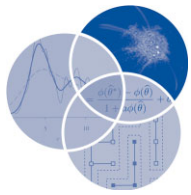
Focus Article



Sturges' rule

David W. Scott*

Histograms are among the most important graphical objects in statistical practice, providing a consistent estimate of any continuous density function with very few assumptions. Restricting attention to bins of equal width, the histogram is sometimes presented as a frequency chart or normalized to be a true density. The construction of a histogram may be specified by either its bin width or by the number of bins. Sturges' rule gives a number-of-bins formula. The formula was the first rule given in the literature and is still widely implemented in software today. This article reviews the underlying rationale for the rule and indicates when it is most appropriate to use in practice. *WIREs Comp Stat* 2009 1 303–306

The construction of a histogram begins with the choice of an interval of support, (a, b). This interval is often taken to be the smallest and largest values of the continuous random sample, $x_1, x_2, \ldots, x_n$, or some convenient 'nice' values that cover all the data. Given the interval (a, b), an equally spaced histogram or frequency diagram is constructed from the bin counts in k equally spaced intervals or bins. The width of each bin is denoted by h, where

$$h = \frac{b-a}{k}. \tag{1}$$

Clearly, the larger the sample size, n, the more bins that can be formed. Sturges[1] proposed the following estimate for the number of bins, m:

$$\hat{k} = 1 + \log_2(n), \tag{2}$$

where the logarithm is taken to base 2. This simple rule is easily implemented and is widely employed in statistics software today. The rationale for Sturges' rule is the focus of this article.

DERIVATION OF STURGES' RULE

Sturges' 1926 article in the *Journal of the American Statistical Association* is but a page in length, and his derivation of Eq. (2) is given verbally. In modern parlance, Sturges' formula is a normal-reference rule. That is, the rule is designed to be exact for normal or Gaussian data. For most other types of data, more bins will be required.

Because a histogram takes a continuous sample and summarizes the data by k bin counts, it is natural to search for a discrete density that well approximates a normal curve. From the central limit theorem, there are many discrete densities that could serve this purpose, but the binomial density is the obvious top candidate. As we have already used the symbol n for the sample size, let us choose the symbol m to denote the number of trials in the binomial process. Thus, the binomial density $B(m, p)$ will be approximately normal when $mp > 5$, and is most normal when $p = 1/2$. Let $Y \sim B(m, p)$; then

$$f(y) = P(Y = y) = \binom{m}{y} p^y (1-p)^{m-y}, \quad y = 0, 1, \ldots, m. \tag{3}$$

But if $p = 1/2$, then the Binomial density is simply

$$f(y) = \binom{m}{y} 2^{-m}. \tag{4}$$

Thus, the fact that the Binomial density is approximately normal results from the fact that the binomial coefficients, $\binom{m}{y}$, for fixed m, are bell-shaped.

Sturges simply suggested that an 'ideal' normal sample would have bin counts, ν_i, given by the binomial coefficients, $\binom{m}{i}$, for $i = 0, 1, \ldots, m$. Such an association to binomial coefficients gives a histogram with $m + 1$ bins.

*Correspondence to: scottdw@rice.edu

Department of Statistics MS-138, Rice University, Houston, TX 77251-1892, USA

DOI: 10.1002/wics.035

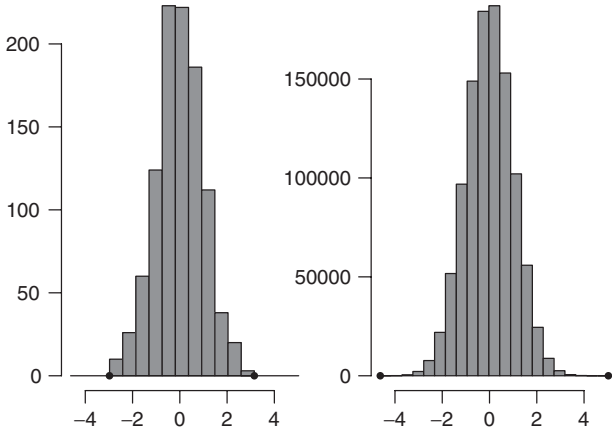


FIGURE 1 | Two examples of Sturges' rule in action with normal data and sample sizes $n = 2^{10}$ and $n = 2^{20}$ with 11 and 21 bins, respectively. The points along the x-axes depict the support intervals, (a, b), of the frequency charts.

Next the total sample size may be computed as

$$n = \sum_{i=0}^{m} v_i = \sum_{i=0}^{m} \binom{m}{i} = \sum_{i=0}^{m} \binom{m}{i} 1^i 1^{m-i} = (1+1)^m = 2^m, \quad (5)$$

using the well-known binomial expansion for the polynomial $(a+b)^m$, with both $a = b = 1$. Solving for m in terms of n, we have

$$m = \log_2(n). \quad (6)$$

Finally, as the number of bins k equals $m+1$, we arrive at Sturges' rule (2)

$$k = m + 1 = 1 + \log_2(n). \quad (7)$$

See Figure 1 for two examples.

