

How to calculate the number of bins in a histogram?

Authored by
stats writer

December 12, 2025

RECOMMENDED CITATION

stats writer (2025). *How to calculate the number of bins in a histogram?*. PSYCHOLOGICAL SCALES. Retrieved from <https://scales.arabpsychology.com/?p=107262>

Determining the appropriate number of **bins** is a critical step in constructing a meaningful **histogram**. A **histogram** serves as a powerful graphical tool used in statistics to visualize the underlying frequency **distribution** of a continuous numerical **data set**.

The number of bins directly dictates the granularity of the visualization. While a simple calculation involves determining the range of your **data set** and dividing it by a predetermined **bin width**, this approach often introduces subjectivity and potential bias. The resulting quotient approximates the required number of intervals needed to span the entire range, but relying solely on arbitrary width selection can obscure key features of the data.

To move beyond arbitrary choices and toward an optimal visual representation, statisticians utilize formalized rules. These rules are designed to balance the need for detail--preventing the hiding of significant patterns--with the need for generalization--avoiding the visualization of mere random noise. This structured approach ensures that the resulting **histogram** provides an accurate and interpretable depiction of the frequency distribution.

The Crucial Role of Bin Selection in Data Visualization

The primary goal of a **histogram** is to communicate the shape and characteristics of a dataset's **distribution**. Effective bin selection is paramount because the chosen number of bins acts as a lens through which the data is viewed. Each bin groups together data points that fall within a specific numerical interval, and the height of the bar represents the frequency of observations in that interval.

If the number of bins is too low, the visualization becomes overly generalized, leading to a loss of detail. Important features, such as multimodality (multiple peaks) or skewness, might be smoothed over and hidden within broad categories. This results in a misleadingly simple representation that fails to capture the complexity inherent in the underlying measurements.

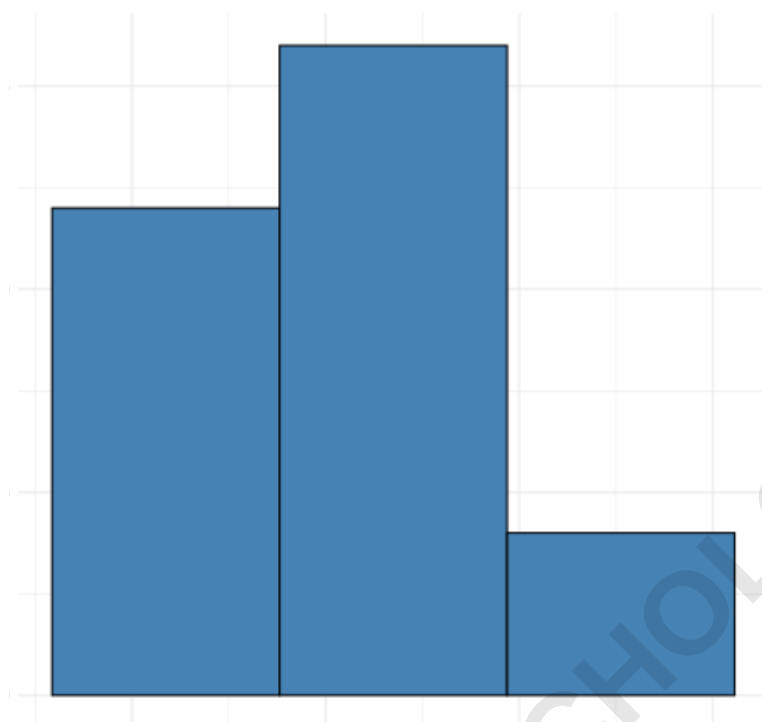
Conversely, selecting an excessively large number of bins creates a highly volatile visualization. When bins are very narrow, many bins may contain only one or zero observations, causing the histogram bars to fluctuate dramatically. This visual instability often highlights random variations inherent in the sampling process rather than the true characteristics of the population distribution, effectively visualizing noise rather than signal.

Understanding the Impact of Bin Count on Interpretation

The visual impact of inadequate bin selection is immediately apparent when comparing visualizations of the same dataset using different bin counts. This demonstrates why the choice of bin count is not merely an aesthetic decision but a methodological one that directly influences data interpretation and statistical conclusions.

