Aesthetic Frequency Classifications
Author(s): David P. Doane
Source: *The American Statistician*, Vol. 30, No. 4 (Nov., 1976), pp. 181-183
Published by: Taylor & Francis, Ltd. on behalf of the American Statistical Association
Stable URL: http://www.jstor.org/stable/2683757
Accessed: 15-04-2017 14:49 UTC

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.

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

American Statistical Association, Taylor & Francis, Ltd. are collaborating with JSTOR to digitize, preserve and extend access to *The American Statistician*

This content downloaded from 136.206.5.34 on Sat, 15 Apr 2017 14:49:02 UTC
All use subject to http://about.jstor.org/terms

Statistical Computing

This Department will carry articles of high quality on all aspects of computation in statistics.
Papers describing new algorithms, programs, or statistical packages will not contain listings of the program, although the completely documented program must be available from the author. Review of the paper will always include a running test of the program by the referee.
The description of a program or package in this Department should not be construed as an endorsement of it by the American Statistical Association or its Committees, nor is any warranty implied about the validity of the program.
The Editorial Committee will be pleased to confer with authors about the appropriateness of topics or drafts of possible articles.

To Our Readers:

In view of widespread interest in statistical computing, *The American Statistician* will expand its coverage in this area. In addition to nontechnical articles about statistical computing for the general readership, we plan to publish:

1. Summaries of selected committee reports dealing with statistical computing.
2. Announcements of new program packages and updates.
3. Brief notices of sources for further information.
4. Announcements and selected reviews of new computing products which may be of assistance to statisticians.

Authors, producers, or distributors wishing to have such materials announced or reviewed are invited to submit information according to the following guidelines:

(1) *Reports of Committees:* Summaries of reports of ASA committees dealing with statistical computing will be considered for publication if they are of widespread interest to the readers of *The American Statistician*. ["Criteria and Considerations in the Evaluation of Statistical Program Packages," or committee evaluations of statistical computing software and hardware.]

(2) *Announcements and Selected Reviews of New Program Packages or Updates:* Complete information concerning availability, costs involved, language, transportability, special features, and where the software has been used must be provided. The authors must state their intent that the material will be available to all requesters for a minimum of a two-year period. The editorial committee reserves the right to request running tests of the programs. [SPSS version X, BMDP revision, or "An Interactive Forecasting System".]

(3) *Notices of Sources for Further Information:* Brief references to other sources of information which may be of interest to readers. [ACM's SIG/SIC bulletins, and TOMS, IMSL newsletter.]

(4) *Announcements and Selected Reviews of New Products:* Complete information on such new items as hand calculators, programmable desk calculators, or minicomputers, particularly devices which extend statistical analysis capability, are of interest. Information on availability, costs, and special features should be provided.

Information on committee reports, new program packages or updates, sources for further information, and new products should be submitted to Associate Editor Thomas J. Boardman, Department of Statistics, Colorado State University, Fort Collins, Colorado 80523. Articles on statistical computing should continue to be sent to the Editor.

Editor

Aesthetic Frequency Classifications

DAVID P. DOANE*

I. Introduction

How do you teach a computer to look at a set of sample observations on one variable and make a frequency classification with the "right" number of classes, "nice" class limits, and "round" interval widths? No problem would seem more elementary, yet the advice proffered by statistics textbooks is vague, with liberal use of the terms "judgment" and "common sense." A computer algorithm which can aesthetically classify any univariate data set must make the rules more specific. Statisticians and others should find food for thought in the approach suggested here.

II. How Many Classes?

In his original brief exposition (total length: three paragraphs), Herbert Sturges [1] proposed a simple rule for classifying a series of N items. The expectation is that a normally distributed variable can be appropriately divided so that the class frequencies comprise a binomial series for all N which are even powers of 2. To use Sturges' own example, 16 observations will be divided into 5 classes with frequencies of 1, 4, 6, 4, 1. The optimal number of classes, in general, is $K = 1 + \log_2(N)$. Sturges' unambiguous rule has become a guideline for researchers, even where it is inappropriate.

Sample data are seldom symmetric, let alone normally distributed. Sturges' Rule does not always provide enough classes to reveal the shape of a severely skewed distribution. At a minimum, the real-world researcher would want to modify Sturges' Rule to reflect skewness. The statistic

$$\sqrt{b_1} = \frac{\Sigma(X - \bar{X})^3}{[\Sigma(X - \bar{X})^2]^{3/2}}$$

is a well-known measure of departure from the symmetric normal distribution. Since Sturges' formula provides for translating continuous, symmetric, normal data into discrete, symmetric, binomial classes, it is appropriate to use $\sqrt{b_1}$ to modify his rule. If $\sqrt{b_1}$ for a particular sample is more than so-and-so many

* School of Econ. & Management, Oakland Univ., Rochester, MI 48063.

