Peertechz

LIFE SCIENCES GROUP

ANNALS OF Environmental Science and Toxicology OPEN ACCESS JOURNAL

ISSN: 2641-2969 DOI: https://dx.doi.org/10.17352/aest

Review Article

Determination of the number and width of class intervals using R

José Moral de la Rubia*

School of Psychology, Autonomous University of Nuevo León, Monterrey, Mexico

Received: 10 June, 2024
Accepted: 22 June, 2024
Published: 24 June, 2024

***Corresponding author:** José Moral de la Rubia, School of Psychology, Autonomous University of Nuevo León, Monterrey, Mexico, E-mail: jose.morald@uanl.edu.mx ; jose_moral@hotmail.com

ORCiD: http://orcid.org/0000-0003-1856-1458

Keywords: Histogram; Frequency tables; Chi-square goodness-of-fit test; Number of class intervals; Bin width

Copyright License: © 2024 De la Rubia JM. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

https://www.agriscigroup.us

Check for updates

Abstract

The histogram and frequency table are fundamental tools for describing continuous variables or discrete variables with many values. Most statistical programs are not flexible, nor do they explicitly state the rules they use to construct histograms or provide guidelines for constructing interval tables. However, by programming or applying the appropriate procedures, this can be achieved with Excel, MATLAB, and R. The objective of this methodological article is to provide a script for the R program to calculate the number and width of class intervals using eight rules that provide a uniform width (four depending on the sample size and four based on optimal width). The script automates the selection of the rule to produce an interval table and a histogram with overlaid density and normal curves. Additionally, symmetry is assessed using the D'Agostino test, mesokurtosis with the Anscombe-Glynn test, and normality with the Lilliefors, Anderson-Darling, and Shapiro-Francia tests. Furthermore, three rules are calculated that provide variable width: one for samples of 25 to 39 data points (multiple of 5) and two for samples of at least 40 data points (Mann-Wald and Moore). Once one of these three rules is chosen, it is applied to the normality check using the likelihood ratio test. Additionally, an optimal histogram provided by R from its basic library is computed. The script is applied to two examples and is adapted to the small samples (< 25 data points) in a third example. It is concluded that this script can be of practical and didactic use.

Introduction

Two basic descriptive tools for describing continuous variables are the frequency table with class intervals and the histogram [1]. There are many rules for determining the number of class intervals and their width when constructing a frequency table or histogram from a random sample of n data points drawn from a continuous quantitative variable or from a sample of discrete values when their variability is large [2,3].

The domain of a continuous quantitative variable is understood to span the entire real line or a subset of the real numbers, while the domain of a discrete quantitative variable spans the integers or a subset of the integer numbers [4]. Additionally, both types of variables, being quantitative, have a unit of measurement, such as the standard deviation when the scale is centered on the arithmetic mean [5].

Most programs are not flexible and do not clearly state the rules they use to construct histograms. Moreover, many do not provide rules for constructing an interval table [4,6,7]. In this regard, the spreadsheet program Excel may be one of the best options [8], alongside the mathematical program MATLAB [9]. Another option is the statistical program R, developed by the mathematical community and available for free since 2000. It has undergone continuous updates and improvements, but its programmability is often viewed as a drawback [10], leading to its underutilization in various scientific disciplines [11], especially in the social sciences [12].

This methodological article aims to present a script developed for the R program [13] to determine the number of class intervals (k) and their width, which can be uniform (w) or variable (w_i). Eleven rules are shown, which can be grouped

Citation: De la Rubia JM (2024) Determination of the number and width of class intervals using R. Ann Environ Sci Toxicol 8(1): 022-042.
DOI: https://dx.doi.org/10.17352/aest.000077

into three sets [2], along with an optimal histogram provided by R [14], and additional inferential information about the distribution and randomness of the sample.

Frequency tables with class intervals and histograms are fundamental tools in data analysis, offering a structured, clear, and insightful way to represent and understand quantitative data. They play a critical role in identifying patterns, understanding distributions, detecting anomalies, and supporting further statistical analysis. However, the problem of determining the number and width of the class interval arises. Thus, having a script that facilitates these decisions and provides the table and histogram is a valuable tool for researchers. In addition, the script tests for randomness, which is a fundamental assumption for statistical inference, and for normality, skewness, and kurtosis, which are key data for deciding on which tests to use for hypothesis testing. The freely available R program is particularly suitable for these purposes with the difficulty that the scripts must be written by the researcher, which is done in the present article. It should be noted that the script is developed for samples of at least 25 data points, but it is also adapted for situations involving small samples (n < 25 data points).

Rules for determining the number and width of class intervals included in the script

The first group consists of four rules. In these rules, the number of class intervals (k) is established first, and then the uniform width (w) is determined by the quotient between the range and the number of intervals: $a = (max\text{-}min) / k$. The frequency per interval is variable (n_i; i = 1, 2, ..., k). The four rules included in this group are square root [15], Rice University [16], Sturges [17], and Doane [18].

The second group also consists of four rules, but these are based on algorithms for optimizing a uniform width per class interval or bin. First, the uniform width (w) is established and from there the number of class intervals (k) is determined by the quotient between the sample range and the uniform width, rounded up: $k = [max - min) / w]$. The frequency per interval is variable (n_i; i = 1, 2, ..., k). The four rules included in this group are those of Scott [19], Freedman and Diaconis [20], Rudemo [21], and Shimazaki and Shinomoto [22].

The third group consists of three rules that achieve a uniform frequency (n_i) per class interval when the number of class intervals (k) is established, resulting in variable widths (w_i). These rules were developed for the application of Pearson's chi-square test in normality testing, based on its asymptotic approximation to Pearson's chi-square distribution with k - 3 degrees of freedom, where k is the number of class intervals [23]. This approximation requires that no observed frequency be less than 0, no expected frequency be less than 1, and that at least 80% of the expected frequencies be greater than 5 [24].

In the latter group, there are three rules: the multiple of 5 for sample sizes from 25 to 39, and those of Mann-Wald [25] and Moore [26] for sample sizes of at least 40. The multiple of 5 rule is developed in this script to complement the Mann-Wald and Moore rules. Additionally, the script does not use the chi-square test, but the likelihood ratio test as it is more powerful [27], and the Williams' [28] correction is optionally applied, as suggested not only for 2 × 2 tables [29] but also for one-way goodness-of-fit tests when the degrees of freedom are small [30].

The script also includes the R program's automatic option for optimal bins. The computational algorithm to obtain this histogram minimizes the squared error among the values of each class interval or bin and the average of the bins (default) or minimizes the mean squared error by dividing the squared error by the width of the bin [14].

The eleven rules previously mentioned and executable with the script are briefly described below. This is followed by a presentation of the script, which is structured into three parts and applied to two random samples of 51 participants: one drawn from a normal distribution and the other from an unknown distribution. Finally, some aspects of the script are discussed, conclusions are drawn, and suggestions for application are provided.

Rules for establishing k-class intervals with uniform width and variable frequencies: The square root rule was introduced in 1892 [15] by the English mathematician and father of contemporary statistics, Karl Pearson (1857-1936). The number of class intervals or bins is obtained by taking the square root of the sample size and then rounding up ($[\sqrt{n}]$). This method has no distributional assumptions and is recommended for small samples, typically those smaller than 100 [31]. Its advantage is the ease of calculation and distributional universality, particularly recommended for the arcsine distribution, but its disadvantage is that it provides a very large number of class intervals when the sample size is very large. Refer to Equation 1, in which x represents a sample of size n drawn randomly from a quantitative variable X.

$$\mathrm{x} = \left\{x_i\right\}_{i=1}^{n} \subseteq \mathrm{X};\ k = \left[\sqrt{n}\right] \Rightarrow w = \left[max\left(\mathrm{x}\right) - min\left(\mathrm{x}\right)\right] / k \quad (1)$$

The Rice University rule was developed in the statistics department of William Marsh Rice University, a private university based in Houston, Texas [16]. The number of class intervals or bins is obtained by doubling the cube root of the sample size and rounding it up (Equation 2). This method has no distributional assumptions, can be used with any sample size, and is based on the rules of David Warren Scott [19] and Freedman and Diaconis [20], where the cube root of the sample size is used as the denominator to determine the width of the class intervals or bins [2]. Interestingly, David Warren Scott was a professor in the statistics department of this university, where he served as department head from 1990 to 1993, and was recognized as professor emeritus by the university in 2021 [32]. Like the square root rule, its advantage is ease of calculation and distributional universality, especially recommended for the arcsine distribution. However, its disadvantage is that it does not work well with very large samples compared to the rules of optimized width.

Citation: De la Rubia JM (2024) Determination of the number and width of class intervals using R. Ann Environ Sci Toxicol 8(1): 022-042.
DOI: https://dx.doi.org/10.17352/aest.000077

$$k = \left[2 \times \sqrt[3]{n}\right] \Rightarrow w = \left[max\left(x\right) - min\left(x\right)\right] / k \quad (2)$$

The Sturges rule was developed by the German-born American statistician Herbert Sturges (1882-1958) in 1926 [17]. The number of class intervals or bins is obtained by adding 1 to the logarithm base 2 of the sample size, and then rounding up (see Equation 3). This method assumes symmetry and is based on expressing the sample size as a binomial expansion with parameter p = ½ (constant probability of success in n independent trials with two possible outcomes: 1 = success and 0 = failure or no success). Its advantage is that it is simple to calculate and is the default option in the 'hist' function of the R program. Its disadvantage is that it assumes symmetry and an underlying distribution that converges to normality.

$$k = \left[1 + \log_2\left(n\right)\right] \Rightarrow w = \left[max\left(x\right) - min\left(x\right)\right] / k \quad (3)$$

Doane's rule was developed by the American statistician David Patrick Doane [18], Professor Emeritus of Economics at Oakland University. The number of class intervals or bins is defined based on the Sturges rule, with an additional term introduced as a skewness correction (Equation 4). This additional term is determined by the logarithm base 2 of the sum of one and the skewness coefficient, derived from the standardized third central moment in absolute value (Equation 5), divided by its asymptotic standard error (Equation 6). Similar to previous rules, the result is rounded up. Doane's rule does not require a symmetric distribution and is recommended as an alternative to the Sturges rule when the symmetry assumption is not met [33]. This is a more complex calculation rule than Sturges' rule, with the advantage that it does not assume symmetry or convergence to normality. Although it corrects for symmetry, it does not take kurtosis into account.

$$k = \left[1 + \log_2\left(n\right) + \log_2\left(1 + \frac{\left|b_1\left(x\right)\right|}{EE\left[b_1\left(x\right)\right]}\right)\right] \Rightarrow w = \frac{max\left(x\right) - min\left(x\right)}{k} \quad (4)$$

$$b_1\left(x\right) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_n\left(x\right)}\right)^3 ; \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} ; s_n\left(x\right) = \sqrt{\frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{n}} \quad (5)$$

$$EE\left[b_1\left(x\right)\right] = \sqrt{\frac{6\left(n-2\right)}{\left(n+1\right)\left(n+3\right)}} \quad (6)$$

Rules for establishing uniform width, determining *k* class intervals with varying frequencies: Scott's rule was developed by the American statistician David Warren Scott, professor emeritus of the Department of Statistics at Rice University, as previously mentioned [32]. The uniform width of the class intervals or bins is determined by the quotient of 3.49 times the sample standard deviation and the cube root of the sample size (Equation 7). This rule assumes a normal distribution and is based on minimizing the integrated mean square error when estimating the density per interval [19]. The formula is derived by approximating the density estimation using a Gaussian kernel function [34]. However, it performs well with various distributions other than the normal distribution, except for the arcsine distribution [2]. Scott's rule is recommended for large sample sizes [33] and improves with increasing sample size [2]. Its advantage is that it uses an optimization algorithm and is simple to calculate, but its disadvantage is that it assumes a normal distribution and works poorly with small samples.

$$w = \frac{3.49 \times \sqrt{\frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{n-1}}}{\sqrt[3]{n}} \Rightarrow \left[k = \frac{max\left(x\right) - min\left(x\right)}{w}\right] \quad (7)$$

The Freedman-Diaconis rule was developed by the American statisticians David Freedman of the University of California and Persi Diaconis of Stanford University. The uniform width of the class intervals or bins is determined by the ratio of twice the interquartile range to the cube root of the sample size [20]. Refer to Equation 8, where sample quartiles are denoted by $q_p(x)$, and p represents the order of the quantile, which, in the case of quartiles, is 0.25 (lower), 0.5 (middle), and 0.75 (upper).

$$w = \frac{2 \times IQR\left(x\right)}{\sqrt[3]{n}} = \frac{2 \times \left(q_{0.75}\left(x\right) - q_{0.25}\left(x\right)\right)}{\sqrt[3]{n}} \Rightarrow k = \left[\frac{max\left(x\right) - min\left(x\right)}{w}\right] \quad (8)$$

Quantiles can be estimated using rules 6, 7, and 8 in R [35]. The most widely used rule for calculating sample quantiles is rule 6. This rule is based on expressing the order of the quantile p as the arithmetic mean of the statistic of order i in a sample of size n drawn from a standard continuous uniform distribution, which serves as a non-informative prior distribution when estimating a probability in Bayesian inference. Rule 7 is the default in the R program and is based on expressing the order of the quantile p as the mode of the statistic of order i in a sample of size n drawn from a standard continuous uniform distribution. Rule 8 is especially recommended when the distribution of the variable is unknown [36,37], as indicated by the simulation study of Hyndman and Fan [38] and the exploratory data analysis of Tukey [39]. It is based on expressing the order of the quantile p as the median of the statistic of order i in a sample of size n drawn from a standard continuous uniform distribution.