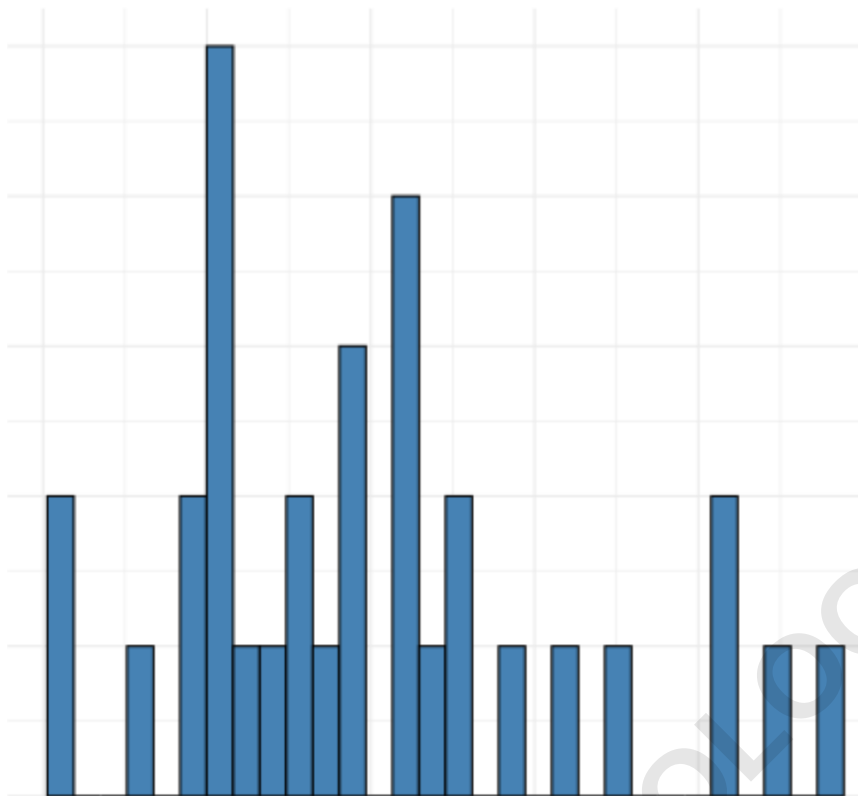
Consider a scenario where the dataset truly exhibits two distinct clusters of values (bimodal distribution). If we use too few bins, these two clusters might merge into a single, broad peak, hiding the true underlying pattern:

Total Bins = 3



In contrast, using an excessively large number of bins results in a jagged, noisy plot. While this plot theoretically shows every small variation, the high frequency of zero-count bins makes it difficult to discern the overall shape or central tendency of the **data set**, obscuring the primary distribution pattern:

Total Bins = 30



Introducing Sturges' Rule for Optimal Bin Estimation

To overcome the subjectivity associated with manual bin selection, formal methodologies have been developed. Among these, **Sturges' Rule** is perhaps the most widely recognized and frequently used approach for determining the optimal number of bins, particularly in scenarios involving normally distributed data.

Proposed by Herbert A. Sturges in 1926, this rule provides a straightforward, sample-size-dependent formula for calculating the suggested number of intervals (k) required to represent the data. The underlying assumption is that the data is generated from a normal distribution, making it an excellent default choice when no prior knowledge about the data's shape is available.

Sturges' Rule mathematically relates the number of bins to the total number of observations (n) in the dataset using a logarithmic relationship. The formula is designed to ensure that as the sample size increases, the number of bins grows slowly, maintaining visual clarity while accommodating larger volumes of data. The rule utilizes the following formula:

$$\text{Optimal Bins } (k) = \lceil \log_2 n \rceil + 1$$

It is important to understand the components of this formula to apply it correctly:

n: Represents the total count of observations in the dataset.

log₂: Denotes the logarithm base 2. Using base 2 relates conceptually to the process of successive binary partitioning of the data range.

⌈ ⌋: These symbols represent the "ceiling" function, meaning the result must be rounded up to the nearest whole integer. Since the number of bins must be an integer, this rounding ensures a sufficient number of intervals to cover the entire data range.

