Comparison of a Few Simple Prescriptions for the Selection of Number of Bins in a Univariate Histogram for Normally Distributed Samples Using Simulations

D Maibam[1]*, Y Sharma[2], A Khardewsaw[1] and A Saxena[1]

[1] Department of Physics, NEHU, Shillong, Meghalaya, India

[2] Department of Physics, Don Bosco College, Tura, Meghalaya, India

E-mail: deveshwori_maibam@outlook.com

Abstract

The mean squared error estimate is used to characterise the performance of 7 methods for determining the number of bins in a histogram, which serves to determine the underlying normal probability density function that generated the samples. The simpler and direct method of simulation is used to achieve the objectives of the present study. Random samples from normal distributions with 10 different values of standard deviation and 20 different sample sizes are used with 1000 runs for each combination. A fitting procedure is performed on the estimated mean squared error, which is then used to identify the prescription that gives the least error– determined to be the rule prescribed by Scott in the present study.

Keywords **– histogram, simulation, least-square fit, mean squared error**

Introduction

One of the simplest methods of estimating the underlying probability distribution of a given set of samples is the histogram[1, 2]. Although, the term histogram is usually associated with the visual representation similar to a bar-graph with touching adjacent sides, but, in the current context it was deemed useful to expand the definition to include the entire process of generating the processed data for the graph from the raw dataset. For the construction of

a histogram for a given dataset (assumed univariate in the present case), the total range of the data values is divided into non-overlapping intervals or 'bins' of equal size[1] and then for each interval the number of data points that lie in that interval is counted; with this data we can generate a frequency table with boundary values of the bins in one column and the corresponding frequencies (number of data points in that particular bin) in another column. The bar-graph plotted with each bar-width enclosing a bin and the bar-height equal to the corresponding frequency of data points is a histogram. Sometimes a few values are either too high or too low in comparison to the bulk of the dataset; such values are called outliers and may be either valid data points or the result of some gross errors in the measurements/calculations; in either case, the outliers are generally removed before the construction of the histogram and analysed separately. In the present case, it is assumed that the dataset has no outliers or that it has been removed. To obtain the probability density values, the frequency in each data bin is divided by the total number of data points, further divided by the bin-width.

[1] The conditions of non-overlapping and equal size for the bins is not a strict requirement but is generally used to simplify the construction of the histogram.

The basic purpose of the histogram is to show the shape of the generating frequency distribution over the given range of values; in its construction, the primary concern is the number of bins to be taken – too few bins could over-smooth the shape leaving out important details and too many bins could under-smooth the shape and generate spurious details. Over the years, there have been a handful of suggestions for determination of the optimal bin-width or alternatively number of bins – some *ad hoc* methods (see table 1) and some based on statistical optimisation procedures. In the present paper, we will analyse the performance of 7 methods of determining the number of bins for a given univariate dataset taken from normal distribution. The comparative effectiveness and the simplicity of the *ad hoc* methods for the estimation of the bin-widths or the number of bins is exemplified in their usage in popular software packages for statistical calculations like Microsoft Excel 2017[2], which uses the Scott's rule.

Methodology

The 7 methods for determining the number of bins or the bin-number for the construction of the histogram used in the present paper are listed in table 1.

[2] from online Microsoft Excel 2017 help at https://support.office.com/

The details about the individual methods may be obtained from the relevant references. A few of the tabulated methods calculate the bin-width (W) instead of the bin-number (N) – the relationship between the two is given by,

$$N = \frac{(range\ of\ data)}{W} \quad \ldots\ldots(1)$$

Table 1: List of the methods used for estimating the bin-number/ number for drawing histogram from data.

