The Choice of a Class Interval
Author(s): Herbert A. Sturges
Reviewed work(s):
Source: *Journal of the American Statistical Association*, Vol. 21, No. 153 (Mar., 1926), pp. 65-66
Published by: American Statistical Association
Stable URL: http://www.jstor.org/stable/2965501
Accessed: 06/04/2012 02:17

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at http://www.jstor.org/page/info/about/policies/terms.jsp

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.

American Statistical Association is collaborating with JSTOR to digitize, preserve and extend access to *Journal of the American Statistical Association.*

http://www.jstor.org

For the meeting in 1927 the arguments and inducements of the young government of Egypt prevailed and the General Assembly finally voted that the seventeenth session should be held in Cairo.

WALTER F. WILLCOX

THE CHOICE OF A CLASS INTERVAL

CASE I. COMPUTATIONS INVOLVING A SINGLE SERIES

In case a single statistical series of range R with N items is involved in a computation, the optimal class interval may be estimated from the formula

$$C = \frac{R}{1 + 3.322 \log N}$$

This formula gives the class interval for the computation of the averages, measures of dispersion, skewness, etc., of frequency distributions. It is based on the principle that the proper distribution into classes is given, for all numbers which are powers of 2, by a series of binomial coefficients. For example, 16 items would be divided normally into 5 classes, with class frequencies 1, 4, 6, 4, 1. Thus if a statistical series had 16 items with values ranging from 20 to 70, or a range of 50 points, it should be divided into 5 classes of 10 points each, that is, the class interval would be 10. Similarly, 64 is the sixth power of 2, so a statistical series containing 64 items should be divided into 6 plus 1, or 7 classes. If such a series had a range of 35 points the class interval would be 5.

The most convenient class intervals are 1, 2, 5, 10, 20, etc., so that in practice the formula for the theoretical class interval may be used as a means of choosing among these convenient ones. In general the next smaller convenient class interval should be chosen, that is, the one next below the theoretically optimal interval. If the formula gives 9, 10 may be chosen, but if the formula indicates 7 or 8, the one actually used should generally be the next lower convenient class interval, 5.

CASE II. COMPUTATIONS INVOLVING TWO SERIES

In computing the coefficient of correlation, if the approximate value, r, of the index is known in advance, the optimal class interval to be used in each series may be estimated from the formula

$$C = \frac{R}{1 + (1.661 + 1.661r) \log N}$$

For example, in many biological correlations the approximate value of r is known to be 0.50, which results in 2.491, or approximately 2.5 as the coefficient of $\log N$ in the formula.

In case there is no advance knowledge of the value of r it should be first assumed to be 0, which gives the largest class intervals and the smallest number of classes, in other words the easiest computation. If the value of r derived from the preliminary computation is much different from 0, it may be used in a more accurate second determination of the class intervals in the series concerned, and r may then be calculated with the indicated classification of the series.

For other cases involving two or more series, working formulas may be derived for estimating the optimal theoretical size of the class interval.

HERBERT A. STURGES

Washburn College

COMMENT ON "A FORMULA FOR FREIGHT RATES"

The note on a formula for freight rates in this JOURNAL for September, page 416, is of interest in suggesting that a hyperbola fits existing freight schedules better than a parabola, but in some respects the note is misleading. A table of ton-mile costs for varying distances is quoted from the Nimmo Internal Commerce Report of 1876 as showing that "haulage costs do not increase at a fixed ratio, but the ratio of increase becomes smaller as mileage increases." The figures quoted show a constant ratio. They are derived from the assumption of a constant haulage factor of 9 mills per ton-mile combined with a fixed terminal cost of 31.62 cents per ton. Thus for 10 miles this gives a total rate of 40.62 cents, or 4.062 cents per ton-mile, as in the table quoted. For 100 miles, the total rate is $1.2162 and the rate per ton-mile is 1.216 cents, as in the table; and for 1,000 miles the total rate is $9.3162, and per ton-mile 0.93162 of a cent, rounded off at 0.932 in the table. In short, when we wish to construct rate scales on the basis of the cost of the service, we should use the formula for a straight line. There is no evidence that, after excluding from consideration the terminal costs and other factors wholly independent of distance, the road haul cost is any less for the fourth, fifth, or sixth hundred miles than it is for the first or second. It has been suggested that the shorter hauls are more expensive because the lightly loaded local trains are relatively more numerous for such distances. But this is not a satisfactory basis for saying that road haul costs decline with distance.

The true basis for a diminishing rate of progression for the road haul factor in the freight charge is in commercial considerations of what the traffic will bear. A lower profit per ton on the long haul widens the area of competition. The real difficulty in constructing rate scales is not in finding a curve that will best fit a given number of rates, but it is in determining what the key rates should be.

M. O. LORENZ

Washington, D. C.

PROGRESS OF WORK IN THE CENSUS BUREAU

A QUICK AND INEXPENSIVE CITY CENSUS

Under the supervision of Edward W. Koch of the Bureau of the Census, the city of Louisville, Kentucky, has recently taken a census through the agency of volunteer enumerators, working without compensation.

The preliminary work was begun December 1, 1925, with the aid of a very active local committee composed of the Mayor, and representatives of the Board of Trade, Bar Association, Real Estate Board, Publicity League, and various Luncheon Clubs. The first appeal for volunteer workers was made at