It should be noted that the Freedman-Diaconis rule does not require any distributional assumptions [20]. It is recommended for large sample sizes [33] and improves with increasing sample size [2]. It performs well with various distributions, although it yields the worst results with the arcsine distribution, similar to Scott's rule. For such bimodal leptokurtic distributions, the square root and Rice University rules are recommended [2]. Therefore, its advantage is that it uses an optimization algorithm, is simple to calculate, and has

Citation: De la Rubia JM (2024) Determination of the number and width of class intervals using R. Ann Environ Sci Toxicol 8(1): 022-042.
DOI: https://dx.doi.org/10.17352/aest.000077

no distributional assumptions, but its disadvantage is that it works poorly with small samples.

Rudemo's rule was developed by the Danish mathematician and engineer Mats Rudemo in 1982 [21], adopting Scott's optimized bin-width approach without assuming a normal distribution. The uniform width of the intervals or bins is determined by minimizing the integrated mean square error using a leave-one-out cross-validation procedure. Refer to Equation 9, where n_i represents the density per interval using kernel estimation, and h is the bandwidth used in this estimation. This bandwidth corresponds to the uniform width of the k-class intervals or bins when estimating the density per interval, which is the sum of the densities of the data points within the interval.

$$\arg\min_{h}\left(\frac{2}{h(n-1)}-\frac{n+1}{n^2h(n-1)}\sum_{i=1}^{k}n_i^2\right)=a\Rightarrow k=\left[\frac{max(x)-min(x)}{w}\right] \tag{9}$$

The most commonly used kernel function is the Gaussian or normal function (Equation 10), which has a bandwidth given by Silverman [40] derived from the minimization of the integrated mean square error. The value of this bandwidth (h) is 0.9 times the minimum between the sample standard deviation (s_{n-1}) and the sample interquartile range divided by the interquartile range of the standard normal distribution (IQR / 1.34), with this weighted minimum divided by the fifth root of the sample size [41]. See Equation 11. However, the kernel function that optimizes the integrated mean square error most effectively is Epanechnikov's parabolic function [42,43]. See Equation 12. In this case, to determine the bandwidth, one can apply the method of Sheather and Jones [44], which is based on the minimization of the asymptotic integrated mean square error [45,46].

$$x\in\{x_i\}_{i=1}^{n}\subseteq X;\hat{f}_h(x)=\frac{1}{nh\sqrt{2\pi}}\sum_{i=1}^{n}e^{-\frac{1}{2}\left(\frac{x-x_i}{h}\right)^2} \tag{10}$$

$$h=\frac{0.9\times min\left(s_n(x),\frac{q_{0.75}(x)-q_{0.25}(x)}{z_{0.75}-z_{0.25}}\right)}{\sqrt[5]{n}}=\frac{0.9\times min\left(s_n(x),\frac{IQR(x)}{1.34}\right)}{\sqrt[5]{n}} \tag{11}$$

$$\hat{f}_h(x)=\frac{3}{4nh}\sum_{i=x-k}^{x+h}\left(1-\left(\frac{x-i}{h}\right)^2\right) \tag{12}$$

The Shimazaki-Shinomoto rule was developed by Japanese physicist Hideaki Shimazaki and Japanese systems engineer and neuroscientist Shigeru Shinomoto in 2007 [22]. It is based on finding the optimal uniform width per interval or bin from the loss function. See Equation 13, where m represents the arithmetic mean, s_n^2 symbolizes the uncorrected variance of bias, n_i is the density per interval obtained by kernel estimation, Δ denotes the bandwidth in the kernel estimation of the point density, n constitutes the sample size, and the product $\Delta \times n$ is the bin width (w). By default, the Gaussian kernel with the bandwidth of Silverman [40] is used, as shown in Equation 10 [47].

$$\arg\min_{\Delta}\left(\frac{2m(n_i)-s_n^2(n_i)}{(\Delta\times n)^2}\right)=\Delta;a=\Delta\times n\Rightarrow k=\left[\frac{max(x)-min(x)}{w}\right] \tag{13}$$

The advantage of the Rudemo and Shimazaki-Shinomoto rules is that they use an optimization algorithm and do not rely on distributional assumptions. However, their disadvantage is that they require software for computation and perform poorly with small samples.

Rules for establishing k with uniform frequency and variable widths: The Mann-Wald rule is applied in the context of normality testing through either the chi-square goodness-of-fit test or the likelihood ratio test [23]. These two American Jewish statisticians of Austro-Hungarian origin propose that the number of class intervals (k) falls within an interval [k_{min}, k_{max}]. The lower bound (k_{min}) is determined by applying Equation 14, and the upper bound (k_{max}) is obtained using Equation 15 [25]. Once the number of intervals is established, the uniform frequency (n_I) is calculated by dividing the sample size by k.

$$k_{min}=2\left(\frac{2n^2}{\left(\ddot{O}^{-1}(\alpha)\right)^2}\right);n_{min}=\left[\frac{n}{k_{min}}\right];n_{exc}=n-k_{min}\times n_{min} \tag{14}$$

$$k_{max}=4\left(\frac{2n^2}{\left(\ddot{O}^{-1}(\alpha)\right)^2}\right);n_{max}=\left[\frac{n}{k_{max}}\right];n_{exc}=n-k_{max}\times n_{max} \tag{15}$$

If the quotient is a whole number, all intervals have the same frequency. If the quotient is a number with decimal places, it is rounded down to yield the common frequency. In this case, an excess frequency (n_{exc}) arises, which is distributed by adding an element to the central intervals. Therefore, the presence of a surplus means that the $n - k \times n_I$ central intervals contain $n_I + 1$ elements.

In principle, the value of k must be a divisor of n for the quotient to be an integer, and it must be greater than or equal to 5; therefore, the sample size must be at least 40. To meet these requirements, n can be decomposed into prime factors, and the product of factors chosen should fall within the interval [k_{min}, k_{max}], be as close as possible to the lower limit (≥ k_{min}), and be greater than or equal to 5.

To construct the k class intervals, the n data are sorted in ascending order. The minimum value becomes the lower limit of the first interval. The next $n_I - 1$ elements are included, with the last of these values becoming the upper limit of the first interval. The subsequent value becomes the lower limit of the second interval, and the process continues until the k-th interval, where the upper limit is the maximum value in the

Citation: De la Rubia JM (2024) Determination of the number and width of class intervals using R. Ann Environ Sci Toxicol 8(1): 022-042.
DOI: https://dx.doi.org/10.17352/aest.000077

sample. If there is a surplus, the $n - k \times n_I$ central intervals contain $n_I + 1$ elements. If k is odd, one element is added to the center interval, alternating between left and right for the placement of the surplus. If k is even, there are two center intervals. The process begins with the left-center interval and continues with the right-center interval when placing the surplus, alternating from left to right for the placement of one element per interval. Ideally, a value for k that does not result in a surplus is sought. It is important to note that with this rule, the width per interval or bin always varies.

If the significance level of the goodness-of-fit test (α) is set to 0.05, the k_{min} is $[1.88 \times n^{0.4}]$ and the k_{max} is $[3.77 \times n^{0.4}]$. Because the number of class intervals (k) that provides the best adequacy to the goodness-of-fit test is close to the lower limit, Moore [26] recommends using $2 \times n^{0.4}$, rounded to the nearest integer, to obtain the number of class intervals. Once k is defined, the uniform frequency (n_I) and the excess frequency (n_{exc}) are determined, resulting in the variable width of the bins (w_i). See Equation 16.

$$k = \left[2 \times n^{0.4}\right]; n_I = [n/k]; n_{exc} = n - k \times n_I \quad (16)$$

What to do with a sample smaller than 40? If the sample size is at least 25 but less than 40, you can apply what is referred to in this article as the "rule of the multiple of 5". Divide the sample size by 5 and round down to obtain the uniform or common frequency per interval: $n_I = [n / 5]$. The number of class intervals is determined by dividing the sample size by the uniform frequency: $k = [n / n_I]$. If the quotient $n / 5$ results in an integer, all intervals have the same frequency (n_I), which happens when n ends in 0 or 5. If a decimal is obtained, the integer part (rounded down) represents the common frequency of the k intervals, and there is an excess frequency: $n_{exc} = n - k \times n_I$, which is distributed by incrementing one element in each of the $n - k \times n_I$ intervals at the center.

The advantage of these last three rules is that they are designed for calculating the chi-square test or likelihood ratio from a class interval table, ensuring adequate convergence of the test statistic to the chi-square distribution. However, their disadvantage is that they produce an uninformative histogram that resembles a uniform distribution, so its representation is omitted in the script.

Method

First, the script is developed using the R programming language [13]. The script is structured into three parts: 1) uniform amplitude rules, from which a frequency table and histogram are derived; 2) the optimal histogram of Kreider [14]; and 3) uniform frequency rules, from which a frequency table and the likelihood ratio test for normality are derived.

It should be noted that the developed script can be executed using the R or RStudio programs, with the eight required libraries (DEoptim, pracma, ggplot2, randtests, modeest, moments, scales, and nortest) installed on the personal computer. This way, the first high-resolution histogram is obtained as a JPEG file, and the second plot can be saved as a high-quality JPEG file. Another option is to run the script online at the following address: https://rdrr.io/snippets/. This method is more convenient, but the two graphics are of low resolution.

This script is applied to two samples of 51 data points each. The first sample is highlighted in blue at the beginning of the script. For practical significance, the data correspond to scores on a visual attention capacity test on a D scale (with a mean of 5.5 and a standard deviation of 2). It was created using Excel's random number generator by drawing a random sample of 51 data points from a normal distribution with a location parameter $\mu = 5$ and scale $\sigma = 2$. The second sample was generated from the same normal distribution, seeking an output away from the Gaussian bell, slightly modified to increase its abnormality, not by outliers, as would be the case with a contaminated normal distribution, but by depression at its center, as could a mixture of two truncated normal distributions with different location parameters [48]. This modification aims to create a random sample of the same size but with an unknown, non-normal distribution.

The script developed is for a sample of at least 25 data points, but a simplified script for small samples ($n < 25$) is shown at the end of the Results section. This script is applied to a random sample of 16 data points generated from an arcsine distribution modified to provide a larger mode on the left and a smaller mode on the right at the extremes of a bounded range.

Result

The script for R for a sample of at least 25 data points

In the first part, the four previously seen rules for establishing the number of class intervals (k) with a uniform width (w), determined by k, and with a variable frequency (n_i) are calculated, as well as the four previously shown rules for establishing the uniform width (w) with the number of intervals (k) determined by w and with a variable frequency (n_i). These eight rules are summarized in a four-column table: rule, number of class intervals (k), uniform width (w), and frequency per interval (n_IC). Using the 'mlv' function of the 'modeest' library to obtain the mode by the value with the highest frequency, k, and w are selected. If there is no unique modal value for k and the hypotheses of symmetry, mesokurtosis, and normality at a 5% significance level hold, Scott's rule is adopted. If there is neither a unique mode nor normality, the Freedman-Diaconis rule is chosen. However, if the distribution shows a U-shaped profile corresponding to an arcsine distribution, it is recommended to use the square root or Rice University rules instead of the Freedman-Diaconis rule. It appears in the script as non-executable code (due to the # symbol before it).

The testing of symmetry is done using D'Agostino's test [49] and that of mesokurtosis using the Anscombe-Glynn test [50]. For this purpose, it is necessary to load the 'moments' library. In addition, using the 'nortest' library, normality is tested by at least one of three tests: Lilliefors' test [51], based

Citation: De la Rubia JM (2024) Determination of the number and width of class intervals using R. Ann Environ Sci Toxicol 8(1): 022-042.
DOI: https://dx.doi.org/10.17352/aest.000077

on the maximum linear distance between the empirical and theoretical cumulative distribution functions; Anderson-Darling's test [52], based on the standardized quadratic distance between the two functions; and Shapiro-Francia test [53], based on the shared variance between the empirical and theoretical quantiles. These three tests were chosen because they are recommended for their power and adequacy [54-56]. Statistical power is calculated for all inferential tests. It is obtained through a bootstrap simulation procedure, except in the case of the likelihood ratio test or G-test. For this test, the complementary cumulative distribution function of a non-central chi-square distribution is used.