This content downloaded from 136.206.5.34 on Sat, 15 Apr 2017 14:49:02 UTC
All use subject to http://about.jstor.org/terms

standard deviations away from zero, one would wish to reject the hypothesis upon which Sturges' Rule is predicated. The standard deviation of $\sqrt{b_1}$ depends only upon sample size [2], becoming smaller as sample size increases:

$$\sigma\sqrt{b_1} = \sqrt{\frac{6(N-2)}{(N+1)(N+3)}}$$

The proposed rule for adding extra classes is

$$K_e = \log_2\left(1 + \frac{\sqrt{b_1}}{\sigma\sqrt{b_1}}\right).$$

If $\sqrt{b_1} = 0$, no extra classes are added. As departure from the symmetric normal distribution becomes more obvious, classes are added, but at a decreasing rate.

This formula is motivated by the theory of information coding. The entropy of a message (in bits) is given by the formula $-\log_2(1/M)$ where M is the number of different equiprobable symbols which may occur. [3] As more symbols are used to encode a message, entropy increases at a decreasing rate. A frequency classification is appropriately viewed as a mapping from the real line to a finite set of symbols. Therefore, classification is a coding process subject to diminishing returns. The proposed modification of Sturges' Rule is consistent with this view.

III. Formal Problem Definition

Assume that there are no open-ended classes, that all classes must be of equal width, and that adjacent classes have common limits. With these simplifications, the problems may be completely expressed as finding an optimal trio (K^*, S^*, L^*) where K^* is the number of classes, S^* is the width of each class, and L^* is the lower limit of the first class. Every other class limit may then be determined. For a given K, it is proposed that the best choices S^* and L^* will satisfy the following conditions:

1. S^* is the roundest number such that $0 < K{\cdot}S^* - R \leqq S^*$ (R = observed data range)
2. L^* is the roundest number such that
   a. $L^* + K{\cdot}S^* > H$ (H = highest observed value)
   b. $0 \leqq L - L^* \leqq S^*$ (L = lowest observed value)

The first condition merely states that the effective range $K{\cdot}S^*$ must cover the true range, but may not exceed it by more than one interval width. Excessive waste is non-functional, and hence unaesthetic. The conditions on L^* require (a) that the range be centered so as to include the highest observed value, and (b) that waste in the first interval not exceed one interval width (implying the same for the last interval, given the condition on S^*). It is assumed that S^* is chosen first. It remains to specify the meaning of "roundest."

IV. Roundness

A simple hierarchy of roundness is proposed, applicable to all imaginable situations. Two definitions will suffice:

DEFINITION 1: A *very round number* X has the form $X = A{\cdot}10^B$ where $A \in \{5, 2, 1\}$ and B is any integer.

DEFINITION 2: The *roundness* of a number Y, denoted $R(Y)$, is the greatest very round exact divisor of Y.

The first definition establishes the infinite hierarchy

$$\{. . . , 1000, 500, 200, 100, 50, 20, 10, 5, 2, 1, .5, .2, .1, . . .\}.$$

The second definition assigns a roundness to every number except irrationals or repeating decimals, since any other number has an exact divisor in the hierarchy of very round numbers. These definitions correspond to commonsense notions about what roundness should mean. For example, we might say that 700 is "round to the hundreds" while 695 is only "round to the fives" (implying a relatively large difference in degree of aesthetic desirability for labeling a frequency classification).

V. The Algorithm

The preceding rules permit selection of an interval size S^* and a lower limit L^* for any *given* K. By departing slightly from the ideal K, we might obtain significant improvements in the roundness of S^* and L^*, thus increasing aesthetic satisfaction. Researchers do this intuitively. Arbitrarily, the computer is instructed to find S^* and L^* for integer K's in the range $K_{\text{ideal}} \pm \hat{K}$, where $\hat{K}$ is the nearest integer to $\log_2(K_{\text{ideal}})$. Then we must choose the best combination of (K, S^*, L^*) where several K values are considered.

From here on, the task is eclectic. One possibility is to have the computer print histograms for all the K values, and let the researcher decide which has the best class limits. But an algorithm using the following priority rules works very well:

1. Choose the K which produces the best (S^*, L^*) combination in terms of *relative* roundness of S^* and L^*:
   a. $R(S^*) \leqq 2\,R(L^*)$ — Most Aesthetic
   b. $R(S^*) = 2.5\,R(L^*)$ — ↕
   c. $R(S^*) > 2.5\,R(L^*)$ — Least Aesthetic
2. If the first rule produces no unique choice, choose the K value which (in order of priority):
   a. maximizes $R(S^*)$
   b. maximizes $R(L^*)$
   c. minimizes the distance from K to the ideal K (modified Sturges' Rule)

This content downloaded from 136.206.5.34 on Sat, 15 Apr 2017 14:49:02 UTC
All use subject to http://about.jstor.org/terms

The first priority rule says that if S^* is much rounder than L^*, the result will be unaesthetic *regardless* of the roundness of either one. The reader may construct examples to illustrate this point using the hierarchy of round numbers, and to verify the relevance of the particular ratios indicated. The priorities in the second rule are applied sequentially, as necessary, to narrow the choices until a unique K is determined.