EVALUATION OF STURGES' RULE

In this section, we compare Sturges' rule to other rules that have been developed since 1979. First, we attempt to transform Sturges' formula from a number-of-bins to a bin-width rule. It is well-known that if $Z \sim N(0, \sigma^2)$, then

$$\mathrm{E}\left[\max_{1 \le i \le n} |Z_i|\right] \approx \sigma\sqrt{2\log(n)}. \quad (8)$$

Since the largest and smallest order statistics are essentially independent, and approximately $n/2$ points are greater than 0, then a useful approximation to the expected value of the sample range, R, is given by

$$\mathrm{E}[R] = \mathrm{E}\left[\max_{1 \le i \le n} Z_i - \min_{1 \le i \le n} Z_i\right] \approx 2\sigma\sqrt{2\log(n/2)}. \quad (9)$$

Practical application of a number-of-bins rule is complicated when the sample has outliers, which results in a number of empty bins over the interval (a, b). One suggestion is to only count non-empty bins toward the $1 + \log_2(n)$ rule. An alternative suggestion is to trim such outliers from the sample and use the remaining points to determine the interval (a, b), from which a bin width h can be determined by using Eq. (1). Typical software implementations do not use such ideas. For normal data, we combine Eqs (9), (2), and (1) to obtain the Sturges bin-width rule

$$h_{\mathrm{Sturges}} = \frac{2\sigma\sqrt{2\log(n/2)}}{1 + \log_2(n)}. \quad (10)$$

A more modern treatment of choice of histogram smoothing considers a criterion such as integrated mean square error (IMSE). The height of each histogram bin has two error components: variance and bias. Variance may be reduced by increasing the number of points in a bin, either by collecting more data or making the bins wider. Bias, on the other hand, may be reduced by making the bins narrower. The IMSE criterion balances these errors by integrating the pointwise mean squared errors. Scott[2] proved that the optimal IMSE bin width for a histogram is given by

$$h_{\mathrm{Scott}} = \left(\frac{6}{n \int f'(x)^2 \mathrm{d}x}\right)^{1/3} \approx 3.5\sigma n^{-1/3} \quad (11)$$

for a normal density, since $\int f'(x)^2 \mathrm{d}x = (4\sqrt{\pi}\sigma^3)^{-1}$ in that case. In the left frame of Figure 2, the bin width for Sturges' and Scott's rules are displayed for $10 < n < 10^6$. For textbook examples (n not much more than 100), both rules give about the same value. However, for large samples, Sturges' rule suggests much wider bins than Scott's rule.

Next, we compare Sturges' rule with another number-of-bins rule by Terrell and Scott[3]. They investigated possible values of the integral in the denominator of Eq. (11) for densities $f(x)$ with a

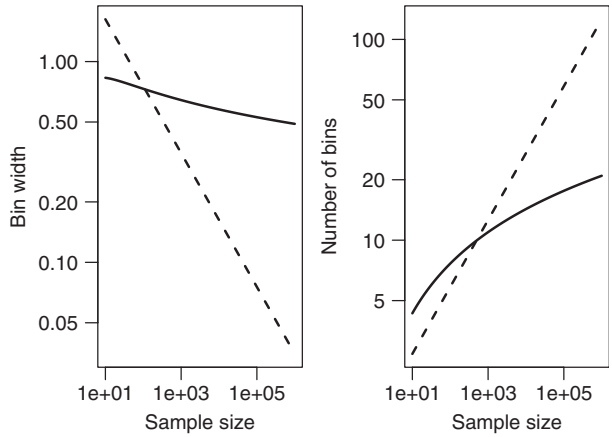


FIGURE 2 | (Left) Bin width for standard normal data as a function of sample size for Sturges' rule (solid line) and Scott's rule (dashed line). (Right) Number of bins as a function of sample size for Sturges' rule (solid line) and over the oversmoothed density (dashed line).

known interval of support (a, b). Although there is no upper bound on this integral, there is a lower bound which occurs for the density

$$f_1(x) = \frac{3}{4}(1 - x^2), \qquad -1 \le x \le 1 \qquad (12)$$

when $(a, b) = (-1, 1)$. This lower bound on the roughness of f translates into an upper bound on the bin width h, which in turn translates into a lower bound on the number of bins $(b - a)/h$. In a real sense, the density $f_1(x)$ is the easiest density to estimate with a histogram. For any other density, more bins will be required. Thus, the authors refer to $f_1(x)$ as the smoothest density for a histogram and the resulting rules 'oversmoothed.'