The quantiles for the Freedman-Diaconis rule are calculated using rule 8, as recommended by Hyndman and Fan [38] and Tukey [39]. In contrast, rule 9 is used for the Shapiro-Francia test to obtain the quantiles, since these are the theoretical quantiles for a sample randomly drawn from a normal distribution [53]. Additionally, the sample's randomness is tested using the Wald-Wolfowitz runs test [57], which requires the 'randtests' library.

It should be noted that the significance level of 0.05 can be modified, for example, to 0.1 for small samples of 20 to 30 and to 0.01 for large samples of 1000 or more data points. It is not recommended to use the script with a sample smaller than 20; preferably, the sample should be at least 25. Once the script has selected the number and uniform width of the class intervals by the convergence of results or other conditions, it automatically produces the interval table and histogram with the overlaid density and normal curves. The plot is of high resolution as it is created using the 'ggplot2' library. The 'scales' library is also necessary for this plot.

Rudemo's rule [21] in this script is not available in any statistical program. Its code requires the 'DEoptim' library to find the optimum in the presence of multiple local minima. By default, a range between 0.001 and the sample maximum for the function being minimized is used. However, in the case of a very large amplitude that determines a very low number of intervals compared to the Shimazaki-Shinomoto rule [22] and other optimization rules, a range between 0.001 and a quarter of the sample maximum is provided as an option. Here, we are considering a value close to the semi-interquartile range, as an approximation to the standard deviation [58,59], which represents a distance four times larger than that proposed by Fisher [60] for the uniform width per class interval. The Shimazaki-Shinomoto rule is indeed programmed in R and requires the 'pracma' library.

The second part of the script is very short. It allows the creation of an optimal histogram using Kreider's algorithm [14], which is available among the basic functions of R. This plot is of low resolution. If the script is run with the R program installed on the computer, the R graphics device (ACTIVE), when expanded, offers a "save as" function in its toolbar under "File". Among the saving options is the high-quality JPEG format, which overcomes this limitation.

In the third part of the script, three rules are computed to define the number of class intervals with uniform frequency and variable width. The Mann-Wald [25] and Moore [26] rules are computed, which require a sample of at least 40 data points. For the Mann-Wald (MW) rule, its lower bound, upper bound, and an additional value are computed. This additional value appears only if there is an integer, the closest to the lower limit without exceeding the upper limit, which gives a uniform frequency of at least 5 cases and no excess. Otherwise, the message "NC = condition not met" appears. Additionally, these two rules are complemented by a third one, called the multiple of 5, for samples between 25 and 39 data points. Therefore, it is recommended that the minimum size of the random sample should preferably be 25 to use the script.

These three rules are presented in a four-column summary table: rule, number of class intervals (k), variable width (w_i), and uniform frequency per interval (n_IC). The user then chooses the rule (multiple of 5, Mann-Wald, or Moore) by assigning values to three parameters: k = number of class intervals, n_I = uniform frequency per interval, and n_{exc} = excess frequency. Under these parameters, the frequency table and the likelihood ratio test for normality [61] with the continuity correction of Williams [28] and a significance level of 5% are calculated. Modifiable parts of the script are highlighted in blue.

```
# Replaces the vector of scores with its sample, with the data separated by commas.

x <- c(8.89, 2.36, 7.98, 8.25, 8.05, 5, 3.85, 4.69, 4.7, 5, 5.97, 7.36, 5.88, 7.94, 1.82, 5.47, 5.33, 7.54, 3.97, 7.54, 4.57, 5.83, 3.97, 5.01, 8.47, 2.09, 4.85, 6.05, 2.56, 3.63, 7.91, 6.43, 1.85, 6.69, 0.8, 6.73, 5.02, 6.31, 3.7, 8.48, 6.59, 7.42, 2.8, 9.93, 2.52, 5.93, 10.29, 6.76, 6.01, 4.28, 4.43)

# Square root rule (Pearson, 1892)

n <- length(x)

k_sqrt <- ceiling(sqrt(n))

R <- max(x) - min(x)

w_sqrt <- R / k_sqrt

if (k_sqrt == 1) {n_sqrt <- n} else {n_sqrt <- "variable"}

# Rice University Rule (Lane, 2015)

k_Rice <- ceiling(2 * n^(1/3))

w_Rice <- R / k_Rice

if (k_Rice == 1) {n_Rice <- n} else {n_Rice <- "variable"}

# Sturges' Rule (1926)

k_Sturges <- ceiling(1 + log2(n))

w_Sturges <- R / k_Sturges

if (k_Sturges == 1) {n_Sturges <- n} else {n_Sturges <- "variable"}
```


Citation: De la Rubia JM (2024) Determination of the number and width of class intervals using R. Ann Environ Sci Toxicol 8(1): 022-042.
DOI: https://dx.doi.org/10.17352/aest.000077

```
# Doane's Rule

b1 <- mean((x - mean(x))^3) / sqrt(mean((x - mean(x))^2))^3

se_b1 <- sqrt((6*(n - 2)) / ((n + 1)*(n + 3)))

k_Doane <- ceiling(1 + log2(n) + log2(1 + abs(b1) / se_b1))

w_Doane <- R / k_Doane

if (k_Doane == 1) {n_Doane <- n} else {n_Doane <- "variable"}

# Scott's Rule

sd <- sd(x)

w_Scott <- 3.49 * sd / n^(1/3)

k_Scott <- ceiling(R / w_Scott)

if (k_Scott == 1) {n_Scott <- n} else {n_Scott <- "variable"}

# Freedman-Diaconis Rule

IQR <- quantile(x, probs = 0.75, type = 8) - quantile(x, probs = 0.25, type = 8)

w_FD <- 2 * IQR / n^(1/3)

k_FD <- ceiling(R / w_FD)

if (k_FD == 1) {n_FD <- n} else {n_FD <- «variable»}

# Rudemo's Rule

set.seed(123)

minimize_function <- function(h, x) {n <- length(x)

xi <- sort(x)

di <- density(xi, kernel = "gaussian")

range <- max(x) - min(x)

k <- ceiling(range / h) + 1

interval_densities <- rep(0, k)

for (i in 1:k) {interval_min <- min(x) + (i - 1) * h

interval_max <- min(x) + i * h

interval_densities[i] <- sum(di$y[xi >= interval_min & xi < interval_max])}

result <- 2 / (h * (n - 1)) - (n + 1) / (n^2 * h * (n - 1)) * sum(interval_densities^2)

return(result)}

library(DEoptim)

control_params <- DEoptim.control(trace = FALSE, itermax = 100, NP = 50)

result <- DEoptim(fn = minimize_function, lower = 0.001, upper = max(x), x = x, control = control_params)$optim$bestmem

#result <- DEoptim(fn = minimize_function, lower = 0.001, upper = max(x)/4, x = x, control = control_params)$optim$bestmem # In case of a very small number of class intervals

w_Rudemo <- result

k_Rudemo <- ceiling((max(x) - min(x)) / w_Rudemo)

if (k_Rudemo == 1) {

n_Rudemo <- n} else {n_Rudemo <- "variable"}

# Shimazaki-Shinomoto Rule

library(pracma)

k <- ceiling(4 * (max(x) - min(x)) / sd(x))

SS <- histss(x, n = 10, plotting = FALSE)

k_SS <- length(histss(x, n = k, plotting = FALSE)$counts)

w_SS <- R / k_SS

if (k_SS == 1) {n_SS <- n} else {n_SS <- "variable"}

# Testing the randomness of the sample sequence using the Wald-Wolfowitz runs test

library(randtests)

runse <- runs.test(x, alternative = "two.sided", threshold = median(x), pvalue = 'exact')

runsa <- runs.test(x, alternative = "two.sided", threshold = median(x), pvalue = 'normal')

alpha <- 0.05

ww_power <- function(x, alpha, B = 1000) {n <- length(x)

p_values <- numeric(B)

for (i in 1:B) {bootstrap_sample <- sample(x, replace = TRUE)

result <- runs.test(x, alternative = "two.sided", threshold = median(x), pvalue = 'exact')

p_values[i] <- result$p.value}

power <- mean(p_values < alpha)

return(power)}

set.seed(123)

power <- ww_power(x, alpha)

cat("Wald-Wolfowitz runs test. Criterion: median", "\n")

cat("Number of runs: r =", runse$runs, "\n")
```


Citation: De la Rubia JM (2024) Determination of the number and width of class intervals using R. Ann Environ Sci Toxicol 8(1): 022-042.
DOI: https://dx.doi.org/10.17352/aest.000077

```
cat("n_0 = #(x_i < mdn(x)) =", runse$parameter["n1"], "and", "n_1 = #(x_i > mdn(x)) =", runse$parameter["n2"], "\n")

cat("n = n_0 + n_1 =", runse$parameter["n"], "\n")

cat("Two-tailed exact probability: p =", round(runse$p.value, 4), "\n")

cat("Mean: M(R|n_0, n_1) =", runse$mu, "and", "Standard deviation: SD(R|n_0, n_1) =", round(sqrt(runse$var), 4), "\n")

cat("Standardized number of runs: z_r =", round(runse$statistic, 4), "\n")

cat("Two-tailed asymptotic probability: p =", round(runsa$p.value, 4), "\n")

cat("Statistical power for the Wald-Wolfowitz runs test using bootstrap simulation: ϕ =", power, "\n")

# Statistical data for the calculation of the eight uniform width rules

cat("Statistical data for calculation of uniform width rules", "\n")

cat("Sample size: n =", n, "\n")

cat("Sample range: R(x) = max(x) - min(x) =", R, "\n")

cat("Sample interquartile range (quartiles by rule 8): IQR(x) =", round(IQR, 4), "\n")

cat("Skewness coefficient based on the standardized third central moment: b_1(x) = m_3 / (m_2)^(3/2) =", round(b1, 4), "\n")

cat("Asymptotic standard error of b_1: se(b_1) =", round(se_b1, 4), "\n")

cat("Sample standard deviation (with Bessel's correction): sd(x) =", round(sd, 4), "\n")

# Testing for symmetry

library(moments)

cat("Testing for symmetry", "\n")

agostino.test(x, alternative = "two.sided")

agostino_power <- function(x, alpha, B = 1000) {n <- length(x)

p_values <- numeric(B)

for (i in 1:B) {bootstrap_sample <- sample(x, replace = TRUE)

result <- agostino.test(bootstrap_sample)

p_values[i] <- result$p.value}

power <- mean(p_values < alpha)

return(power)}

set.seed(123)

power <- agostino_power(x, alpha)

cat("Statistical power for the D'Agostino skewness test using bootstrap simulation: ϕ =", power, "\n")

agostino <- agostino.test(x, alternative = "two.sided")

is_symmetric <- agostino$p.value > alpha

# Testing for mesokurtosis

cat("Testing for mesokurtosis", "\n")

anscombe.test(x, alternative = "two.sided")

ag_power <- function(x, alpha, B = 1000) {n <- length(x)

p_values <- numeric(B)

for (i in 1:B) {bootstrap_sample <- sample(x, replace = TRUE)

result <- anscombe.test(bootstrap_sample, alternative = "two.sided")

p_values[i] <- result$p.value}

power <- mean(p_values < alpha)

return(power)}

set.seed(123)

power <- ag_power(x, alpha)

cat("Statistical power for the Anscombe-Glynn kurtosis test using bootstrap simulation: ϕ =", power, "\n")

ag <- anscombe.test(x, alternative = "two.sided")

is_mesokurtic <- ag$p.value > alpha

# Testing for normality

library(nortest)

cat("Testing for normality by three tests with different rationales", "\n")

lillie.test(x)

lillie_power <- function(x, alpha, B = 1000) {n <- length(x)

p_values <- numeric(B)

for (i in 1:B) {bootstrap_sample <- sample(x, replace = TRUE)

result <- lillie.test(bootstrap_sample)

p_values[i] <- result$p.value}
```


