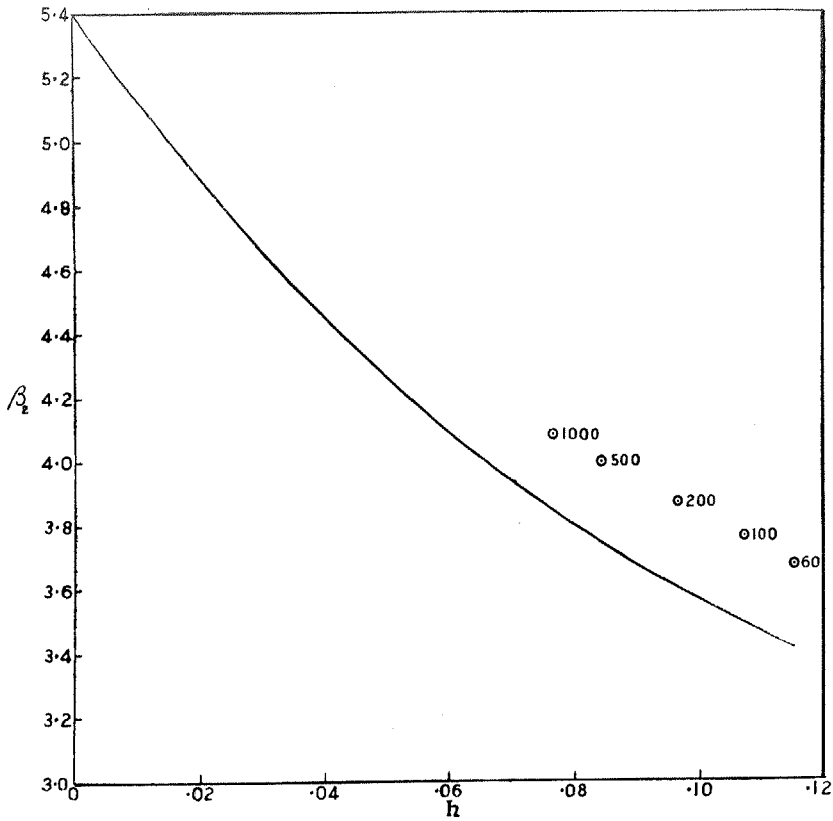

Fig. 3. Change in β_2 with sample size as indicated by the penultimate formula, with actual values for samples up to 1000.

6. *Summary.*

The limiting distribution, when n is large, of the greatest or least of a sample of n, must satisfy a functional equation which limits its form to one of two main types. Of these one has, apart from size and position, a single parameter h, while the other is the limit to which it tends when h tends to zero.

The appropriate limiting distribution in any case may be found from the manner in which the probability of exceeding any value x tends to zero as x is increased. For the normal distribution the limiting distribution has $h = 0$.

From the normal distribution the limiting distribution is approached with extreme slowness; the final series of forms passed through as the ultimate form is approached may be represented by the series of limiting distributions in which h tends to zero in a definite manner as n increases to infinity.

Numerical values are given for the comparison of the actual with the penultimate distributions for samples of 60 to 1000, and of the penultimate with the ultimate distributions for larger samples.