Volume: 07 Number: 02 Category: 04 Page: 77-86 ISSN No: 1303-6319 Cilt: 07 Sayı: 02 Kategori: 04 Sayfa: 77-86

DETERMINATION OF THE NUMBER OF BINS/CLASSES USED IN HISTOGRAMS AND FREQUENCY TABLES: A SHORT BIBLIOGRAPHY

Nurhan DOĞAN* **İsmet DOĞAN****

ABSTRACT

The histogram is the oldest and most popular tool for graphical display of a univariate set of data. An important parameter that needs to be specified when constructing a histogram is the bin width (also called the interval width or class width). This is simply the length of the subintervals of the real line, sometimes called "bins" or "intervals" (also called "classes"), on which the histogram is based. Frequency distributions facilitate the organization and presentation of data. A major issue with all classifying techniques is how to select the number of classes. There is no "correct" answer for every set of data. Each case must be treated separately; each frequency table must be designed individually. The number of bins / classes increases as the sample size increases. Larson's (1975) formula had the lowest number of bins / classes and the Ishikawa's (1986) formula had the highest one for n>300. The aim of this study is to give a short bibliography for bins/class numbers.

Keywords: Bin width, Class number, Frequency tables, Histogram.

1. INTRODUCTION

Statistics is the science of assembling, classifying, tabulating and analyzing numerical facts or data. A major issue with all tabulating and classifying techniques is how to select the number of classes. Questions are frequently asked about how many classes are useful and what size they should be. There are many criteria and guidelines for approaching the problem. Unfortunately, no standart, objective selection procedure exists. There is no "correct" answer for every set of data. Each case must be treated separately; each frequency table must be designed individually. More detail can be shown in larger number of classes; however, if the number of classes is too large, the classification loses its effectiveness as a means of summarizing data. On the other hand, if the number of classes is too small, the data will be condensed so much that little or no insight can be gained into the nature of the pattern of variation (Plane and Oppermann, 1981).

Frequency distributions facilitate organization and presentation of data. It is generally recognized that the process of tabulating data in classes may give rise to a considerable degree of distortion of the original data. Even when the class limits are chosen with the utmost care, so as to secure an adequate scatter in each class, the measures of central tendency, dispersion and the type of distribution may be none too dependable. Nevertheless, frequency classifications are indispensable, both as means of summing up a large array of data compactly and as a form of generalization.

* Assist.Prof., Afyon Kocatepe University, Medical Faculty, Biostatistics Department, Afyonkarahisar, Turkey, e-mail: ndogan@aku.edu.tr
** Prof., Afyon Kocatepe University, Medical Faculty, Biostatistics Department, Afyonkarahisar, Turkey, e-mail: idogan@aku.edu.tr

Very often it happens that the statistician must work with tabulations made up from original sources which he cannot conveniently consult. Hence, the analysis of frequency distributions is likely to remain one of the important branches of statistics (Davies, 1929).

The composition of a numerical random sample is conveniently pictured by its histogram. A histogram conveys visual information of both the frequency and relative frequencies of observations; that is the essence of a density function (Scott, 1992). For many classes of data one expects the underlying population to be approximately normal and hence the histogram of the sample also to be approximately normal. If so, it may be further convenient to smooth the histogram by approximating it by a suitable normal density curve (Brown and Hwang, 1993).

The framework of the classical histogram is useful for conveying the general flavor of nonparametric theory and practise. The histogram is most often displayed on a nondensity scale: either as bin counts or as a stem-and-leaf plot. The classical frequency histogram is formed by constructing a complete set of nonoverlapping intervals, called bins, and counting the number of points in each bin. In order for the bin counts to be comparable, the bins should all have the same width (Scott, 1992).

The histogram is the classical nonparametric density estimator, probably dating from the mortality studies of John Graunt in 1662. Today the histogram remains an important statistical tool for displaying and summarizing data. In addition, it provides a consistent estimate of the true underlying probability density function (Scott, 1979).

The histogram is the oldest and most popular tool for graphical display of a univariate set of data. It is taught in virtually all elementary data analysis courses and available in most statistical computing packages. An important parameter that needs to be specified when constructing a histogram is the bin width (also called the interval width or class width). This is simply the length of the subintervals of the real line, sometimes called “bins” or “intervals” (also called “classes”), on which the histogram is based. It is not very difficult to see that the choice of the bin width has an enormous effect on the appearence of the resulting histogram. The choice of a very small bin width results in a jagged histogram, with a separate block for each distinct observation. On the other hand, a very large bin width results in a histogram with a single block. Intermediate bin widths lead to a variety of histogram shapes between these two extremes. Ideally, the bin width should be chosen so that the histogram displays the essential structure of the data, without giving too much credence to the data set at hand. Essentially it amounts to choosing the bin width (h);

$$h = \frac{range\ of\ data}{number\ of\ class}$$ (Wand, 1997).

The choice of bin width is directly related to the number of classes. It is best if all bins are of the same width. However, in certain cases, unequal bin width must be used (Plane and Oppermann, 1981).

The histogram is a statistical technique with a long history. Unfortunately, there exist only a few explicit guidelines, which are based on statistical theory, for choosing the number of bins that appear in the histogram (He and Meeden, 1991).

Since the well-known formula by Sturges (1926), many authors have proposed rules for choosing the bin-width (class width) or the number of bins. These rules have always been regarding the histogram as an estimator of an underlying density and relying on large-n asymptotics. Histograms constructed according to these rules have been compared with other density estimators, notably kernel and spline estimators (Hirai, 1990; Denby and Mallows, 2009).

To construct a frequency table or a histogram for grouped data, most introductory statistics books agree on determining the four crucial elements. These four elements are: The range, the number of class intervals, the class interval width and the starting point (Lohaka, 2007).

How many classes? There are too many numbers being recommended for the number of classes. To start with, a number of textbooks recommended as a general rule that 10 classes to be taken as optimal and 30 as the maximum. A few others urged between 10 and 20 classes. According to some other authors, the number of classes can be situated either between 10 and 14, or 5 and 15. Besides these suggestions, some formulas have been developed in choosing an appropriate number of classes, which is denoted by the letter *k*.

Table 1. Formulas for the number of bins / classes

Order	*k*	Reference	Order	*k*	Reference
1	$1+\left[3.3\times\log_{10}(n)\right]$	Sturges, 1926	13	Rudemo suggested a cross validation technique for selecting the number of bins.	Rudemo, 1982
2	$4\times\sqrt[5]{2\times(n-1)^2/c^2}$ $c=\frac{1}{\sqrt{2\pi}}\int_c^{\infty}e^{-\frac{1}{2}t^2dt}$	Mann and Wald, 1942	14	Gislason and Goldfield proposed in their study a simple method for determining the optimum number of bins to be used in a histogrammic representation of a function of a continuous variable.	Gislason and Goldfield, 1984
3	$\sqrt{n/5}$	Cohran, 1954	15	Suzuki used a rounded number for the class interval and also the endpoint of a class.	Suzuki, 1985
4	$\sqrt[3]{n}$	Cencov, 1962	16	$\sqrt[3]{2n}$	Terrel and Scott, 1985
5	$1.87\times(n-1)^{0.4}$	Bendat and Piersol, 1966	17	$6+(n/50)$	Ishikawa, 1986
6	Mori proposed a method by minimizing the mean squared error of a histogram estimate $\hat{f}_n$ of the true density $f(x)$.	Mori, 1974	18	Using minimum description lenght principle, Rissanen estimated the density.	Rissanen, 1992
7	$1+\left[2.2\times\log_{10}(n)\right]$	Larson, 1975	19	In their note, they gave a decision theoretic approach (using loss function and stepwise Bayes rule based on the Bayesian bootstrap) to the problem of choosing the number of bins in a histogram.	He and Meeden, 1997
8	$1+\log_2 n+\log_2\gamma$ $\gamma=\sqrt{(n+1)\times(n+3)\times\sum_{i=1}^{n}(x-\bar{x})^2}/\sqrt{6\times(n-2)\times\sum_{i=1}^{n}(x-\bar{x})^2}$	Doane, 1976	20	Using Bayesian probability theory, Knuth derived a straightforward algorithm that computed the posterior probability of the number of bins for a given data set.	Knuth, 2006
9	$2\times\sqrt{n}\quad for\ \ n\le 100$ $10\times\log_{10} n\quad for\ \ n\ge 100$	Velleman, 1976	21*	$2\times\sqrt[3]{n}$	Rice
10	$\sqrt{n}$	Mosteller and Tukey, 1977	22*	$2.5\times\sqrt[4]{n}$	Anonymous1
11	$\left[\left(range\ \ of\ \ data\times\sqrt[3]{n}\right)/(3.49\times\sigma)\right]$ $\sigma=s\tan dart\ \ deviation\ \ of\ \ sample$	Scott, 1979	23*	$2^k\ge n$ *k* is the smallest integer number of classes and *n* being the number of observations	Anonymous2
12	$\left[range\ \ of\ \ data/2\times(IQ)\times n^{-1/3}\right]$ $IQ(Interquartile)=Q_3-Q_1$	Freedman and Diaconis, 1981			

* Directly taken from Lohaka, 2007.

Sturges proposed a simple rule for classifying a series of n item. Sturges' rule probably survived as long as it did because, for moderate *n* (less than 200), it gives similar results to the alternative rules, and so produces reasonable histograms. However, it does not work for large *n* (Hyndman, 1995).

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. Sturges' unambiguous rule has become a guideline for researchers, even when 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 realworld researcher would want to modify Sturges' Rule to reflect skewness. The statistic;

$$\sqrt{b_1} = \frac{\sum (x - \bar{x})^2}{\left[\sum (x - \bar{x})^2\right]^{3/2}}$$

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

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

The proposed rule for adding extra classes is given below:

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

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 (Doane, 1976).

2. MATERIALS AND METHODS

In this article, 14 formulas are used to compute the number of bins / classes taken from Table 1. Data were generated to simulate a number of experimental scenarios. The simulation results were analysed by using table representations. All computations were performed with Excel software for Windows.

3. RESULTS AND DISCUSSION

After determining the range of scores and distributions, the main question before a set of data can be converted into a grouped-data frequency table is determining the number of classes. The number of classes to be used is primarily not only dependent on the number of observations in the data set, but also on the range of observed scores. That is, larger numbers of observations require a larger number of class groups. In general, however, the frequency distribution should have at least five class groupings, but no more than 15. If there are not enough class groupings or if there are too many, little information would be obtained (Berenson and Levine, 1992).

Table 2 and Table 3 present the results that are obtained with the Scott's (1979) formula and Freedman and Diaconis's (1981) formula, respectively. According to the results of both Scott's formula and Freedman and Diaconis's formula, the number of classes increase rapidly as both range and standart deviation / interquartile of data goes up, particularly when *n* is in thousands.

Table 2. General results of Scott's (1979) formula

n	Range	Standart Deviation	Number of Bins/ Classes
Increase	Increase	Increase	Increase
Increase	Increase	Decrease	Increase
Increase	Decrease	Increase	Decrease
Increase	Decrease	Decrease	Decrease
Decrease	Increase	Increase	Decrease
Decrease	Increase	Decrease	Decrease
Decrease	Decrease	Increase	Decrease
Decrease	Decrease	Decrease	Decrease

Table 3. General results of Freedman and Diaconis' (1981) formula

n	Range	Interquartile	Number of Bins/ Classes
Increase	Increase	Increase	Increase
Increase	Increase	Decrease	Decrease
Increase	Decrease	Increase	Decrease
Increase	Decrease	Decrease	Decrease
Decrease	Increase	Increase	Increase
Decrease	Increase	Decrease	Increase
Decrease	Decrease	Increase	Increase
Decrease	Decrease	Decrease	Decrease

Table 4 lists a few numbers of observed values, *n*, for each of the 12 formulas proposed to compute an appropriate or suitable number of classes, *k*. It can be observed that there is a great diversity of *k* values obtained.

Table 4. Value of *k* for selected *n* numbers

n	Sturges	Cohran	Cencov	Bendat and Piersol	Larson	Velleman	Mosteller and Tukey	Terrel and Scott	Ishikawa	Rice	Anonymous1	Anonymous2
10	5	2	3	5	4	7	4	3	7	5	5	4
20	6	2	3	7	4	9	5	4	7	6	6	5
30	6	3	4	8	5	11	6	4	7	7	6	5
40	7	3	4	9	5	13	7	5	7	7	7	6
50	7	4	4	9	5	15	8	5	7	8	7	6
60	7	4	4	10	5	16	8	5	8	8	7	6
70	8	4	5	11	6	17	9	6	8	9	8	7
80	8	4	5	11	6	18	9	6	8	9	8	7
90	8	5	5	12	6	19	10	6	8	9	8	7
100	8	5	5	12	6	20	10	6	8	10	8	7
125	8	5	5	13	6	21	12	7	9	10	9	7
150	9	6	6	14	6	22	13	7	9	11	9	8
175	9	6	6	15	6	23	14	8	10	12	10	8
200	9	7	6	16	7	24	15	8	10	12	10	8
250	9	8	7	17	7	24	16	8	11	13	10	8
300	10	8	7	19	7	25	18	9	12	14	11	9
350	10	9	8	20	7	26	19	9	13	15	11	9
400	10	9	8	21	7	27	20	10	14	15	12	9
450	10	10	8	22	7	27	22	10	15	16	12	9
500	10	10	8	23	7	27	23	10	16	16	12	9
1000	11	15	10	30	8	30	32	13	26	20	15	10
1500	12	18	12	35	8	32	39	15	36	23	16	11
2000	12	20	13	40	9	34	45	16	46	26	17	11
2500	13	23	14	43	9	34	50	18	56	28	18	12
5000	14	32	18	57	10	37	71	22	106	35	22	13
10000	15	45	22	75	10	40	100	28	206	44	25	14
20000	16	64	28	99	11	44	142	35	406	55	30	15
40000	17	90	35	130	12	47	200	44	806	69	36	16
50000	17	100	37	142	12	47	224	47	1006	74	38	16
100000	18	142	47	187	12	50	317	59	2006	93	45	17

When setting up a grouped-data frequency table, it makes a big difference how many classes are used. The choice of the number of class is quite arbitrary. It is even highly subjective as it is a matter of personel judgement. There exist guidelines to help the researchers with this, but they remain vague in that they are subject to personel interpretations, tastes and preferences.

It is important to remember that frequency distribution tables are employed to reveal or emphasize a group pattern. Either too many or too few classes may blur that pattern and thereby work against the researcher who seeks to add clarity to the analysis. In sum, the researcher generally makes a decision as to the number of classes based on the set of data and personal objectives, factors that may vary considerably from one research situation to another (Lohaka, 2007).

A considerable amount of work has been done on how to set up the best histogram. However, there is no agreement as to which method should work best (Gislason and Goldfield, 1984).

4. REFERENCES

Bendat, S. J., Piersol, A.G., 1966. Measurements and Analysis of Random Data. John Wiley & SONS, Inc., New York.

Berenson, M. L., Levine, D.M., 1992. Basic Business Statistics: Concepts and Applications. Prentice Hall Englewood Cliffs, New Jersey.

Brown, L.D., Hwang, J.T., 1993. How to Approximate a Histogram by a Normal Density. The American Statistician, 47 (4), 251-255.

Cencov, N.N., 1962. Evaluation of an Unknown Distribution Density From Observations. Soviet Mathematics, 3, 1559-1562.

Cohran, W.G., 1954. Some Methods for Strengthening the Common Chi Square Test. Biometrics, 10 (4), 417-451.

Denby, L., Mallows, C., 2009. Variations on the Histogram. Journal of Computational and Graphical Statistics, 18(1), 21-31.

Davies, G.R., 1929. The Analysis of Frequency Distributions. Journal of the American Statistical Association, 24 (168), 349-366.

Doane, D.P., 1976. Aesthetic Frequency Classifications. The American Statistician, 30 (4), 181-183.

Freedman, D., Diaconis, P., 1981. On The Histogram as a Density Estimator: L_2 Theory. Zeitschrift fur Wahrscheinlichkeitstheorie und Verwandte Gebiete, 57, 453-476.

Gislason, E. A., Goldfield, E. M., 1984. Determination of the Optimum Number of Bins to Use in a Histogrammic Representation of a Probability Density Function. Journal of Chemical Physics, 80(2), 701-704.

He, K., Meeden, G., 1997. Selecting the Number of Bins in a Histogram: A Decision Theoretic Approach. Journal of Statistical Planning and Inference, 61, 59-69.

Hirai, Y., Some Remarks on Class Interval of Histograms, 2009. http://eprints.lib.okayama-u.ac.jp/ 9696/1/082_0113_0117.pdf.

Hyndman, R. J., The Problem with Sturges' Rule for Constructing Histograms, 1995, Monash University, www.robjhyndman.com/papers/sturges.pdf.

Ishikawa, K., 1986. Guide to Quality Control. White Plains, New York: Unipub, Kraus International.

Knuth, K.H., Optimal Data-Based Binning for Histograms, Draft Paper, 2006. http://www.huginn.com/knuth/papers/knuth-histo-draft-060221.pdf.

Larson, H.J., 1975. Statistics: An Introduction. John Wiley & SONS, Inc., New York.

Lohaka, H.O., 2007. Making a Grouped Data Frequency Table: Development and Examination of the Iteration Algorithm. PhD Thesis, College of Education, Ohio University (unpublished).

Mann, H. B., Wald, A., 1942. On the Choice of the Number of Class Intervals in the Application of the Chi Square Test. The Annals of Mathematical Statistics, 13 (3), 306-317.

Mori, T., 1974. An Optimal Length of Class Interval for Histogram. Japan Journal of Applied Statistics, 4(1), 17-24.

Mosteller, F., Tukey, J.W., 1977. Data Analysis and Regression, A Second Course in Statistics. Addison-Wesley, Reading, MA.

Plane, D. R., Oppermann, E.B., 1981. Business and Economic Statistics. Business Publications, Inc., Plano, Texas.

Rissanen, J., 1992. Stochastic Complexity in Statistical Inquary. World Scientific, Singapore.

Rudemo, M., 1982. Empricial Choice of Histograms and Kernel Density Estimators. Scandinavian Journal of Statistics, 9 (2), 65-78.

Scott, D. W., 1979. On Optimal and Data-Based Histograms. Biometrika, 66(3), 605-610.

Scott, D. W., 1992. Multivariate Density Estimation: Theory, Practice and Visualization. John Wiley & SONS, Inc., New York.

Sturges, H.A., 1926. The Choice of a Class Interval. Journal of the American Statistical Association, 21 (153), 65-66.

Suzuki, G., 1985. Effective Use of Graphical Representation in Statistics. Japan Journal of Applied Statistics, 14(1), 27-37.

Terrel, G.R., Scott, D.W., 1985. Oversmoothed Nonparametric Density Estimates. Journal of the American Statistical Association, 80 (389), 209-214.

Velleman, P.F., 1976. Interactive Computing for Exploratory Data Analysis I: Display Algorihms. 1975 Proceedings of the Statistical Computing Section, 142-147, Washington DC: American Statistical Association.

Wand, M.P., 1997. Data-Based Choice of Histogram Bin Width. The American Statistician, 51 (1), 59-64.

HİSTOGRAMLARDA VE SIKLIK TABLOLARINDA KULLANILAN SÜTUN / SINIF SAYILARININ BELİRLENMESİ: KISA BİR BİBLİYOGRAFYA

ÖZET

Tek değişkenli bir veri setinin grafiksel gösterimi için en eski ve en popüler yöntem histogramdır. Bir histogramın oluşturulmasında belirlenmesi gereken önemli bir parametre ise aralık genişliği veya sınıf genişliği olarak da bilinen sütun genişliğidir. Histogramların esası olan sütun genişliği basit olarak sütun, aralık veya sınıf olarak isimlendirilen alt grupların uzunluğudur. Sıklık dağılımları ise verinin düzenlenmesi ve sunulmasında yardımcı olur. Sınıflandırma tekniklerinin hemen tamamında temel sorun sınıf sayısının nasıl seçileceğidir. Sınıf sayısı ile ilgili her veri seti için geçerli doğru bir cevap bulunmamaktadır. Her bir durum ayrı ayrı dikkate alınmalı, her bir sıklık tablosu kendine özgü olarak düzenlenmelidir. Örnek büyüklüğü arttıkça sütun / sınıf sayısı artmaktadır. n>300 olması durumunda en düşük sütun / sınıf sayısı Larson (1975) formülünden, en yüksek sütun / sınıf sayısı ise Ishikawa (1986) formülünden elde edilmekedir. Bu çalışmanın amacı, sütun / sınıf sayısının belirlenmesi ile ilgili kısa bir bibliyografya vermektir.

Anahtar Kelimeler: Sütun genişliği, Sınıf sayısı, Sıklık tablosu, Histogram.