Citation: De la Rubia JM (2024) Determination of the number and width of class intervals using R. Ann Environ Sci Toxicol 8(1): 022-042.
DOI: https://dx.doi.org/10.17352/aest.000077

```
power <- mean(p_values < alpha)

return(power)}

set.seed(123)

power <- lillie_power(x, alpha)

cat("Statistical power for the Lilliefors normality test using bootstrap simulation: ϕ =», power, «\n»)

ad.test(x)

ad_power <- function(x, alpha, B = 1000) {n <- length(x)

p_values <- numeric(B)

for (i in 1:B) {bootstrap_sample <- sample(x, replace = TRUE)

result <- ad.test(bootstrap_sample)

p_values[i] <- result$p.value}

power <- mean(p_values < alpha)

return(power)}

set.seed(123)

power <- ad_power(x, alpha)

cat("Statistical power for the Anderson-Darling normality test using bootstrap simulation: ϕ =», power, «\n»)

sf.test(x)

sf_power <- function(x, alpha, B = 1000) {n <- length(x)

p_values <- numeric(B)

for (i in 1:B) {bootstrap_sample <- sample(x, replace = TRUE)

result <- sf.test(bootstrap_sample)

p_values[i] <- result$p.value}

power <- mean(p_values < alpha)

return(power)}

set.seed(123)

power <- sf_power(x, alpha)

cat("Statistical power for the Shapiro-Francia normality test using bootstrap simulation: ϕ =", power, "\n")

lillie <- lillie.test(x)

ad <- ad.test(x)

sf <- sf.test(x)

is_normal <- lillie$p.value > alpha | ad$p.value > alpha | sf$p.value > alpha

if (is_symmetric && is_mesokurtic && is_normal) {normality <- 2

} else if (is_symmetric && is_mesokurtic) {normality <- 1

} else {normality <- 0}

interpretation <- c("No, at 5% significance level", "Ambiguous. The hypotheses of symmetry and mesokurtosis, but not normality, hold at 5% significance level.", "Yes, at 5% significance level.")

cat("Normality:", normality, "=", interpretation[normality + 1], "\n")

# Summary table of the eight uniform width rules

rule_table1 <- data.frame(Rule = c("Square root", "Rice University", "Sturges", "Doane", "Scott", "Freedman-Diaconis","Rudemo", "Shimazaki-Shinomoto"),

k = c(k_sqrt, k_Rice, k_Sturges, k_Doane, k_Scott, k_FD, k_Rudemo, k_SS),

w = c(round(w_sqrt, 4), round(w_Rice, 4), round(w_Sturges, 4), round(w_Doane, 4), round(w_Scott, 4), round(w_FD, 4), round(w_Rudemo, 4), round(w_SS, 4)),

n_IC = c(n_sqrt, n_Rice, n_Sturges, n_Doane, n_Scott, n_FD, n_Rudemo, n_SS))

cat("Table 1: Summary of number, width, and frequency per class interval", "\n")

print(rule_table1)

cat("Note. k = number of class intervals, w = uniform width per interval, and n_IC = absolute frequency per interval.", "\n")

#Selection of the number and uniform width of the intervals by convergence of the results

cat("Rule selection by convergence of results (Table 1)", "\n")

library(modeest)
```

Table 1: Summary of number, width, and frequency per class interval.

Rule	k	W	n_IC
Square root	8	1.1862	variable
Rice University	8	1.1862	variable
Sturges	7	1.3557	variable
Doane	7	1.3557	variable
Scott	5	2.0738	variable
Freedman-Diaconis	6	1.8274	variable
Rudemo	3	3.8400	variable
Shimazaki-Shinomoto	3	3.1633	variable

Note. k = number of class intervals, w = uniform width per interval, and n_IC = absolute frequency per interval.

Citation: De la Rubia JM (2024) Determination of the number and width of class intervals using R. Ann Environ Sci Toxicol 8(1): 022-042.
DOI: https://dx.doi.org/10.17352/aest.000077

```
mod_k <- mlv(rule_table1$k, method = "mfv")

mod_w <- mlv(rule_table1$w, method = "mfv")

mode_k <- as.numeric(mod_k)

mode_w <- as.numeric(mod_w)

if (length(mode_k) == 1) {cat("Most frequent number of class intervals: k =", mode_k, "\n")

cat("Most frequent class interval width: w =", mode_w, "\n")

} else if (length(mode_k) > 1 && normality == 2) {mode_w <- w_Scott

mode_k <- k_Scott

cat("Since there is no unique mode, but the distribution is normal, Scott's rule is used: k =", mode_k, "and w =", mode_w, "\n")

} else if (length(mode_k) > 1 && normality < 2) {mode_w <- w_FD

mode_k <- k_FD

cat("Since there is neither a unique mode nor normality, the Freedman-Diaconis rule is used: k =", mode_k, "and w =", mode_w, "\n")}

#cat("Since the sample was drawn from an arcsine distribution, square root rule is used: k =", k_sqrt, "and w =", w_sqrt, "\n")

#mode_k <- k_sqrt

#mode_w <- w_sqrt

# Frequency distribution table (convergence of uniform width rules)

intervals <- seq(min(x), max(x), mode_w)

if (max(intervals) < max(x)) {intervals <- c(intervals, max(intervals) + mode_w)}

interval_labels <- paste0("[", round(intervals[-length(intervals)], 3), ", ", round(intervals[-1], 3), ")")

interval_labels[length(interval_labels)] <- paste0("[", round(intervals[length(intervals) - 1], 3), ", ", round(max(x), 3), "]")

frequencies <- table(cut(x, breaks = intervals, include.lowest = TRUE, right = FALSE))

x_i <- round((intervals[-length(intervals)] + intervals[-1]) / 2, 3)

n_i <- as.vector(frequencies)

f_i <-round(n_i / length(x), 4)

p_i <- paste0(round(f_i * 100, 1), "%")

N_i <- cumsum(n_i)

F_i <- round(cumsum(f_i), 4)

P_i <- paste0(round(F_i * 100, 1), "%")

table <- data.frame(Interval = interval_labels, x_i, n_i, f_i, p_i, N_i, F_i, P_i)

cat("Table 2: Frequency distribution", "\n")

print(table)

cat("Note. x_i = class mark, n_i = absolute frequency, f_i = relative frequency, p_i = percentage, ", "\n")

cat("N_i = cumulative absolute frequency, F_i = cumulative relative frequency, and P_i = cumulative percentage.", "\n")

# Histogram with overlaid density and normal curves

library(ggplot2)

library(scales)

density <- density(x, kernel = "epanechnikov", bw = "SJ")

x_values <- seq(mean(x) - 4 *sd(x), mean(x) + 4 *sd(x), length = 1000)

y_values <- dnorm(x_values, mean = mean(x), sd = sd(x))

intervals <- seq(min(x), max(x), length.out = mode_k + 1)

histogram <- ggplot(data = data.frame(x = x), aes(x = x)) +

geom_histogram(aes(y = after_stat(density)), binwidth = NULL, breaks = intervals, fill = "darkolivegreen2", color = "black") +

geom_line(data = data.frame(x = density$x, y = density$y), aes(x = x, y = y), color = "darkblue", linewidth = 1) +
```

Table 2: Frequency distribution.

i	Interval	x_i	n_i	f_i	p_i	N_i	F_i	P_i
1	[0.8, 2.874)	1.837	8	0.1569	15.7%	8	0.1569	15.7%
2	[2.874, 4.948)	3.911	11	0.2157	21.6%	19	0.3726	37.3%
3	[4.948, 7.021)	5.984	18	0.3529	35.3%	37	0.7255	72.6%
4	[7.021, 9.095)	8.058	12	0.2353	23.5%	49	0.9608	96.1%
5	[9.095, 10.29]	10.132	2	0.0392	3.9%	51	1	100%

Note. x_i = class mark, n_i = absolute frequency, f_i = relative frequency, p_i = percentage, N_i = cumulative absolute frequency, F_i = cumulative relative frequency, and P_i = cumulative percentage.

Citation: De la Rubia JM (2024) Determination of the number and width of class intervals using R. Ann Environ Sci Toxicol 8(1): 022-042.
DOI: https://dx.doi.org/10.17352/aest.000077

```
geom_line(data = data.frame(x = x_values, y = y_values), aes(x = x, y = y), color = "red", linewidth = 1) +

labs(x = "X values", y = "Density") +

theme(panel.background = element_rect(fill = "white"), axis.text.x.bottom = element_text(size = 8), axis.text.y = element_text(size = 8), axis.title.x = element_text(size = 9), axis.title.y = element_text(size = 9), axis.line = element_line(color = "black")) +

scale_y_continuous(labels = label_number(accuracy = 0.01))

jpeg("Histogram1.jpeg", width = 800, height = 600, units = "px", res = 300)

print(histogram)

dev.off()

histogram

#Optimal histogram

par(mar = c(5, 6, 4, 2) + 0.1)

hist(x, bincol=NULL, main = NULL, xlab = "X values", ylab = "Frequency", col="darkolivegreen2", border = "black", lwd = 2.5, cex.axis = 2.5, cex.lab = 2.5)

# Rule of the multiple of 5 (proposed in this article)

n <- length(x)

n_I <- floor(n / 5)

w_I <- "variable"

k_I <- floor(n / n_I)

exc <- n - k_I * n_I

# Mann-Wald Rule

k_MW_min <- round(2*((2*n^2) / qnorm(0.05)^2)^(1/5), 0)

k_MW_max <- round(4*((2*n^2) / qnorm(0.05)^2)^(1/5), 0)

w_MW_min <- "variable"

w_MW_max <- "variable"

n_MW_min <- floor(n / k_MW_min)

n_MW_max <- floor(n / k_MW_max)

exc_MW_min <- n - k_MW_min * n_MW_min

exc_MW_max <- n - k_MW_max * n_MW_max

# Function to obtain prime factors

prime_factors <- function(n) {if (n <= 1) {return(NULL)}

factors <- c()

d <- 2

while (n > 1) {while ((n %% d) == 0) {factors <- c(factors, d)

n <- n / d}

d <- d + 1

if (d * d > n) {if (n > 1) {factors <- c(factors, n)}

break}}

return(factors)}

# Function to generate all possible products of the factor vector

generate_products <- function(factors) {len <- length(factors)

products <- c()

for (i in 1:(2^len - 1)) {selected_factors <- factors[as.logical(intToBits(i)[1:len])]

product <- prod(selected_factors)

products <- c(products, product)}

return(unique(products))}

# Function to obtain the factor closest to the lower limit that gives a uniform frequency of at least 5 cases and no excess.

find_k_MW <- function(n, k_MW_min) {factors <- prime_factors(n)

if (is.null(factors)) {return(NA)}

products <- generate_products(factors)

valid_products <- products[products >= k_MW_min & floor(n / products) >= 5 & n %% products == 0]

if (length(valid_products) == 0) {return(NA)}

return(min(valid_products))}

k_MW <- find_k_MW(n, k_MW_min)

if (is.na(k_MW)) {k_MW <- "NC"

w_MW <- "NC"

n_MW <- "NC"

exc_MW <- "NC"} else {w_MW <- "variable"

n_MW <- floor(n / k_MW)

exc_MW <- n - k_MW * n_MW}

# Moore's Rule
```


Citation: De la Rubia JM (2024) Determination of the number and width of class intervals using R. Ann Environ Sci Toxicol 8(1): 022-042.
DOI: https://dx.doi.org/10.17352/aest.000077

```
k_Moore <- round(2 * n^0.4, 0)

w_Moore <- "variable"

n_Moore <- floor(n / k_Moore)

exc_Moore <- n - k_Moore * n_Moore

central_int <- function(exc) {if (exc == 0) {return("0 in central CIs")}

else if (exc == 1) {return("1 datum in the central CI")}

else {return(paste("1 datum in the ", exc, " central CIs", sep = ""))}}

rule_table2 <- data.frame(Rule = c("Chi-2 test (24 < n < 40)", "Mann-Wald_min (n > 39)", "Mann-Wald (n > 39)", "Mann -Wald_max (n > 39)", "Moore (n > 39)"),

k = c(k_I, k_MW_min, k_MW, k_MW_max, k_Moore),

w_i = c(w_I, w_MW_min, w_MW, w_MW_max, w_Moore),

n_CI = c(paste(n_I, "&", central_int(exc)), paste(n_MW_min, "&", central_int(exc_MW_min)), paste(n_MW, "&", central_int(exc_MW)), paste(n_MW_max, "&", central_int(exc_MW_max)), paste(n_Moore, "&", central_int(exc_Moore))))

cat("Table 3: Summary of the number of class intervals, their widths, and uniform and excess frequency ", "\n")

print(rule_table2)

cat("Note. k = number of class intervals, w_i = variable width per interval, and n_IC = absolute frequency per class interval (NC = Non-Compliance with the condition).", "\n")

# Table of class intervals with uniform frequency and variable widths and testing for normality using g-test

# Input values provided by the user (choose rule: multiple of 5, MW, Moore)

k <- 10 # Number of class intervals

n_I <- 5 # Uniform frequency per interval

n_exc <- 1 # Excess frequency (one element per center interval)

cat("Criterion. Moore:", "k =", k, ",", "n_I =", n_I, "y", "n_exc =", n_exc, "\n") # Indicate criterion

x_sorted <- sort(x)

n <- length(x)

interval_limits <- numeric(k)

central_intervals <- (k - n_exc) / 2

for (i in 1:k) {if (i <= central_intervals || i > (central_intervals + n_exc)) {

lower_limit <- x_sorted[(i - 1) * n_I + 1]

upper_limit <- x_sorted[min(i * n_I, n)]} else {lower_limit <- x_sorted[(i - 1) * n_I + 1]

upper_limit <- x_sorted[min((i - 1) * n_I + n_I + 1, n)]}

interval_limits[i] <- upper_limit}

frequency_table <- data.frame(Interval = character(k), n_o = numeric(k), n_e = numeric(k), g_i = numeric(k), stringsAsFactors = FALSE)

for (i in 1:k) {if (i == 1) {lower_limit <- min(x_sorted)

F_z_LS_prev <- 0} else {lower_limit <- interval_limits[i - 1]

F_z_LS_prev <- pnorm(interval_limits[i - 1], mean = mean(x), sd = sd(x))}

upper_limit <- interval_limits[i]

F_z_LS <- pnorm(upper_limit, mean = mean(x), sd = sd(x))

if (i == k) F_z_LS <- 1

if (i == 1) {interval_label <- pasteo("[", round(lower_limit, 2), ", ", round(upper_limit, 2), ")")} else if (i == k) {interval_label <- pasteo("(", round(lower_limit, 2), ", ", round(upper_limit, 2), "]")} else {interval_label <- pasteo("(", round(lower_limit, 2), ", ", round(upper_limit, 2), ")")}

n_o <- if (i > central_intervals && i <= (central_intervals + n_exc)) {n_I + 1} else {n_I}

n_e <- n * (F_z_LS - F_z_LS_prev)

g_i <- if (n_e == 0) 0 else n_o * log(n_o / n_e)

frequency_table[i, ] <- c(interval_label, n_o, n_e, g_i)}

frequency_table$n_o <- as.numeric(frequency_table$n_o)

frequency_table$n_e <- as.numeric(frequency_table$n_e)

frequency_table$g_i <- as.numeric(frequency_table$g_i)

sums <- c("Sum", sum(frequency_table$n_o), sum(frequency_table$n_e), sum(frequency_table$g_i))

frequency_table <- rbind(frequency_table, sums)

cat("Table 4: Class intervals with uniform frequency, variable widths, and calculations for the G-test", "\n")

print(frequency_table)

cat("Note. n_o = observed frequency, n_e = expected
```


Citation: De la Rubia JM (2024) Determination of the number and width of class intervals using R. Ann Environ Sci Toxicol 8(1): 022-042.
DOI: https://dx.doi.org/10.17352/aest.000077

```
frequency, and g_i = n_o * ln(n_o / n_e).", "\n")

    # G-test for normality (Woolf, 1957) using Williams' continuity correction.

    cat("Likelihood ratio test or G-test for normality", "\n")

    g <- 2 * sum(as.numeric(frequency_table$g_i[-(k + 1)]))

    p <- pchisq(g, df = k - 3, lower.tail = FALSE)

    power_g <- 1 - pchisq(qchisq(alpha, df = k - 3, lower.tail = FALSE), df = k - 3, ncp = g, lower.tail = TRUE, log.p = FALSE)

    q <- 1 + (k^2 - 1) / (6 * n * (k - 3))

    g_cc <- g / q

    p_c <- pchisq(g_cc, df = k - 3, lower.tail = FALSE)

    power_g_cc <- 1 - pchisq(qchisq(alpha, df = k - 3, lower.tail = FALSE), df = k - 3, ncp = g_cc, lower.tail = TRUE, log.p = FALSE)

    cat("G-test statistic: g = 2 * sum(g_i) =", round(g, 4), "\n")

    cat("Asymptotic p-value for G-test statistic: p =", round(p, 6), "\n")

    if (p < alpha) {cat(sprintf("The null hypothesis of normality is rejected at a significance level of %.2f with the G-test statistic.", alpha), "\n")

    } else {cat(sprintf("The null hypothesis of normality is maintained at a significance level of %.2f with the G-test statistic.", alpha), "\n")}

    cat("The right-tailed statistical power for the alternative hypothesis of non-normality for the G-test: ϕ =", round(power_g, 4), "\n")

    cat("Williams' continuity correction: q =", round(q, 4), "\n")

    cat("Williams' continuity-corrected G-test statistic: g_cc = g / q =", round(g_cc, 4), "\n")

    cat("Asymptotic p-value for continuity-corrected G-test statistic: p =", round(p_c, 6), "\n")

    if (p_c < alpha) {cat(sprintf("The null hypothesis of normality is rejected at a significance level of %.2f with the continuity-corrected G-test statistic.", alpha), "\n")

    } else {cat(sprintf("The null hypothesis of normality is maintained at a significance level of %.2f with the continuity-corrected G-test statistic.", alpha), "\n")}

    cat("The right-tailed statistical power for the alternative hypothesis of non-normality for the G-test with Williams' continuity correction: ϕ =", round(power_g_cc, 4), "\n")
```

Example 1: random sample drawn from a normal distribution: In two-tailed tests at a 5% significance level, the sample of 51 data points from the script can be considered random by the Wald-Wolfowitz test [57], symmetric (H_0: β_1 = 0) by the D'Agostino test [49], mesokurtic (H_0: β_2 = 3) by the Anscombe-Glynn test [50], and normally distributed according to the following tests: Lilliefors' D [51], Anderson-Darling's A^2 [52], Royston's W [53], and Woolf's G [61]. There is no unique mode, but there is normality, so Scott's rule [19] is used (Table 1). The histogram with the overlaid density and normal curves can be seen in Figure 1, and the frequency distribution in Table 2. The histogram and density curve reveal a normal profile.

Figure 2 shows Kreider's optimal histogram [14] with a similar profile, but with more bins, which allows for a worse appreciation of the normal bell shape. Moore's rule [26] (Table 3) is chosen to construct the frequency table and compute the G-test [61] (Table 4) since the sample size is larger than 39 and the Mann-Wald rule [25] does not yield an integer without excess. Moore's rule gives 10 class intervals with a common frequency of 5 and an excess frequency of 1, located in the fifth interval.

```
Wald-Wolfowitz runs test. Criterion: median

Number of runs: r = 28

n_0 = #(x_i < mdn(x)) = 25 and n_1 = #(x_i > mdn(x)) = 25

n = n_0 + n_1 = 50
```

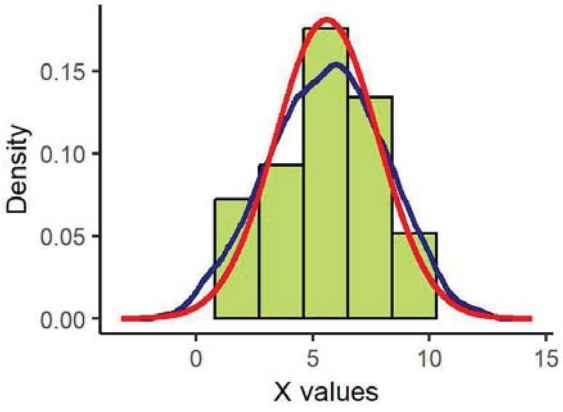

Figure 1: Histogram with overlaid density and normal curves (Scott's rule).

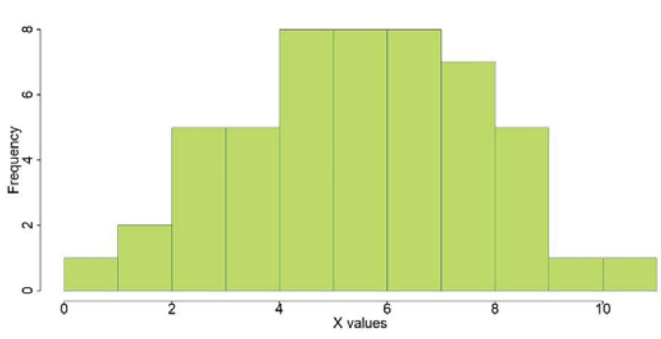

Figure 2: Kreider's optimal histogram.

Citation: De la Rubia JM (2024) Determination of the number and width of class intervals using R. Ann Environ Sci Toxicol 8(1): 022-042.
DOI: https://dx.doi.org/10.17352/aest.000077

Table 3: Summary of the number of class intervals, their widths, and uniform and excess frequency.

Rule	k	w_i	n_IC
Chi-2 test (24 < n < 40)	5	variable	10 & 1 datum in the central CI
Mann-Wald_min (n > 39)	9	variable	5 & 1 datum in the 6 central CIs
Mann-Wald (n > 39)	NC	NC	NC & 1 datum in the NC central CIs
Mann-Wald_max (n > 39)	18	variable	2 & 1 datum in the 15 central CIs
Moore (n > 39)	10	variable	5 & 1 datum in the central CI

Note. k = number of class intervals, w_i = variable width per interval, and n_IC = absolute frequency per class interval (NC = Non-Compliance with the condition).

Table 4: Class intervals with uniform frequency, variable widths, and calculations for the G-test.

i	Interval	n_o	n_e	g_i
1	[0.8, 2.36)	5	3.6135	1.6238
2	(2.36, 3.7]	5	6.3071	-1.1612
3	(3.7, 4.43]	5	5.2790	-0.2715
4	(4.43, 5]	5	4.8456	0.1569
5	(5, 5.83]	6	7.5927	-1.4125
6	(5.83, 6.01]	5	1.6439	5.5618
7	(6.01, 6.69]	5	5.9028	-0.8300
8	(6.69, 7.54]	5	6.1723	-1.0532
9	(7.54, 8.05]	5	2.8646	2.7850
10	(8.05, 9.93]	5	6.7785	-1.5216
	Sum	51	51	3.8775

Note. n_o = observed frequency, n_e = expected frequency, and g_i = n_o * ln(n_o / n_e).

Two-tailed exact probability: p = 0.6707

Mean: M(R|n_0, n_1) = 26 and Standard deviation: SD(R|n_0, n_1) = 3.4993

Standardized number of runs: z_r = 0.5715

Two-tailed asymptotic probability: p = 0.5676

Statistical power for the Wald-Wolfowitz runs test using bootstrap simulation: ϕ = 0

Statistical data for calculation of uniform width rules

Sample size: n = 51

Sample range: R(x) = max(x) - min(x) = 9.49

Sample interquartile range (quartiles by rule 8): IQR(x) = 3.3883

Skewness coefficient based on the standardized third central moment: b_1(x) = m_3 / (m_2)^(3/2) = -0.0693

Asymptotic standard error of b_1: se(b_1) = 0.3236

Sample standard deviation (with Bessel's correction): sd(x) = 2.2035

Testing for symmetry

D'Agostino skewness test

skew = -0.06933, z = -0.22483, p-value = 0.8221

Statistical power for the D'Agostino skewness test using bootstrap simulation: ϕ = 0.009

Testing for mesokurtosis

Anscombe-Glynn kurtosis test

kurt = 2.43281, z = -0.78376, p-value = 0.4332

Statistical power for the Anscombe-Glynn kurtosis test using bootstrap simulation: ϕ = 0.13

Testing for normality by three tests with different rationales

Lilliefors (Kolmogorov-Smirnov) normality test

D = 0.062539, p-value = 0.8875

Statistical power for the Lilliefors normality test using bootstrap simulation: ϕ = 0.174

Anderson-Darling normality test

A = 0.19672, p-value = 0.8846

Statistical power for the Anderson-Darling normality test using bootstrap simulation: ϕ = 0.153

Shapiro-Francia normality test

W = 0.99112, p-value = 0.9234

Statistical power for the Shapiro-Francia normality test using bootstrap simulation: ϕ = 0.073

Normality: 2 = Yes, at 5% significance level.

Rule selection by convergence of results (Table 1)

Since there is no unique mode, but the distribution is normal, Scott's rule is used: k = 5 and w = 2.073755

Likelihood ratio test or G-test for normality

G-test statistic: g = 2 * sum(g_i) = 7.755

Asymptotic p-value for G-test statistic: p = 0.3547

The null hypothesis of normality is maintained at a significance level of 0.05 with the G-test statistic.

The right-tailed statistical power for the alternative hypothesis of non-normality for the G-test: ϕ = 0.487

Williams' continuity correction: q = 1.0462

Williams' continuity-corrected G-test statistic: g_cc = g / q = 7.4124

Asymptotic p-value for continuity-corrected G-test statistic: p = 0.3872

The null hypothesis of normality is maintained at a significance level of 0.05 with the continuity-corrected G-test statistic.