Method	Abbreviation used in the present work	Formula for the number(N)/width(W) of bins*
H. A. Sturges' formula [3]	Sturges	$N = 1 + \log_2(n)$ ..... (2)
D. P. Doane's formula [4]	Doane	$N = 1 + \log_2(n) + \log_2\left(1 + \sqrt{\frac{\sum(x-\mu)^3}{(n-1)\sigma^3}}\right)$ ... (3)
D. W. Scott's rule [5]	Scott	$W = \frac{3.49 \times \sigma}{\sqrt[3]{n}}$ ..... (4)
D. Freedman & P. Diaconis' formula [6]	Freedman	$W = \frac{2 \times IQR}{\sqrt[3]{n}}$ ..... (5)
K. H. Knuth's formula [7]	Knuth	Maximum of the logarithm of the posterior probability of the histogram model from data
F. Mosteller & J. W. Tukey's formula3[4] [8]	Mosteller	$N = \sqrt[2]{n}$ ..... (6)
Rice's rule [1]	Rice	$N = 2\sqrt[3]{n}$ ..... (7)

*n is the number of elements in the data set, μ is the mean and σ is the standard deviation, IQR is the Inter Quartile Range

Since, the bin-number has to be an integer, therefore the computed values are rounded off to the nearest integer. The current work tries to rank the performances based on the results of simulation. All calculations and simulations have been done in the MATLAB computing environment.

Since, most of the above methods have been formulated with the normal distribution in mind, we have therefore carried out the simulations only for normal distribution with different values of the standard deviation (σ); since, the mean (μ) plays no role in the shape of the distribution, the same value of mean was used throughout.

The simulations were performed

[4] also known as the square root choice

as follows 10 different normal distributions were considered with values ranging from 0.5 to 50 with almost exponential spacing viz. 0.5, 1, 2, 3, 5, 8, 13, 20, 32 and 50. For each of the normal distribution constructed, 20 different values of sample sizes ranging from 30 to 10000 were studied, again with almost exponential spacing between them. A certain number –the sample size – of pseudo-random numbers are then generated from the given normal distribution using MATLAB. 1000 simulations were done for each type. Using the various prescriptions, the histogram is constructed for each simulated sample and, from the data of the histogram, the probability distribution is computed. To estimate the difference of the computed probability distribution with the actual probability distribution function, we have used the mean squared error (MSE) as a measure of the difference. The MSE [2] is the mean value of the squared difference between the values of the constructed and actual probability densities at each evaluated point for each sample considered. The lower the MSE value the better is the agreement between the computed and actual values.

Results and Discussion

The mean value (of the 1000 simulation runs) of the probability density for each of the bin-number prescriptions along with standard deviation (shown as error-bars) is plotted in figure 1.It may be mentioned here that the prescriptions as given by Sturges, Mosteller and by Rice are purely functions of the sample size and hence the simulations do not affect their output i.e. bin-numbers prescribed, therefore, there are no error-bars in the corresponding plots in figure 1. Plots corresponding to different standard deviation (σ) values (of normal distribution for the same method are plotted in similar colour immediately adjacent to each other. It can be observed that the prescriptions of Mosteller and that of Freedman have relatively larger bin-numbers for higher sample sizes.

The mean values of the MSE score as a function of the sample size for different σ values are plotted in figure 2 as separate graphs for each prescription. We observe the expected pattern of higher MSE values at lower sample sizes and lower MSE values at higher sample sizes i.e. the MSE values decreases almost steadily with increasing sample size (n). This decrease in the MSE values with increasing n is found to nearly follow an inverse law i.e. the MSE values decreases as the reciprocal of n, therefore we decided to fit the MSE values from the various prescriptions to a function of the form:

$$\text{MSE} = \frac{1}{n}\, a \qquad \ldots\ldots (8)$$

where, '*a*' is a parameter to be determined from least square fitting.

The values of the fitting parameter '*a*' obtained after performing a least square fit (at 95% confidence bounds) between the MSE values and *n* are plotted in figure 3; the error-bar represents the lower and upper confidence bounds. The parameter '*a*' also shows a non-linear decrease with increasing values, similar to (1/σ) trend; this can be understood as the result of the difficulty in capturing sharper peaks of a normal distribution curve (characteristic of small σ values) with relatively few equally spaced histogram bins.

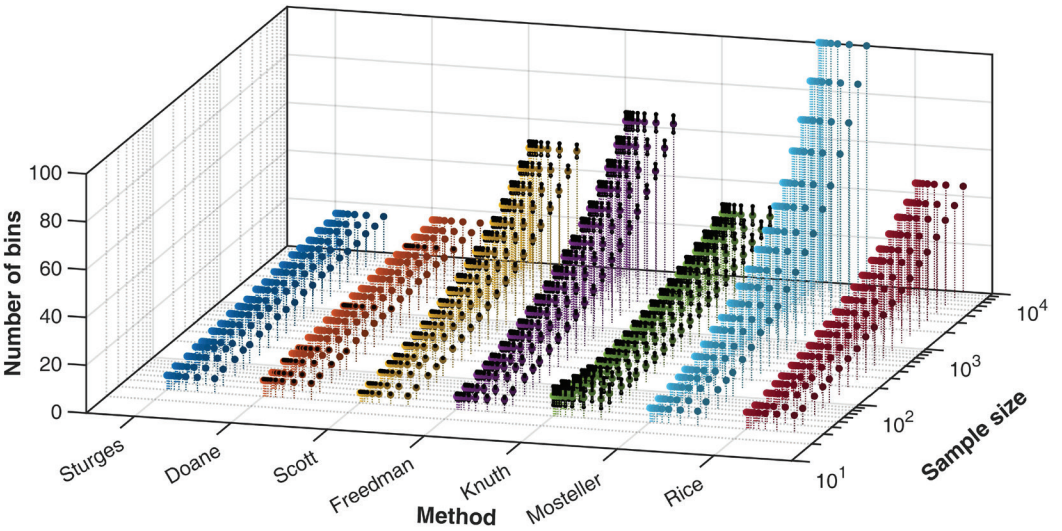

Figure 1: Plot of the number of bins of histogram for different sample sizes and methods of determining them. The name of the method employed is plotted in the x-axis and is self-evident. The 10 different values of σ are plotted in the same colour adjacently with their separation proportional to the value of σ .

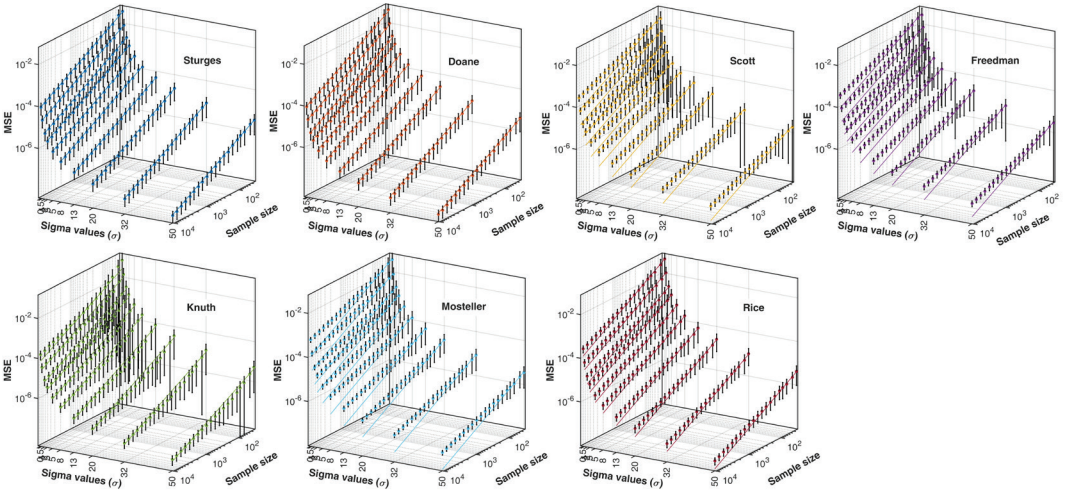

Figure 2: Plot of the MSE for different sample sizes and methods of determining them shown as coloured dots with error-bars (in black) representing standard deviation. The coloured solid lines are the best-fit lines according to equation (8).

The best-fit curve as obtained from the above fit is shown in figure 2 as a solid curve across the sample size. While noting the discrepancy between the actual data points and the best-fit curve in figure 2, we must remember that the graph has been logarithmically scaled.