Applying Sturges' Rule: A Detailed Example

To illustrate the practical application of **Sturges' Rule**, consider a small but representative dataset where we have measured a specific variable for 31 subjects. The total number of observations, therefore, is $n = 31$.

Suppose we have the following set of observations. Although the values themselves are not necessary for the bin calculation, the count is essential:

Data Values
21.0
21.0
22.8
21.4
18.7
18.1
14.3
24.4
22.8
19.2
17.8
16.4
17.3
15.2
10.4
10.4
14.7
32.4
30.4
33.9
21.5
15.5
15.2
13.3
19.2
27.3
26.0
30.4
15.8
19.7
15.0

We apply the formula directly to determine the optimal number of bins (k):

$$\text{Optimal Bins} = \lceil \log_2(31) \rceil + 1$$

First, we calculate the logarithm base 2 of 31. This calculation yields a result close to 4.954:

$$\log_2(31) \approx 4.954$$

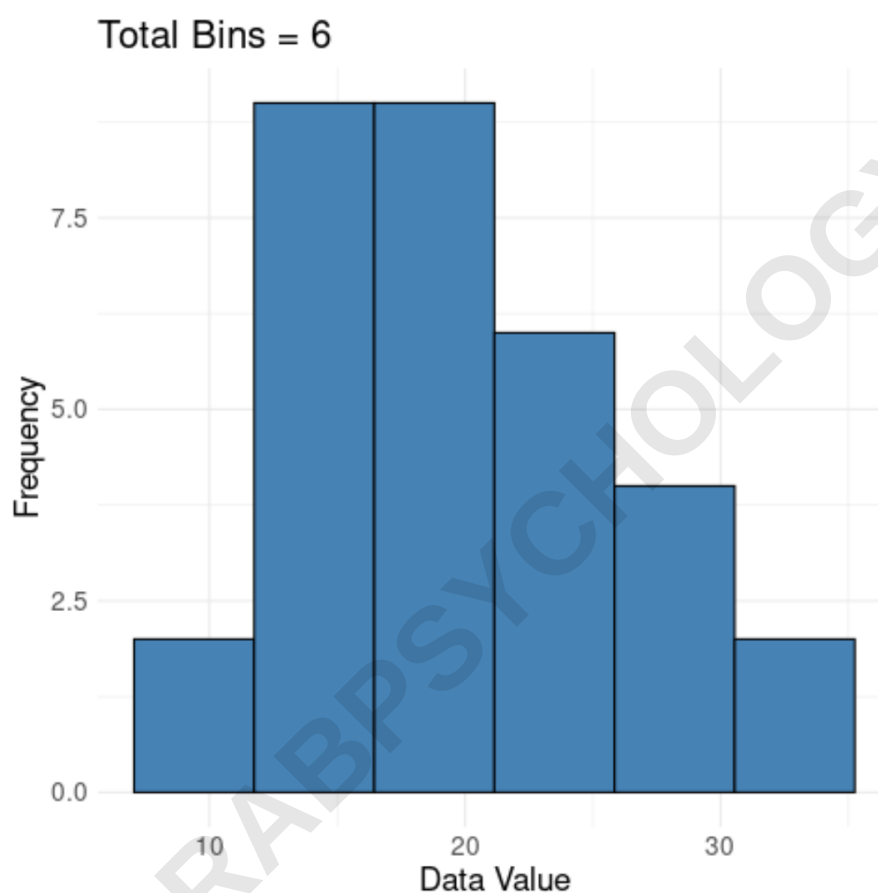
Next, we add 1 to the result:

$$4.954 + 1 = 5.954$$

Finally, we apply the ceiling function, rounding the result up to the nearest integer. The ceiling of 5.954 is 6:

Optimal Bins = $\lceil 5.954 \rceil = 6$.

According to **Sturges' Rule**, the most appropriate visualization of this specific **data set** ($n=31$) would utilize 6 bins. This approach provides a balance, offering enough resolution to reveal the shape of the **distribution** without introducing excessive noise, as demonstrated in the resulting visualization:



Key Components and Limitations of Sturges' Rule

While **Sturges' Rule** is widely used due to its simplicity and reliance only on the sample size, it operates under the strong assumption that the data is approximately normally distributed. This underlying assumption dictates its primary limitation: for datasets that are highly skewed or contain significant outliers, the rule may underestimate the required number of bins needed to accurately portray the data's shape.

Furthermore, the rule suggests that the number of bins must increase with the logarithm of the

sample size. For very large datasets, the increase in bins provided by Sturges' Rule can be too slow. For instance, increasing the sample size by a factor of 10 only adds roughly 3.3 bins (since $\log_2(10) \approx 3.3$). If a dataset contains hundreds of thousands of observations, the resulting histogram might still appear overly smooth, masking important fine-grained details.

Despite these limitations, Sturges' Rule remains a powerful starting point, particularly for moderate sample sizes (n between 50 and 200). It is often utilized as a reliable default setting in statistical software packages when a user does not explicitly define the bin count, providing an initial structure that is statistically grounded.

A Reference Guide for Sturges' Rule Values

To facilitate quick interpretation and planning, the following table provides pre-calculated optimal bin counts based on various sample sizes (n), according to Sturges' Rule. This reference highlights how the required number of bins scales logarithmically with the sample size:

Sample Size	Number of Bins
30	6
50	7
100	8
200	9
500	10
1,000	11
5,000	14
10,000	15

Analyzing this table shows that doubling the sample size only slightly increases the calculated number of bins. For example, moving from $n=64$ to $n=128$ only increases the bin count from 7 to 8. This logarithmic growth emphasizes the rule's focus on maintaining visual consistency across varying dataset sizes, prioritizing the visualization of the overall shape rather than micro-level variations.

Advanced Alternatives to Sturges' Rule

While Sturges' Rule is the most common default method, its dependence on the normality assumption often necessitates the use of alternative rules for non-normal or highly dispersed data. These alternatives incorporate additional statistical measures beyond just the sample size, such as

the variability and spread of the data, to provide a more robust calculation of the optimal **bin width** and count.

Three prominent alternatives are frequently employed in statistical analysis:

The Square-root Rule: This method is exceptionally simple and often serves as the default in many programming languages and statistical environments when data characteristics are unknown. It is particularly useful for large datasets where computational speed is a concern.

Number of Bins = $\sqrt{n}$

The Rice Rule: This rule is slightly more conservative than Sturges' Rule, often suggesting a higher number of bins, particularly for smaller sample sizes. It is designed to expose more detail in the data structure, offering a more granular look at the frequency distribution.

Number of Bins = $2 * \sqrt[3]{n}$

The Freedman-Diaconis' Rule: This is generally considered one of the most robust methods because it is resistant to outliers and skewness. Unlike Sturges' Rule, the Freedman-Diaconis rule incorporates the data's variability using the **interquartile range (IQR)**, which measures the spread of the middle 50% of the data. The formula calculates the optimal **bin width** (h), and the number of bins (k) is then determined by the data range divided by h.

Bin Width (h) = $(2 * IQR) / \sqrt[3]{n}$

Since the number of bins (k) is calculated by dividing the data range (Max - Min) by the bin width (h), this rule tailors the visualization specifically to the spread and density characteristics of the measured data, rather than just the sample count.

Implementing Bin Calculation in Practice

In modern statistical analysis, manual calculation of the number of bins is rarely necessary, as software packages handle these decisions automatically. However, understanding the underlying rules allows the analyst to override the default settings when necessary, ensuring the visualization accurately reflects complex data structures.

Most software environments, such as R, Python (with libraries like Matplotlib or Seaborn), and specialized statistical programs, allow users to choose which bin calculation algorithm to employ (e.g., Sturges, Freedman-Diaconis, or Scott's Rule). Furthermore, these programs often provide interactive visualization tools that permit the analyst to manually adjust the number of bins dynamically.

When working with programming environments, you can use built-in functions to automatically apply **Sturges' Rule** or its alternatives to determine the optimal number of bins for your histogram based on the size and characteristics of your **data set**. This reliance on computational methods ensures consistency and accuracy across different analyses.

ARABPSYCHOLOGY.COM