Citation: De la Rubia JM (2024) Determination of the number and width of class intervals using R. Ann Environ Sci Toxicol 8(1): 022-042.
DOI: https://dx.doi.org/10.17352/aest.000077

The right-tailed statistical power for the alternative hypothesis of non-normality for the G-test with Williams' continuity correction: $\phi = 0.4662$

Example 2: random non-normal sample: x <- c(2.7, 4.71, 6.13, 8.22, 7.59, 6.08, 4.97, 6.4, 4.56, 4.24, 5.2, 6.8, 5.31, 4.83, 5.09, 4.56, 4.84, 6.84, 8.21, 4.82, 3.64, 6.5, 5.05, 6.28, 6.21, 6.62, 5.44, 4.13, 8.93, 4.59, 7.84, 4.47, 2.91, 1.28, 4.62, 6.69, 5, 6.7, 4.86, 9.57, 7, 4.48, 4.13, 5.32, 4.94, 6.3, 10.05, 4.56, 1.24, 4.61, 4.24)

In two-tailed tests at a 5% significance level, the sample of 51 data points can be considered random by the Wald-Wolfowitz test (r = 27, p_{exact} = 0.8843, z(r) = 0.2858, p_{asymp} = 0.7751, ϕ = 0), symmetric by the D'Agostino test (b_1 = 0.23270, z = 0.74847, p-value = 0.4542, ϕ = 0.152), and mesokurtic by the Anscombe-Glynn test (b_2 = 3.6156, z = 1.2594, p-value = 0.2079, ϕ = 0.087). However, it does not conform to a normal distribution according to the tests of Lilliefors (D = 0.1275, p-value = 0.0376, ϕ = 0.837), Anderson-Darling (A^2 = 0.9619, p-value = 0.0141, ϕ = 0.876), Royston (W = 0.9541, p-value = 0.0466, ϕ = 0.756), and Woolf (g = 19.9095, p = 0.0058, ϕ = 0.9266; g_cc = 19.03, p = 0.0081, ϕ = 0.9131). There is convergence in the number of intervals to 8 by the square root [15], Rice [16], and Doane [18] rules, with a uniform width of 1.1013 (Table 5). See the histogram with the overlaid density and normal curves in Figure 3 and the frequency distribution in Table 6. The histogram and density curve reveals a bimodal profile, with a higher mode located around 5 and a lower mode around 7.

Figure 4 shows Kreider's optimal histogram [14] with a very similar profile. As in the previous example, Moore's rule [26] is chosen to construct the frequency table and compute the G-test (Table 7), since the sample size is the same (Table 4). Moore's rule [26] gives 10 class intervals with a common frequency of 5 and an excess frequency of 1, located in the fifth interval.

A simplified script for small sample sizes: What should be done with a small sample size of less than 25? In this case, only the first part of the script would apply. The number of class intervals can be determined using the square root rule or Rice University's rule. With both rules, the uniform width is calculated by dividing the range by the number of class intervals.

Table 5: Summary of number, width, and frequency per class interval.

Rule	k	W	n_IC
Square root	8	1.1013	variable
Rice University	8	1.1013	variable
Sturges	7	1.2586	variable
Doane	8	1.1013	variable
Scott	6	1.6904	variable
Freedman-Diaconis	9	1.1002	variable
Rudemo	10	0.9625	variable
Shimazaki-Shinomoto	11	0.8009	variable

k = number of class intervals, w = uniform width per interval, and n_IC = absolute frequency per interval.

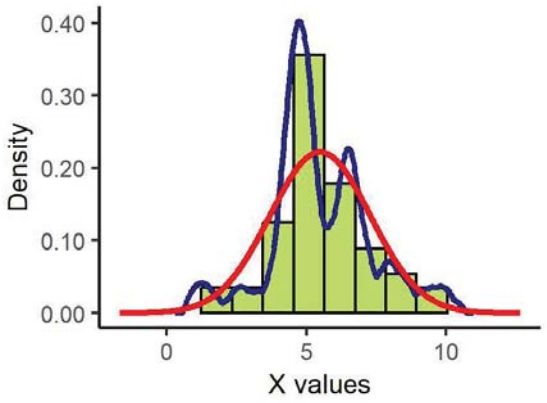

Figure 3: Histogram with overlaid density and normal curves (convergence of criteria).

When testing for randomness with the Wald-Wolfowitz runs test, exact probabilities must be applied. For normality testing, the Shapiro-Wilk test can be used. The D'Agostino skewness test can be used with a sample size greater than 8, and the Anscombe-Glynn kurtosis test with a sample size of at least 20. Additionally, the significance level should be increased to 0.1.

Below is a script that follows these suggestions. It is applied to a sample generated by inverse transformation sampling from an arcsine distribution with threshold parameters: a = 0.5 and b = 2. The distribution is modified to have a major mode on the left and a minor mode on the right within its bounded range. These data, to give them specific context, represent the average diameter of the cavernous artery (measured in millimeters) of the penis in 16 men subjected to visual stimuli with homosexual content.

```
# Vector of scores

x <- c(0.657, 0.687, 0.513, 1.179, 1.976, 1.611, 0.571, 0.604, 1.482, 0.524, 1.746, 0.754, 2, 1.843, 0.501, 0.889)

cat("Rules for determining the number of class intervals (k) and uniform width (w)", "\n")

# Square root rule

k_sqrt <- ceiling(sqrt(length(x)))

w_sqrt <- (max(x) - min(x)) / k_sqrt

cat("Square root rule: k =", k_sqrt, "w =", w_sqrt, "\n")

# Rice University Rule

k_Rice <- ceiling(2 * length(x)^(1/3))

w_Rice <- (max(x) - min(x)) / k_Rice

cat("Rice University Rule: k =", k_Rice, "w =", w_Rice, "\n")

# Testing for randomness

library(randtests)
```


Citation: De la Rubia JM (2024) Determination of the number and width of class intervals using R. Ann Environ Sci Toxicol 8(1): 022-042.
DOI: https://dx.doi.org/10.17352/aest.000077

Table 6: Frequency distribution.

i	Interval	x_i	n_i	f_i	p_i	N_i	F_i	P_i
1	[1.24, 2.341)	1.791	2	0.0392	3.9%	2	0.0392	3.9%
2	[2.341, 3.443)	2.892	2	0.0392	3.9%	4	0.0784	7.8%
3	[3.443, 4.544)	3.993	7	0.1373	13.7%	11	0.2157	21.5%
4	[4.544, 5.645)	5.095	20	0.3922	39.2%	31	0.6079	60.7%
5	[5.645, 6.746)	6.196	10	0.1961	19.6%	41	0.8040	80.3%
6	[6.746, 7.848)	7.297	5	0.0980	9.8%	46	0.9020	90.1%
7	[7.848, 8.949)	8.398	3	0.0588	5.9%	49	0.9608	96%
8	[8.949, 10.05]	9.500	2	0.0392	3.9%	51	1	100%

Note. x_i = class mark, n_i = absolute frequency, f_i = relative frequency, p_i = percentage, N_i = cumulative absolute frequency, F_i = cumulative relative frequency, and P_i = cumulative percentage.

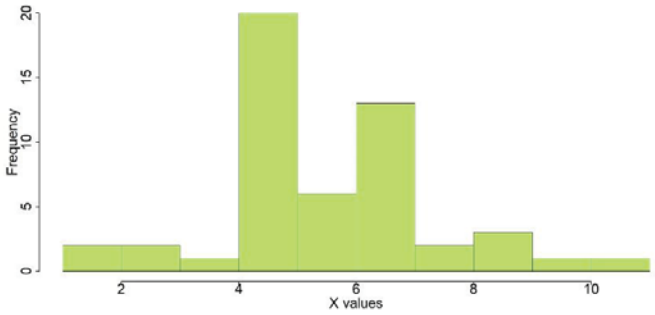

Figure 4: Kreider's optimal histogram.

Table 7: Class intervals with uniform frequency, variable widths, and calculations for the G-test.

i	Interval	n_o	n_e	g_i
1	[1.24, 3.64)	5	7.7262	-2.1759
2	(3.64, 4.47]	5	6.8103	-1.5450
3	(4.47, 4.59]	5	1.1781	7.2276
4	(4.59, 4.83]	5	2.4723	3.5215
5	(4.83, 5.05]	6	2.3764	5.5570
6	(5.05, 5.32]	5	3.0117	2.5346
7	(5.32, 6.28]	5	10.5910	-3.7528
8	(6.28, 6.69]	5	3.9778	1.1435
9	(6.69, 7.59]	5	6.6749	-1.4446
10	(7.59, 9.57]	5	6.1813	-1.0604
	Sum	51	51	10.0055

Note. n_o = observed frequency, n_e = expected frequency, and g_i = n_o * ln(n_o / n_e).

```
result_ww <- runs.test(x, alternative = "two.sided", threshold = median(x), pvalue = 'exact')

print(result_ww)

alpha <- 0.1

if (result_ww$p.value < alpha) {cat(sprintf("The null hypothesis of randomness is rejected at a significance level of %.2f using Wald-Wolfowitz runs test.", alpha), "\n")

} else {cat(sprintf("The null hypothesis of randomness is maintained at a significance level of %.2f using Wald-Wolfowitz test.", alpha), "\n")}

ww_power <- function(x, alpha, B = 1000) {n <- length(x)

p_values <- numeric(B)

for (i in 1:B) {

bootstrap_sample <- sample(x, replace = TRUE)

result <- runs.test(x, alternative = "two.sided", threshold = median(x), pvalue = 'exact')

p_values[i] <- result$p.value}

power <- mean(p_values < alpha)

return(power)}

set.seed(123)

power <- ww_power(x, alpha)

cat("Statistical power for the Wald-Wolfowitz runs test using bootstrap simulation: ϕ =", power, "\n")

# Testing for symmetry

library(moments)

result_agostino <- agostino.test(x, alternative = "two.sided")

print(result_agostino)

if (result_agostino$p.value < alpha) {cat(sprintf("The null hypothesis of symmetry is rejected at a significance level of %.2f using D'Agostino skewness test.", alpha), "\n")

} else {cat(sprintf("The null hypothesis of symmetry is maintaineagostino_power <- function(x, alpha, B = 1000) {n <- length(x)

p_values <- numeric(B)

for (i in 1:B) {

bootstrap_sample <- sample(x, replace = TRUE)
```


Citation: De la Rubia JM (2024) Determination of the number and width of class intervals using R. Ann Environ Sci Toxicol 8(1): 022-042.
DOI: https://dx.doi.org/10.17352/aest.000077

```
result <- agostino.test(bootstrap_sample)

p_values[i] <- result$p.value}

power <- mean(p_values < alpha)

return(power)}

set.seed(123)

power <- agostino_power(x, alpha)

cat("Statistical power for the D'Agostino skewness test using bootstrap simulation: ϕ =", power, "\n")

# Testing for mesokurtosis

result_ag <- anscombe.test(x, alternative = "two.sided")

print(result_ag)

if (result_ag$p.value < alpha) {cat(sprintf("The null hypothesis of mesokurtosis is rejected at a significance level of %.2f using Anscombe-Glynn kurtosis test.", alpha), "\n")

} else {cat(sprintf("The null hypothesis of mesokurtosis is maintained at a significance level of %.2f using Anscombe-Glynn kurtosis test test.", alpha), "\n")}

ag_power <- function(x, alpha, B = 1000) {n <- length(x)

p_values <- numeric(B)

for (i in 1:B) {

bootstrap_sample <- sample(x, replace = TRUE)

result <- anscombe.test(bootstrap_sample, alternative = "two.sided")

p_values[i] <- result$p.value}

power <- mean(p_values < alpha)

return(power)}

set.seed(123)

power <- ag_power(x, alpha)

cat("Statistical power for the Anscombe-Glynn kurtosis test using bootstrap simulation: ϕ =", power, "\n")

# Testing for normality

library(nortest)

result_sw <- shapiro.test(x)

print(result_sw)

if (result_sw$p.value < alpha) {

cat(sprintf("The null hypothesis of normality is rejected at a significance level of %.2f using Shapiro-Wilk test.\n", alpha))

} else {cat(sprintf("The null hypothesis of normality is maintained at a significance level of %.2f using Shapiro-Wilk test.\n", alpha))}

shapiro_power <- function(x, alpha, B = 1000) {n <- length(x)

p_values <- numeric(B)

for (i in 1:B) {

bootstrap_sample <- sample(x, replace = TRUE)

result <- shapiro.test(bootstrap_sample)

p_values[i] <- result$p.value}

power <- mean(p_values < alpha)

return(power)}

set.seed(123)

power <- shapiro_power(x, alpha)

cat("Statistical power for the Shapiro-Wilk normality test using bootstrap simulation: ϕ =", power, "\n")

# Selection of the number and uniform width of the intervals

k <- 6

w <- (24983 - 2498) / 90000

# Frequency distribution table

intervals <- seq(min(x), max(x), w)

if (max(intervals) < max(x)) {intervals <- c(intervals, max(intervals) + w)}

interval_labels <- paste0("[", round(intervals[-length(intervals)], 3), ", ", round(intervals[-1], 3), ")")

interval_labels[length(interval_labels)] <- paste0("[", round(intervals[length(intervals) - 1], 3), ", ", round(max(x), 3), "]")

frequencies <- table(cut(x, breaks = intervals, include.lowest = TRUE, right = FALSE))

x_i <- round((intervals[-length(intervals)] + intervals[-1]) / 2, 3)

n_i <- as.vector(frequencies)

f_i <-round(n_i / length(x), 4)

p_i <- paste0(round(f_i * 100, 1), "%")

N_i <- cumsum(n_i)

F_i <- round(cumsum(f_i), 4)

P_i <- paste0(round(F_i * 100, 1), "%")

table <- data.frame(Interval = interval_labels, x_i, n_i,
```