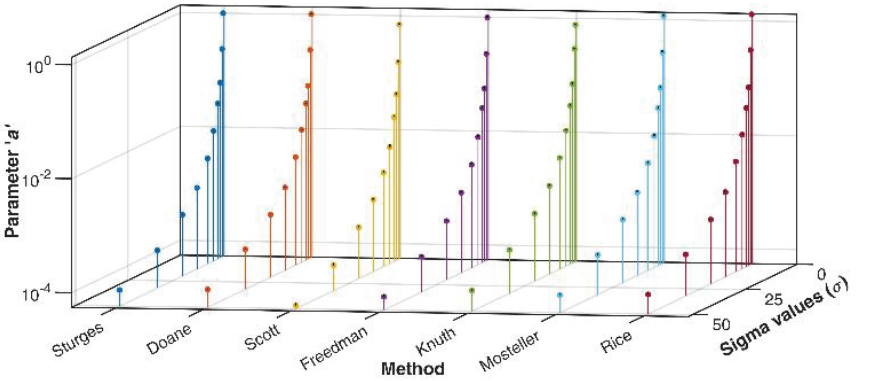

Figure 3: The least square fitting parameter 'a' for different methods at different values σ

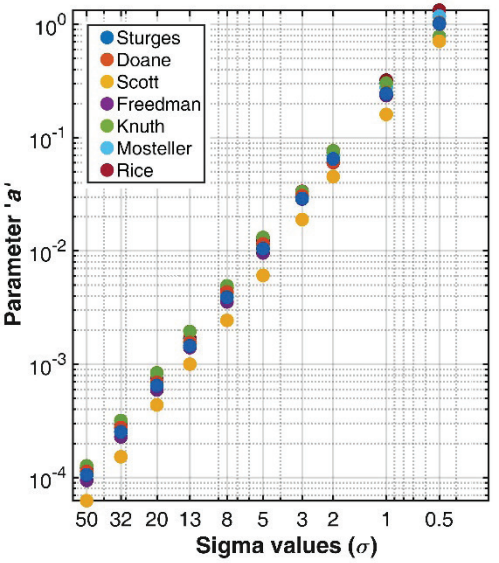

Figure 4: The values of the parameter 'a' for the different values of σ with colour coding for different methods employed in the estimation of the bin-number.

The figure 3 viewed as a 2D plot (as seen from the side) is given in figure 4. In this figure, the y-axis representing σ has been changed to logarithmic scale and since the 95% confidence bounds fall within the marker size of the scatter plot (as seen in figure 3), we have removed the error-bars in this plot. The figure 4 is plotted to show the variation of the MSE values between the different prescriptions. It is easily seen that the prescription given by Scott has the lowest '*a*' value for every σ value used in the present simulation.

From equation (8), it is easily seen that the value of '*a*' controls the rate of decrease of MSE with increase in *n* – lower the value '*a*' the faster is the decrease and vice-versa. Thus, we can take the value of '*a*' as the overall performance index of a given prescription in terms of the MSE measure. From figure 3, the Scott's prescription is found to have the lowest value of '*a*' and thus can be considered the optimal method as per the MSE measure.

Conclusion

The decrease in discrepancy in the estimation of the probability density function of a normal distribution using the method of histograms in terms of the

MSE measure is seen to have a non-linear rate of decrease viz. a decrease proportional to the reciprocal of the sample size with reasonable accuracy. This relation can be used to infer the number of samples that would be required to estimate the underlying normal probability density function for a given experiment to a specified level of accuracy. Further, from among the different methods for the estimation of the bin-number of histogram, the Scott's prescription is found to perform the best in our simulations in terms of the MSE score.

References

[1] Lane D M, Scott D, Hebl M, Guerra R, Osherson D and Zimmer H 2014. Introduction to statistics. United States

[2] Scott D W 2015. Multivariate density estimation: theory, practice, and visualization. United States, John Wiley & Sons

[3] Sturges H A 1926. The choice of a class interval. Journal of the american statistical association 21(153): 65-66

[4] Doane D P 1976. Aesthetic frequency classifications. The American Statistician 30(4): 181-183

[5] Scott D W 1979. On optimal and data-based histograms. Biometrika 66(3): 605-610

[6] Freedman D and Diaconis P, 1981. On the histogram as a density estimator: L2 theory. Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete 57(4): 453-476

[7] Knuth K H 2006. Optimal data-based binning for histograms. arXiv preprint physics/0605197

[8] Mosteller F and Tukey J W 1977. Data analysis and regression: a second course in statistics. United States, Addison-Wesley Pub. Co.