ZASTOSOWANIA MATEMATYKI
APPLICATIONES MATHEMATICAE
Hugo Steinhaus Jubilee Volume
X (1969)

F. A. HAIGHT (Los Angeles)

TWO PROBABILITY DISTRIBUTIONS CONNECTED WITH ZIPF'S RANK-SIZE CONJECTURE

1. Definition. Consider, for fixed $\beta > 0$, $Z > 1$, a table of decimal values of $ZN^{-\beta}$, $N = 1, 2, 3, \ldots$ The number N_n of values nearest the integer n (counting half integers as belonging to the upper category) is the number of N which satisfy the inequality

$$\tfrac{1}{2}(2n-1) \leqslant ZN^{-\beta} < \tfrac{1}{2}(2n+1).$$

Writing the inequality in the form

$$\left(\frac{2Z}{2n-1}\right)^{1/\beta} \geqslant N > \left(\frac{2Z}{2n+1}\right)^{1/\beta}$$

we find that

$$(1) \qquad N_n = \left[\left(\frac{2Z}{2n-1}\right)^{1/\beta}\right] - \left[\left(\frac{2Z}{2n+1}\right)^{1/\beta}\right]$$

where $[x]$ means the integral part of x. Since the total number of non-zero values of N_n is $(2Z^{1/\beta})$, the quantities

$$(2) \qquad p_n(Z) = (2Z)^{-1/\beta} N_n, \qquad n = 1, 2, 3, \ldots$$

have the formal properties of a discrete probability distribution. Letting $Z \to \infty$ and writing $\sigma = 1/\beta$, we obtain

$$p_n = p_n(\infty) = (2n-1)^{-\sigma} - (2n+1)^{-\sigma}, \qquad n = 1, 2, \ldots$$

which we shall call the *Zeta distribution with parameter* σ.

The first two moments of the Zeta distribution can be expressed in terms of the Riemann Zeta Function:

$$m = \Sigma n p_n = (1-2^{-\sigma})\zeta(\sigma),$$

$$m_2 = \Sigma n^2 p_n = (1-2^{1-\sigma})\zeta(\sigma-1).$$

In the special case in which $\beta = \sigma = 1$, a trivial probability distribution (with infinite mean) results. If, however, the limit $Z \to \infty$ is not invoked, and Z is treated as a parameter, the distribution

$$(3) \qquad \frac{1}{2Z}\left\{\left[\frac{2Z}{2n-1}\right]-\left[\frac{2Z}{2n+1}\right]\right\}, \quad n = 1, 2, 3, \ldots$$

results, which we shall call the *harmonic distribution with parameter Z*. Here, too, the mean is infinite.

2. Zipf's Conjecture. According to a principle which he called the "principle of least effort", Zipf [5] argued that the size of cities (as well as an enormous number of other natural and social phenomena) ought to form, when ranked in order of magnitude, multiples of the harmonic series N^{-1}, or at least of the series $N^{-\beta}$. Since the publication of Zipf's conjecture, there has been a great deal of criticism of the principle upon which he based it, although the rank-size relationship associated with his name is still used.

Unfortunately, empirical testing of Zipf's conjecture most frequently takes the form of visual judgement of geometric plots of city sizes on graph paper. This is partly because of the nature of the hypothesis, but also because of the very large number of cities and towns which are to be found in any respectable sample.

In the present paper we suggest a more convenient and compact method for testing Zipf's conjecture, which is at the same time more adaptable to the format in which such data is supplied by demographic authorities. The distributions (2) and (3), which have been derived, would correspond to Zipf's conjecture, if, for the Zeta distribution, a large number of decimal places had been "rounded off" in the grouping of cities, or, for the harmonic distribution, if Zipf's simplified (harmonic) conjecture is being tested.

3. Examples. In Table 1, we show the sizes of world metropolitan areas (not municipalities) fitted, by the method of moments, both to the Zeta and to the harmonic distributions. Since the count is taken to the nearest million, $Z = 6$ places have been discarded.

In Table 2, we show the populations of accredited United States colleges and universities, fitted to the Zeta distribution. Since the count is taken to the nearest thousand, $Z = 3$ places have been discarded. The mean value is over three in this table, and therefore the harmonic distribution would require an inordinately large value of Z (over five hundred) and thus is not given.

Unfortunately estimation of the parameters in both of these distributions is difficult. The moments estimate of σ for the Zeta distribution requires evaluation of the Riemann Zeta Function. For approximate purposes, one can use the formula

$$(1-2^{-\sigma})\zeta(\sigma) = \tfrac{1}{2}(\sigma-1)^{-1}+0.63518142+$$
$$+0.11634237(\sigma-1)-0.01876574(\sigma-1)^2+\dots$$

which was kindly communicated to me by Mr. Robert E. Shafer of the Lawrence Radiation Laboratory.

TABLE 1. World metropolitan areas, classified to the nearest million of population. Source: Rand McNally Cosmopolitan World Atlas, pp. 170-171

n	Number of areas having population nearest to $n\times 10^6$	Zeta (σ = 1.24)	Harmonic (Z = 20)
1	141	166.64	151.20
2	46	26.92	28.00
3	14	10.37	16.80
4	6	5.37	5.60
5	5	3.24	5.60
6	1	2.14	0.00
7	3	1.51	5.60
8+	8	7.79	11.20

TABLE 2. United States accredited colleges and universities, classified to the nearest thousand of student population. Source: World Almanac and Book of Facts, 1959 Edition

n	Number of institutions having student population nearest to $n\times 10^3$	Zeta (σ = 0.88)
1	500	589.33
2	157	130.95
3	61	59.13
4	58	34.05
5	34	22.27
6	16	15.76
7	22	11.78
8	15	9.15
9	14	7.33
10	11	6.01
11	11	5.02
12	6	4.26
13	10	3.67
14	3	3.19
15	2	2.80
16	2	2.48
17	6	2.21
18	2	1.99
19+	21	39.62

On the basis of Tables 1 and 2, one would not say that Zipf's conjecture (in either form) is appropriate to this data. The poor fit in the first few categories is especially noticeable.

By more extensive calculations, one could, of course, simply compute probabilities for fixed Z directly from (1), using the values ($Z = 6$ for Table 1, and $Z = 3$ for Table 2) chosen in the grouping.

A general evaluation of Zipf's conjecture is due to Rapaport [3]; treatment of other types of data by this method can be found in papers of Haight [1], Horvath [2] and Simon [4].

References

[1] F. A. Haight, *Some statistical problems in connection with word association data*, Journal of Mathematical Psychology 3 (1966), pp. 217-233.

[2] W. J. Horvath, *A stochastic model for word association tests*, Psychological Review 70 (1963), pp. 361-364.

[3] A. Rapaport, *Rank-size relations*, International Encyclopaedia of the Social Sciences, 1967.

[4] H. A. Simon, *On a class of skew distribution functions*, Biometrika 42 (1955), pp. 425-440.

[5] G. K. Zipf, *Human behavior and the principle of least effort*, Addison-Wesley Press, New York 1949.

INSTITUTE OF TRANSPORTATION
UNIVERSITY OF CALIFORNIA, LOS ANGELES

Received on 29. 8. 1967