Citation: De la Rubia JM (2024) Determination of the number and width of class intervals using R. Ann Environ Sci Toxicol 8(1): 022-042.
DOI: https://dx.doi.org/10.17352/aest.000077

```
f_i, p_i, N_i, F_i, P_i)

    cat("Table: Frequency distribution", "\n")

    print(table)

    cat("Note. x_i = class mark, n_i = absolute frequency, f_i = relative frequency, p_i = percentage, ", "\n")

    cat("N_i = cumulative absolute frequency, F_i = cumulative relative frequency, and P_i = cumulative percentage.", "\n")

    # Histogram with overlaid density and normal curves

    library(ggplot2)

    library(scales)

    density <- density(x, kernel = "epanechnikov", bw = "SJ")

    x_values <- seq(mean(x) - 4 *sd(x), mean(x) + 4 *sd(x), length = 1000)

    y_values <- dnorm(x_values, mean = mean(x), sd = sd(x))

    intervals <- seq(min(x), max(x), length.out = k + 1)

    histogram <- ggplot(data = data.frame(x = x), aes(x = x)) +

    geom_histogram(aes(y = after_stat(density)), binwidth = NULL, breaks = intervals, fill = "darkolivegreen2", color = "black") +

    geom_line(data = data.frame(x = density$x, y = density$y), aes(x = x, y = y), color = "darkblue", linewidth = 1) +

    geom_line(data = data.frame(x = x_values, y = y_values), aes(x = x, y = y), color = "red", linewidth = 1) +

    labs(x = "X values", y = "Density") +

    theme(panel.background = element_rect(fill = "white"), axis.text.x.bottom = element_text(size = 8), axis.text.y = element_text(size = 8), axis.title.x = element_text(size = 9), axis.title.y = element_text(size = 9), axis.line = element_line(color = "black")) +

    scale_y_continuous(labels = label_number(accuracy = 0.01))

    jpeg("Histogram.jpeg", width = 800, height = 600, units = "px", res = 300)

    print(histogram)

    dev.off()

    histogram
```

At a 10% significance level, the sample sequence can be considered random by the Wald-Wolfowitz runs test (r = 10, n_1 = 8, n_2 = 8, p_exact = 0.810 > 0.1; statistical power: ϕ = 0) and symmetric by the D'Agostino test ($\sqrt{b_1}$ = 0.438, z = 0.897, p-value = 0.370 > 0.1, ϕ = 0.229). However, it presents platykurtosis according to the Anscombe-Glynn kurtosis test (b_2 = 1.513 < 3, z = -2.337, p-value = 0.019 < 0.1, ϕ = 0.637) and is far from normal by the Shapiro-Wilk test (w = 0.840, p-value = 0.010 < 0.1, ϕ = 0.989). The kurtosis test should be approached with caution, as it is recommended for a minimum sample size of 20. The number of class intervals by the square root rule is 4 (w = 0.37475), and by the Rice University rule, it is 6 (w = 0.24983 = 22485 / 90000 = 1499 / 6000). To construct the frequency table with class intervals (Table 8) and the histogram (Figure 5), the Rice University rule is followed, as it best reflects the generating distribution of the sample, although the square root rule also provides a good representation.

Discussion

The extensive three-section script developed in this article facilitates the calculation of eight uniform amplitude rules. On the one hand, four sample size-based rules are obtained: three classical rules [15, 17-18] and a more recent one from Rice University [16]. The latter simplifies two optimal width

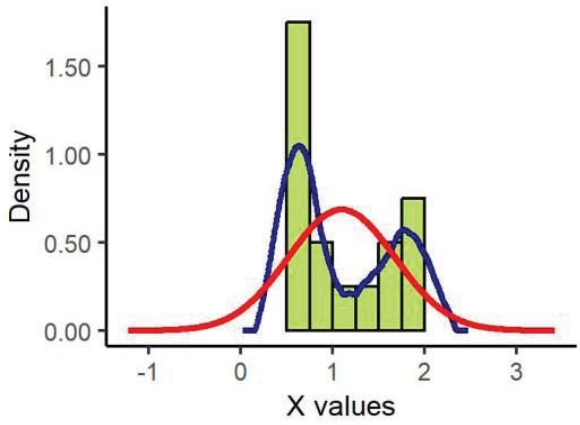

Figure 5: Histogram with overlaid density and normal curves.

Table 8: Frequency distribution.

i	Interval	x_i	n_i	f_i	p_i	N_i	F_i	P_i
1	[0.501, 0.751)	0.626	7	0.4375	43.8%	7	0.4375	43.8%
2	[0.751, 1.001)	0.876	2	0.1250	12.5%	9	0.5625	56.2%
3	[1.001, 1.25)	1.126	1	0.0625	6.2%	10	0.6250	62.5%
4	[1.25, 1.5)	1.375	1	0.0625	6.2%	11	0.6875	68.8%
5	[1.5, 1.75)	1.625	2	0.1250	12.5%	13	0.8125	81.2%
6	[1.75, 2]	1.875	3	0.1875	18.8%	16	1	100%

Note. x_i = class mark, n_i = absolute frequency, f_i = relative frequency, p_i = percentage, N_i = cumulative absolute frequency, F_i = cumulative relative frequency, and P_i = cumulative percentage.

Citation: De la Rubia JM (2024) Determination of the number and width of class intervals using R. Ann Environ Sci Toxicol 8(1): 022-042.
DOI: https://dx.doi.org/10.17352/aest.000077

rules, namely Scott's [19] and Freedman-Diaconis [20] rules. Moreover, four optimal width rules are computed [19-22].

It should be noted that Rudemo's rule [21] is not available in other programs, although the other three optimal amplitude rules are programmed in R: in the basic package via the histogram function (Scott and Freedman-Diaconis) and in the 'pracma' package (Shimazaki-Shinomoto). Rudemo's rule [21] minimizes the integrated mean square error from the leave-one-out cross-validation method. This function has multiple relative minima, hence the DEoptim: Global Optimization by Differential Evolution package is used. By default, the upper limit is set to the sample maximum, but a quarter of the maximum appears as a coded option (highlighted in blue), with the symbol # placed before the result, so that it is not read during the execution of the script. This can be activated when the width is very large and the number of bins very small compared to the Shimazaki-Shinomoto rule [22] and the other two optimization rules, as well as the optimal histogram provided by R from its basic package [14]. The # symbol is put in the previous 'result' and removed in the 'result' shown in blue. A quarter of the range, or approximately the semi-interquartile range, was considered, being an approximation of the standard deviation [58,59] and constituting a distance four times larger than Fisher's [60] proposal for the uniform width of class intervals, which is a quarter of the standard deviation.

For the selection of the number of class intervals, we chose to look for convergence of results among the eight uniform width rules and defined a uniform width bounded within the sample range (between sample minimum and maximum). In the case of non-convergence, the Scott rule [19] is chosen if there is normality, and the Freedman-Diaconis rule [20] is chosen if there is no normality. However, it should be noted that both rules perform poorly with the arcsine distribution [2]. For this distribution, it is advisable to use the square root rule [2], which is coded with the # symbol placed before it so that it is not read during the execution of the script. The distribution can be recognized through the density curve, which will show a U-shape. To change the mode, inactivate mode_k and mode_a by adding the symbol # to the two previous codes, and activate mode_w <- w_sqrt and mode_k <- k_sqrt by removing the symbol #.

There is a third option for determining the number of class intervals. Instead of using a uniform width, this approach leaves the width variable and chooses a uniform frequency when it is variable in the previous eight rules [62]. This third option has been proposed to check normality using Pearson's chi-square test or likelihood ratio test [25,26] with the intention of achieving a uniform frequency of at least five data points per interval with at least five class intervals [24]. Hence, it was separated into a third block, which includes a summary table with the data from these two rules, complemented by a new rule for samples of 25 to 39 data points, the frequency table with the calculations of the likelihood ratio test, and the results of this goodness-of-fit test with or without the Williams' continuity correction [28].

The new rule was named the "multiple of five" because it proposes dividing the sample size by 5 and rounding the result down, resulting in a uniform frequency if there is no surplus, or common frequency if there is surplus. The number of class intervals is the sample size divided by the uniform or common frequency, rounded down. If there is an excess frequency, one data point per interval is allocated to the central intervals. For sample sizes 25 to 29, there are five intervals with a frequency of 5 (there is a surplus with sample sizes 26 to 29); for sample sizes 30 to 34, there are six intervals with a frequency of 6 (there is a surplus with sample sizes 31 to 34); and for sample sizes 35 to 39, there are seven intervals with a frequency of 7 (there is a surplus with sample sizes 36 to 39). In all three cases, the surplus is from 1 to 4 elements. A similar solution is proposed for the Mann-Wald rule [25], seeking an integer, the one closest to the lower limit, that does not exceed the upper limit, that gives a uniform frequency of at least 5 cases, and that does not yield a surplus. If the sample size is greater than 39, it is suggested to use this option, and if it is not feasible, it is recommended to apply Moore's rule [26].

In the case of small samples (less than 25 data points), which is a common situation in the social sciences, a simplified script is presented with two recommended rules for determining the number of class intervals (square root and Rice University), which provide uniform widths. The inferential tests do not include the likelihood ratio test, but they do include the Wald-Wolfowitz runs test using exact probability, the D'Agostino skewness test, the Anscombe-Glynn kurtosis test, and the Shapiro-Wilk normality test instead of the Shapiro-Francia test, as these were developed for small samples.

Conclusion

The extensive three-section script can be used both practically and didactically at a descriptive and inferential level when working with quantitative variables and random samples of at least 25 data points. At the descriptive level, it provides two summary tables for nine rules to determine the number and width of class intervals. The first table is for the eight uniform width rules, and the second is for the three variable width rules. Additionally, it includes two frequency tables (one based on the convergence criterion for the eight uniform width rules and a uniform frequency table) and two histograms (one based on the convergence criterion and another using Kreider's method). At the inferential level, the script includes the density curve (using Epanechnikov's kernel with Sheather-Jones bandwidth) and tests for randomness (Wald-Wolfowitz), symmetry (D'Agostino), mesokurtosis (Anscombe-Glynn), and normality (Lilliefors' D, Anderson-Darling A^2, Royston's W, and Woolf's G).

It can be applied to any type of distribution: whether normal, as found in measures of general intelligence, ability, and expressive attitudes in open societies, as in the first example of the Result section; with positive skewness, as found in measures of psychopathology or stigmatized attitudes; with negative skewness, as observed in measures of academic achievement, prosocial behaviors, or normative attitudes in closed societies; or bimodal, as found in measures of sexual orientation in men (higher mode in exclusively heterosexual

Citation: De la Rubia JM (2024) Determination of the number and width of class intervals using R. Ann Environ Sci Toxicol 8(1): 022-042.
DOI: https://dx.doi.org/10.17352/aest.000077

orientation and lower mode in exclusively homosexual orientation) or political ideology in radicalized democratic societies due to strong socio-economic crises (left-wing vs. right-wing ideology). Other distributional forms can also be observed, as shown in the second example of the Results section, which is very common in the social and health sciences.

A simplified script is also provided for small samples that are common in social research. This script is a condensed version of the extensive three-section script, focusing on the first section. It includes the rules (square root and Rice University) and recommended inferential tests (Wald-Wolfowitz runs test using exact probability, D'Agostino skewness test, Anscombe-Glynn kurtosis test, and Shapiro-Wilk normality test) for small samples. It provides a frequency table with class intervals and a histogram with overlaid density and normal curves.

Acknowledgement

The author expresses gratitude to the reviewers and editor for their helpful comments.

References

1. Peck R, Short T, Olsen C. Introduction to statistics and data analysis (6th ed.). Cengage Learning, Boston, MA, 2020.

2. Moral J. Rice University rule to determine the number of bins. Open Journal of Statistics 2024; 14(1): 119-149. https://doi.org/10.4236/ojs.2024.141006

3. Sahann R, Müller T, Schmidt J. Histogram binning revisited with a focus on human perception. In: Proceeding of the 2021 IEEE Visualization Conference (VIS). Institute of Electrical and Electronics Engineers (IEEE), New Orleans, LA. 2021; 66-70. https://doi.org/10.1109/VIS49827.2021.9623301

4. Cooksey RW. Descriptive statistics for summarising data. In: Illustrating Statistical Procedures: Finding Meaning in Quantitative Data. Springer, Singapore. 2020; 61-139. https://doi.org/10.1007/978-981-15-2537-7_5

5. DeVellis RF, Thorpe CT. Scale development: Theory and applications. Sage publications, Thousand Oaks, CA, 2021.

6. Fang JQ. (Ed.). Statistical methods for biomedical research. World Scientific Publication Co., Singapore, 2021. https://doi.org/10.1142/12060

7. Terrell SR. Statistics translated: A step-by-step guide to analyzing and interpreting data (2nd ed.). Guilford Publications, NewYork, 2021.