Specifically, $\int_{-1}^{1} f_1'(x)^2 dx = 3/2$. Thus, the optimal bandwidth in Eq. (11) satisfies

$$h^* = \left(\frac{6}{n \int f'(x)^2 dx}\right)^{1/3} \le \left(\frac{4}{n}\right)^{1/3}. \qquad (13)$$

The Terrell–Scott inequality for the optimal number of bins is given by

$$k^* = \frac{b - a}{h^*} \ge \frac{2}{(4/n)^{1/3}} = \sqrt[3]{2n} \equiv k_{\mathrm{TS}}. \qquad (14)$$

The Terrell–Scott number-of-bins formula may compared directly with Sturges' rule in Eq. (2). Although the Terrell–Scott rule is not a normal-reference rule, Scott[4] shows that according to IMSE theory, the optimal rule for a normal density and the density $f_1(x)$ are very similar. In the right frame of Figure 2, the number of bins according to Sturges' rule and the Terrell–Scott rule are displayed for $10 < n < 10^5$. As expected, for large samples, Sturges' rule suggests far fewer bins. Thus, histograms constructed according to Sturges' rule will be severely oversmoothed.

We illustrate these ideas with a normal sample of size $n = 1{,}000{,}000$. In Figure 3, we show a portion of the density histogram using Sturges' rule, Scott's rule, and the Terrell–Scott rule, as given in Eqs (2), (11), and (14), respectively. Clearly, the histogram based on Sturges' rule has too few bins. The large bias at the edges of the bins is quite apparent. The middle curve depicts the histogram based on the oversmoothed Terrell–Scott rule. The error is an order of magnitude improved when compared with Sturges' rule. Finally, the top curve shows the application of Scott's rule, which is asymptotically optimal. The actual error is improved for this estimate, although a careful examination shows that there are a number

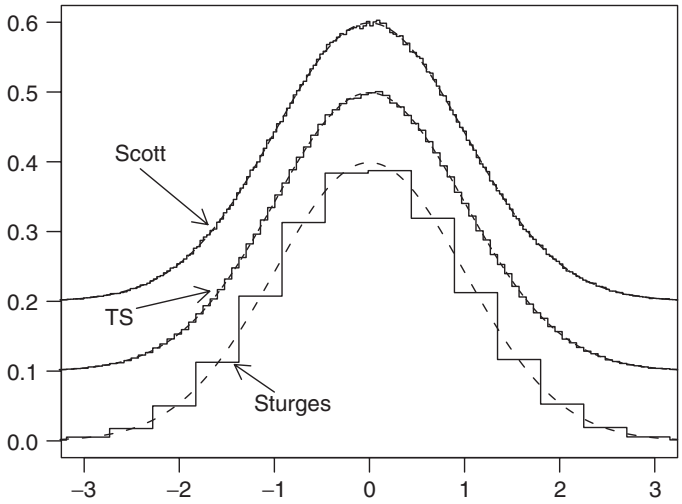


FIGURE 3 | Example of three rules on a normal sample of size 10^6. The sample range is $(-4.93, 4.76)$. The number of bins given by Sturges' rule, the Terrell–Scott rule, and Scott's rule are 21, 126, and 277, respectively. The graph does not display the full extent of the histogram, and the histograms are displaced vertically for clarity. The true normal density is superimposed as a dotted line.

of false (but small) bumps locally in the interval from −0.5 to 0.5. IMSE does not penalize for such bumps. As an aside, the optimal frequency polygon (Scott[5]) does not suffer such extraneous bumps.

CONCLUSIONS

Sturges' rule is the classical formula for providing guidance to the construction of a histogram or frequency curve. It is an example of a normal-reference-rule technique. However, its motivation takes no account of the stochastic nature of a histogram. Thus, its application should be limited, much more limited than its prominent position in statistical software currently offers. By some chance, Sturges' rule happens to coincide with more modern rules for sample sizes $n \approx 100$. Perhaps this explains why early users of Sturges' rule were not dissatisfied with its performance. Modern users more frequently encounter massive datasets, for which the application of Sturges' rule severely oversmooths the histogram, wasting valuable information available in the complete dataset.

ACKNOWLEDGEMENTS

This work was partially supported by NSF award DMS-05-05584 and ONR contract N00014-06-1-0060.

REFERENCES

1. Sturges HA. The choice of a class interval. *Journal of the American Statistical Association* 1926, 21:65–66.
2. Scott DW. On optimal and data-based histograms. *Biometrika* 1979, 66:605–610.
3. Terrell GR, Scott DW. Oversmoothed nonparametric density estimates. *Journal of the American Statistical Association* 1985, 80:209–214.
4. Scott DW. *Multivariate Density Estimation: Theory, Practice, and Visualization*. New York: John Wiley & Sons; 1992.
5. Scott DW. Frequency polygons: theory and application. *Journal of the American Statistical Association* 1985, 80:348–354.