VI. Testing The Algorithm

An algorithm using these suggested rules has proved effective in classifying data from statistics textbooks in the same way the authors of the books have done in their illustrative examples. Textbook examples provide little challenge, because they tend to use symmetric, bell-shaped data. Despite a few "blind spots," the algorithm has performed well in random simulations, and using "strange" data sets submitted by colleagues. Running time is modest (e.g., 7.5 CPU seconds to sort, classify, and print a histogram of 500 observations on a Burroughs 5700). It will work on any data set, measured in millions or millionths, positive or negative. Relaxing the rule on waste (increasing the permissible waste from one to two interval widths) has been found to help somewhat in producing more aesthetic results, but it requires significantly more complicated and eclectic procedures for centering the effective range so as to prevent empty classes. Consequently, the rules stated earlier are preferred, by Occam's Razor.

Examples are not given, because one or two examples would not prove anything about the algorithm's effectiveness. Instead, readers are invited to write for copies of the full computer program (well-documented) and to devise their own challenges and suggestions for improvement. The intent of this paper has been to communicate with those who have uses for such an algorithm, and to raise a few questions. Hopefully, these purposes have been achieved.

REFERENCES

1. Sturges, Herbert A.: "The Choice of a Class Interval," *Journal of the American Statistical Association* (March 1926), p. 65.
2. Pearson, E.S.: "Note on Probability Levels for $\sqrt{b_1}$," *Biometrika,* Vol. 28 (1936), p. 306.
3. Pierce, J. R.: *Symbols, Signals, and Noise: The Nature and Process of Communication* (Harper and Brothers, 1961), pp. 80–86.

Referees

The Editor and Associate Editors are greatly indebted to the following persons for serving as referees for *The American Statistician* through August 19, 1976:

P. O. Anderson
B. C. Arnold
L. A. Aroian
S. K. Badhe
B. Bailar
D. R. Barr
D. Basu
E. Battiste
V. P. Bhapkar
F. L. Bookstein
J. M. Boyett
J. V. Bradley
L. H. Broekhoven
G. R. Bryce
L. F. Burmeister
J. M. Cameron
J. Carlson
B. Causey
V. Chew
B. Chow
W. R. Clarke
W. J. Conover
J. A. Cornell
J. M. Davenport
W. E. Deming
W. J. Dixon
E. Dudewicz
O. Dykstra
K. R. Eberhardt
M. E. Engelhardt
E. Ericksen
J. J. Filliben
S. J. Finch
M. A. Fligner
D. H. Freeman, Jr.
C. E. Fuchs
D. J. Gans
S. Geisser
T. M. Gerig
B. K. Ghosh
L. J. Gleser
A. S. Goldberger
J. H. Goodnight
Z. Govindarajulu
C. W. J. Granger
F. A. Graybill
W. C. Gregory
R. A. Groeneveld
W. Guenther
R. F. Gunst
D. Guthrie
R. Haas
S. J. Haberman
M. Halperin
M. J. Hartley
W. J. Hemmerle
J. Hewett
J. C. Hickman
K. Hinkelmann
D. K. Hildebrand
D. Hoaglin
R. R. Hocking
D. G. Horvitz
R. W. Hoyer
D. Jobson
D. H. Jones
J. D. Kalbfleisch
S. K. Katti
T. Kelley
C. G. Khatri
T. J. Killeen
S. Klugman
G. G. Koch
H. C. Kraemer
W. H. Kruskal
R. J. Kryscio
H. H. Ku
M. Kutner
J. LaBrecque
P. A. Lachenbruch
S. W. Lagakos
R. G. Laha
J. C. Lee
M. Lentner
F. C. Leone
K. J. Levy
R. F. Ling
R. C. Littell
S. Litwin
M. O. Locks
J. T. McClave
R. B. McKennan
J. E. McLean
D. R. McNeil
G. S. Maddala
C. R. Mann
N. Mantel
B. H. Margolin
J. D. Mason
D. M. Maxwell
L. S. Mayer
G. Mayeske
M. F. Miller
G. A. Milliken
A. H. Moore
W. G. Nichols
G. E. Noether
R. T. O'Neill
L. Ott
W. R. Pabst
T. Papaioannou
C. G. Pfeifer
D. A. Pierce
W. Pirie
G. Pledger
G. E. Policello
A. A. Prato
R. H. Randles
M. R. Reynolds
J. R. Rosenblatt
B. Sande, Jr.
P. Schmidt
M. A. Schneiderman
S. R. Searle
D. G. Seigel
A. W. Sherdon
J. Shuster
M.M. Siddiqui
D. O. Siegmund
G. L. Sievers
B. G. Simon
R. D. Snee
J. D. Spurrier
J. H. Stapleton
W. E. Strawderman
S. Sudman
B. V. Sukhatme
P. Tadikamalla
V. S. Taneja
J. E. Triplett
P. V. Tryon
N. S. Urquhart
H. K. Ury
P. Velleman
W. A. Wallis
R. H. Wampler
S. L. Warner
E. J. Wegman
J. Wilkinson
R. F. White
K. M. Wolter
W. M. Wooding
G. L. Yang
M. Zelen
J. zum Brunnen

This content downloaded from 136.206.5.34 on Sat, 15 Apr 2017 14:49:02 UTC
All use subject to http://about.jstor.org/terms