8. Mustafy T, Rahman MTU. Excel. In: Statistics and Data Analysis for Engineers and Scientists. Transactions on Computer Systems and Networks. Springer, Singapore. 2024a; 81-134. https://doi.org/10.1007/978-981-99-4661-7_3

9. Mustafy T, Rahman MTU. MATLAB. In: Statistics and Data Analysis for Engineers and Scientists. Transactions on Computer Systems and Networks. Springer, Singapore, 2024b; 37-80. https://doi.org/10.1007/978-981-99-4661-7_2

10. Venables WN, Smith DM, The R Core Team. An introduction to R. Notes on R: a programming environment for data analysis and graphics. Version 4.4.0. 2024. https://cran.r-project.org/doc/manuals/R-intro.pdf

11. Braun WJ, Murdoch DJ. A first course in statistical programming with R (3rd ed.). Cambridge University Press, Cambridge, UK. 2021. https://doi.org/10.1017/9781108993456

12. Isvoranu AM, Epskamp S, Waldorp L, Borsboom D. (Eds.). Network psychometrics with R: A guide for behavioral and social scientists. Routledge, New York. 2022. https://doi.org/10.4324/9781003111238

13. Sievert C. Interactive web-based data visualization with R, plotly, and shiny. Chapman and Hall/CRC, London. 2020. https://doi.org/10.1201/9780429447273

14. Kreider G. Package 'optbin'. Rdocumentation. 2023. https://cran.r-project.org/web/packages/optbin/optbin.pdf

15. Pearson K. The grammar of science. Walter Scott Publishing Co., London. 1892. https://doi.org/10.1037/12962-000

16. Lane DM. Histograms. In: Online statistics education: a multimedia course of study. Houston, TX: Department of Statistics, Rice University. 2015. https://stats.libretexts.org/Bookshelves/Introductory_Statistics/Book%3A_Introductory_Statistics_(Lane)/02%3A_Graphing_Distributions/2.04%3A_Histograms

17. Sturges HA. The choice of a class interval. J Am Stat Assoc. 1926; 21(153): 65-66. https://doi.org/10.1080/01621459.1926.10502161

18. Doane DP. Aesthetic frequency classification. Am Stat. 1976; 30:181-183. https://doi.org/10.1080/00031305.1976.10479172

19. Scott DW. On optimal and data-based histograms. Biometrika. 1979; 66(3): 605-610. https://doi.org/10.1093/biomet/66.3.605

20. Freedman D, Diaconis P. On the histogram as a density estimator: L_2 theory. Probab Theory Relat Fields. 1981; 57(4): 453-476. https://doi.org/10.1007/BF01025868

21. Rudemo M. Empirical choice of histograms and kernel density estimators. Scand J Stat. 1982; 9(2): 65-78. https://www.jstor.org/stable/4615859

22. Shimazaki H, Shinomoto S. A method for selecting the bin size of a time histogram. Neural Comput. 2007 Jun; 19(6):1503-27. doi: 10.1162/neco.2007.19.6.1503. PMID: 17444758.

23. J, Torabi M. Goodness-of-fit test with a robustness feature. TEST. 2022; 31: 76-100. https://doi.org/10.1007/s11749-021-00772-0

24. Rolke W, Gongora CG. A chi-square goodness-of-fit test for continuous distributions against a known alternative. Comput Stat. 2021; 36(3): 1885-1900. https://doi.org/10.1007/s00180-020-00997-x

25. Mann HB, Wald A. On the choice of the number of class intervals in the application of chi-square test. Ann Mat Stat. 1942; 13(3): 306-317. https://doi.org/10.1214/aoms/1177731569

26. Moore D. Tests of chi-squared type. In: D'Agostino RB, Stephens MA (Eds.), Goodness-of-fit Techniques. Marcel Dekker, New York, NY. 1986; 63-95. https://doi.org/10.1201/9780203753064-3

27. Cahusac PMB. Log likelihood ratios for common statistical tests using the likelihoodR package. The R Journal. 2023; 14(3): 203-213. https://doi.org/10.32614/RJ-2022-051

28. Williams DA. Improved likelihood ratio test for complete contingency tables. Biometrika. 1976; 63(1): 33-37. https://doi.org/10.1093/biomet/63.1.33

29. Cahusac PMB. Likelihood Ratio Test and the Evidential Approach for 2 × 2 Tables. Entropy (Basel). 2024 Apr 28; 26(5):375. doi: 10.3390/e26050375. PMID: 38785625; PMCID: PMC11119089.

30. McDonald JH. Handbook of Biological Statistics (3rd ed.). Sparky House Publishing, Baltimore, MD. 2014.

31. Bonamente M. Goodness of fit and parameter uncertainty for gaussian data. In: Statistics and Analysis of Scientific Data. Singapore: Springer Nature, Singapore. 2022; 233-245. https://doi.org/10.1007/978-981-19-0365-6_12

32. Scott DW. Curriculum vitae. David Warren Scott. Noah Harding Emeritus Professor of Statistics. 2023. https://www.stat.rice.edu/~scottdw/cv.pdf

Citation: De la Rubia JM (2024) Determination of the number and width of class intervals using R. Ann Environ Sci Toxicol 8(1): 022-042.
DOI: https://dx.doi.org/10.17352/aest.000077

33. Honcharenko T, Solovei O. Optimal bin number for histogram binning method to calibrate binary probabilities. In Lytvynenko I, Lupenko, S. (Eds.), Proceedings ITTAP 2023. Information Technologies: Theoretical and Applied Problems 2023. Aachen, Germany: CEUR Workshop Proceedings. 2024. https://ceur-ws.org/Vol-3628/paper18.pdf

34. Heer J. Fast and accurate gaussian kernel density estimation. In: Proceeding of the 2021 IEEE Visualization Conference (VIS). Institute of Electrical and Electronics Engineers (IEEE), New Orleans, LA. 2021; 11-15. https://doi.org/10.1109/VIS49827.2021.9623323

35. R Core Team. Quantile {stats}. R Documentation. Sample quantiles. 2024a. https://stat.ethz.ch/R-manual/R-devel/library/stats/html/quantile.html

36. Linden A. CENTILE2: Stata module to enhance centile command and provide additional definitions for computing sample quantiles. Statistical Software Components. S459262. Boston College Department of Economics. 2023.

37. Sukhoplyuev DI, Nazarov AN. Methods of descriptive statistics in telemetry tasks. In: Proceedings of the 2024 Systems of Signals Generating and Processing in the Field of on-Board Communications. Moscow, Russian Federation. Institute of Electrical and Electronics Engineers (IEEE), New Orleans, LA. 2024. https://doi.org/10.1109/IEEECONF60226.2024.10496798

38. Hyndman RJ, Fan Y. Sample quantiles in statistical packages. Am Stat. 1996; 50(4): 361-365. https://doi.org/10.2307/2684934

39. Tukey JW. Exploratory data analysis. Addison-Wesley, Readings, MA. 1977.

40. Silverman BW. Density estimation for statistics and data analysis. Chapman & Hall/CRC, London. 1986.

41. R Core Team. Density {stats}. R Documentation. Kernel density estimation. 2024b. https://stat.ethz.ch/R-manual/R-devel/library/stats/html/density.html

42. Fadillah N, Dariah PA, Anggraeni A, Cahyani N, Handayani L. Comparison of Gaussian and Epancehnikov kernels. Tadulako Social Science and Humaniora Journal. 2022; 3(1): 13-22. https://doi.org/10.22487/sochum.v3i1.15745

43. Rafajłowicz W. Nonparametric estimation of continuously parametrized families of probability density functions—computational aspects. Algorithms. 2020; 13. https://doi.org/10.3390/a13070164

44. Sheather SJ, Jones MC. A reliable data-based bandwidth selection method for kernel density estimation. J. R. Stat. Soc. Ser. B (Stat. Method). 1991; 53(3): 683-690. https://doi.org/10.1111/j.2517-6161.1991.tb01857.x

45. Venables WN, Ripley BD. Modern applied statistics with S (4th ed.). Springer, New York. 2002. https://doi.org/10.1007/978-0-387-21706-2

46. Ogwu EC, Ojarikre HIA. Comparative study of the rule of thumb, unbiased cross validation and the Sheather Jones-direct plug-in approaches of kernel density estimation using real life data. Int J Innov Res Sci Eng Technol. 2023; 11(3): 1800-1809.

47. Shimazaki H, Shinomoto S. Kernel bandwidth optimization in spike rate estimation. J Comput Neurosci. 2010 Aug;29(1-2):171-182. doi: 10.1007/s10827-009-0180-4. Epub 2009 Aug 5. PMID: 19655238; PMCID: PMC2940025.

48. Jiao Y, Li D, Lu X, Yang JZ, Yuan C. Gas: a gaussian mixture distribution-based adaptive sampling method for pinns. Social Science Research Network (SSRN). 2023. http://dx.doi.org/10.2139/ssrn.4479914

49. D'Agostino RB. Transformation to normality of the null distribution of g_1. Biometrika. 1970; 57(3): 679-681. https://doi.org/10.1093/biomet/57.3.679

50. Anscombe FJ, Glynn WJ. Distribution of the kurtosis statistic b_2 for normal samples. Biometrika. 1983; 70(1): 227-234. https://doi.org/10.1093/biomet/70.1.227

51. Lilliefors H. On the Kolmogorov–Smirnov test for normality with mean and variance unknown. J Am Stat Assoc. 1967; 62: 399-402. https://doi.org/10.1080/01621459.1967.10482916

52. Anderson TW, Darling DA. Asymptotic theory of certain "goodness-of-fit" criteria based on stochastic processes. Ann Math Stat. 1952; 23: 193-212. https://doi.org/10.1214/aoms/1177729437

53. Royston JP. A toolkit of testing for non-normality in complete and censored samples. J R Stat Soc. Series D (The Statistician). 1993; 42(1): 37-43. https://doi.org/10.2307/2348109

54. Demir S. Comparison of normality tests in terms of sample sizes under different skewness and Kurtosis coefficients. International Journal of Assessment Tools in Education. 2022; 9(2): 397-409. https://doi.org/10.21449/ijate.1101295

55. Khatun N. Applications of Normality Test in Statistical Analysis. Open Journal of Statistics 2021; 11(1): 113-122. https://doi.org/10.4236/ojs.2021.111006

56. Wijekularathna DK, Manage AB, Scariano SM. Power analysis of several normality tests: A Monte Carlo simulation study. Commun Stat Simul Comput. 2020; 51(3): 757-773. https://doi.org/10.1080/03610918.2019.1658780

57. Wald A, Wolfowitz J. An exact test for randomness in the non-parametric case based on serial correlation. Ann Mat Stat. 1943; 14(4): 378-388. https://doi.org/10.1214/aoms/1177731358

58. Chattamvelli R, Shanmugam R. Cauchy distribution. In: Continuous Distributions in Engineering and the Applied Sciences-Part I. Springer International Publishing, Cham, Switzerland. 2021; 117-131. https://doi.org/10.1007/978-3-031-02430-6_9

59. Warr RL, Erich RA. Should the interquartile range divided by the standard deviation be used to assess normality?. Am Stat. 2013; 67(4): 242-244. https://doi.org/10.1080/00031305.2013.847385

60. Fisher RA. On the mathematical foundations of theoretical statistics. Philos Trans R Soc Lond A. 1922; 222: 309-368. https://doi.org/10.1098/rsta.1922.0009

61. Woolf B. The log likelihood ratio test (the G-test); methods and tables for tests of heterogeneity in contingency tables. Ann Hum Genet. 1957 Jun;21(4):397-409. doi: 10.1111/j.1469-1809.1972.tb00293.x. PMID: 13435648.

62. Sulewski P. Equal-bin-width histogram versus equal-bin-count histogram. J Appl Stat. 2020 Jun 26; 48(12):2092-2111. doi: 10.1080/02664763.2020.1784853. PMID: 35706612; PMCID: PMC9041617.

Discover a bigger Impact and Visibility of your article publication with Peertechz Publications

Highlights

- Signatory publisher of ORCID
- Signatory Publisher of DORA (San Francisco Declaration on Research Assessment)
- Articles archived in worlds' renowned service providers such as Portico, CNKI, AGRIS, TDNet, Base (Bielefeld University Library), CrossRef, Scilit, J-Gate etc.
- Journals indexed in ICMJE, SHERPA/ROMEO, Google Scholar etc.
- OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting)
- Dedicated Editorial Board for every journal
- Accurate and rapid peer-review process
- Increased citations of published articles through promotions
- Reduced timeline for article publication

Submit your articles and experience a new surge in publication services
https://www.peertechzpublications.org/submission

Peertechz journals wishes everlasting success in your every endeavours.

Citation: De la Rubia JM (2024) Determination of the number and width of class intervals using R. Ann Environ Sci Toxicol 8(1): 022-042.
DOI: https://dx.doi.org/10.17352/aest.000077