What is the estimated standard deviation of any given histogram?

Authored by
stats writer

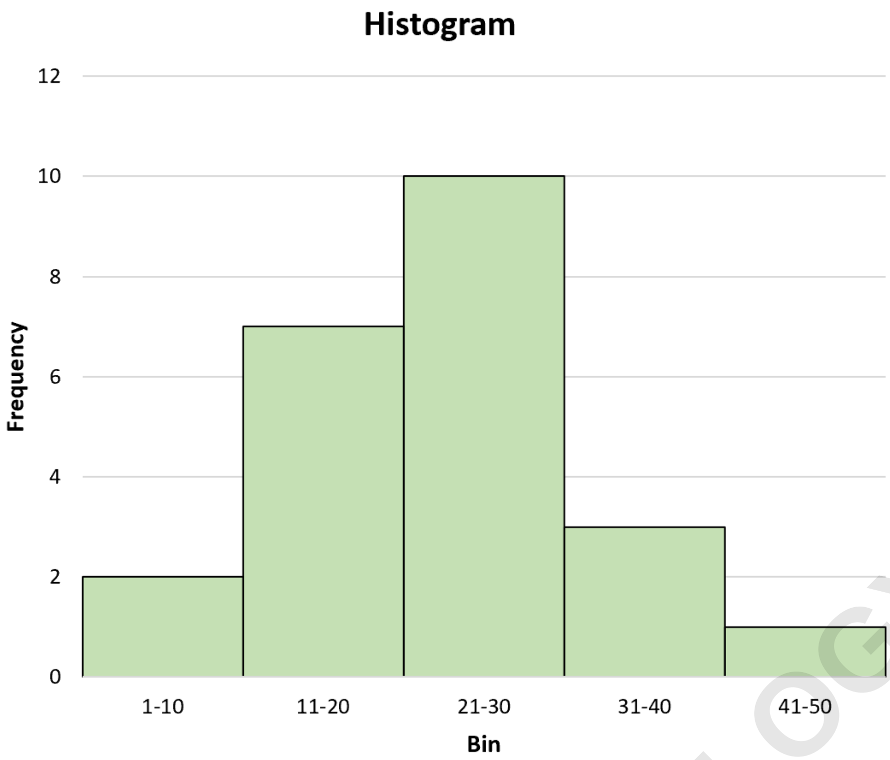
June 29, 2024

The estimated standard deviation of a histogram is a measure of the spread or variability of the data represented by the histogram. It is calculated by determining the average distance of the data points from the mean of the histogram. This value is used to understand the distribution of the data and to make comparisons between different histograms. It provides a numerical representation of the shape and dispersion of the data, allowing for a better understanding of the underlying patterns and trends. The estimated standard deviation is an important statistical tool in analyzing and interpreting histograms.

Estimate the Standard Deviation of Any Histogram

A histogram offers a useful way to visualize the distribution of values in a dataset.

The x-axis of a histogram displays bins of data values and the y-axis tells us how many observations in a dataset fall in each bin.



Since a histogram places observations in bins, it's not possible to calculate the exact standard deviation of the dataset represented by the histogram but it's possible to estimate the standard deviation.

The following example shows how to do so.

Related:

How to Estimate the Standard Deviation of a Histogram

In order to estimate the standard deviation of a histogram, we must first estimate the mean.

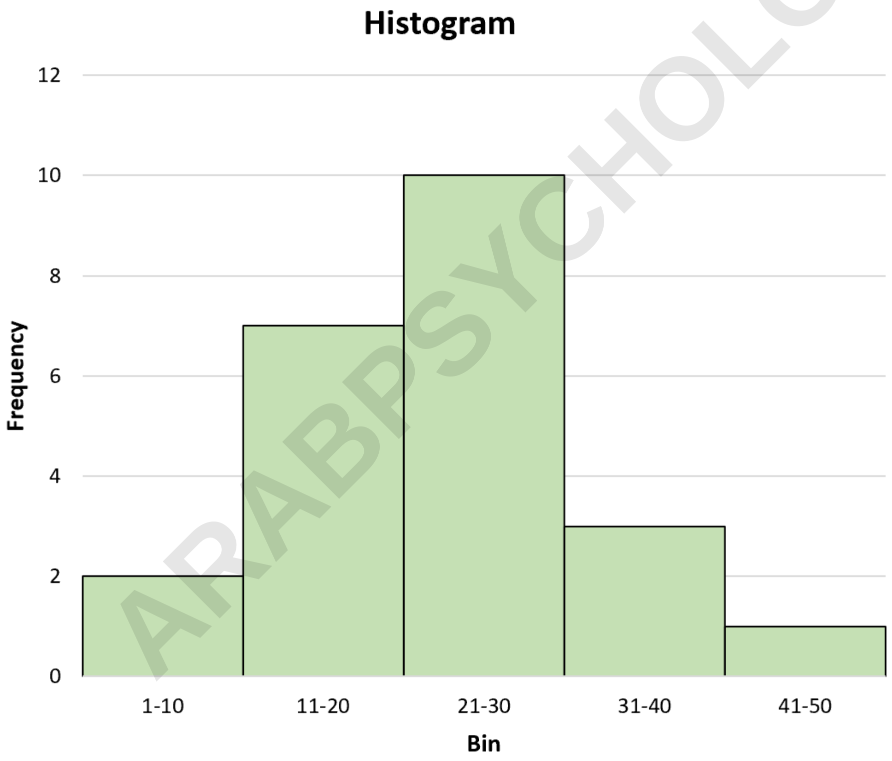
We can use the following formula to estimate the mean:

Mean: Σmini / N

where:

mi: The midpoint of the ith binni: The frequency of the ith binN: The total sample size

For example, suppose we have the following histogram:



Here's how we would estimate the mean value of this histogram:

Range	Frequency (n_i)	Midpoint (m_i)	m_i * n_i
1-10	2	5.5	11
11-20	7	15.5	108.5
21-30	10	25.5	255
31-40	3	35.5	106.5
41-50	1	45.5	45.5

Mean = (11 + 108.5 + 255 + 106.5 + 45.5) / 23 = **22.89**

We estimate the mean to be 22.89.

Now that we have an estimate for the mean, we can use the following formula to estimate the standard deviation:

Standard Deviation: √Σni(mi-μ)2 / (N-1)

where:

ni: The frequency of the ith binmi: The midpoint of the ith binμ: The meanN: The total sample size

Here's how we would apply this formula to our dataset:

Range	Frequency (n_i)	Midpoint (m_i)	$m_i * n_i$	μ	$m_i - μ$	$(m_i - μ)^2$	$n_i(m_i - μ)^2$
1-10	2	5.5	11	22.89	-17.39	302.41	604.82
11-20	7	15.5	108.5	22.89	-7.39	54.61	382.28
21-30	10	25.5	255	22.89	2.61	6.81	68.12
31-40	3	35.5	106.5	22.89	12.61	159.01	477.04
41-50	1	45.5	45.5	22.89	22.61	511.21	511.21

Standard Deviation = √((604.82 + 382.28 + 68.12 + 477.04 + 511.21) / 22) = **9.6377**

We estimate that the standard deviation of the dataset is 9.6377.

Although this isn't guaranteed to match the exact standard deviation of the dataset (since we don't know the of the dataset), it represents our best estimate of the standard deviation.

Additional Resources

The following tutorials explain how to perform other common tasks related to data grouped into bins: