What are the percentages in a box plot?

Authored by

stats writer

November 19, 2025

A box plot, also known as a box-and-whisker plot, is a fundamental tool in descriptive statistics used to visualize the distribution of a dataset. It is highly effective because it concisely displays five critical statistical measures--collectively known as the **five-number summary**--which delineate the spread and central tendency of the data. Understanding the percentages within a box plot is crucial for interpreting data, as these plots are inherently structured around key statistical boundaries: the **median**, the upper and lower quartiles, and the minimum and maximum values.

The percentages embedded within the structure of a box plot reveal the relative positioning of data points. Specifically, the structure is designed to divide the entire dataset into four equal sections, each containing 25% of the observations. This powerful segmentation allows analysts to quickly gauge how tightly or loosely the data is clustered around the center. For instance, the central line of the box marks the 50th percentile, meaning half of the data points lie below this value, while the boundaries of the box itself represent the 25th and 75th percentiles.

By mapping these five points onto a numerical scale, the box plot offers an immediate visual summary of the data's entire range, its symmetry, and the presence of potential **outliers**. This guide will explore in depth how the components of the box plot correspond directly to specific percentages, enhancing your ability to interpret data distribution across various fields, from finance and engineering to academic research.

Understanding the Foundation: The Five-Number Summary

The cornerstone of interpreting any box plot lies in recognizing the **five-number summary**. This statistical framework distills a potentially vast dataset into five essential values that describe its location and spread. These values are meticulously calculated from the raw data and form the visible components of the graphical representation, providing a robust overview suitable for preliminary data analysis and comparison between different groups.

The calculation of these five values requires the data points to be sorted in ascending order. Once sorted, these specific markers are identified. The minimum value marks the smallest observation, while the maximum value denotes the largest observation, thus defining the entire span of the dataset (the range). Crucially, the three remaining values--the first quartile (Q1), the median (Q2), and the third quartile (Q3)--are positional measures that divide the data into four segments, each containing 25% of the data points, which directly translates to their percentile ranks.

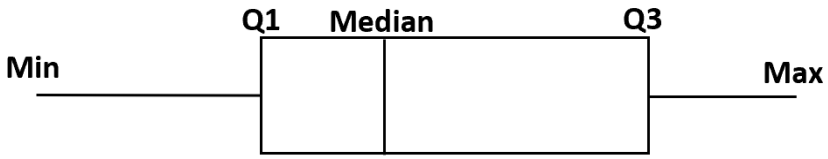
A comprehensive list of these components ensures clarity when beginning any analysis:

A **box plot** is a type of plot that displays the five number summary of a dataset, which includes:

The **Minimum Value** (0th Percentile)
The **First Quartile** (Q1, 25th Percentile)

The **Median Value** (Q2, 50th Percentile)
The **Third Quartile** (Q3, 75th Percentile)
The **Maximum Value** (100th Percentile)

A typical box plot looks like this, visually confirming the positions of the five-number summary:



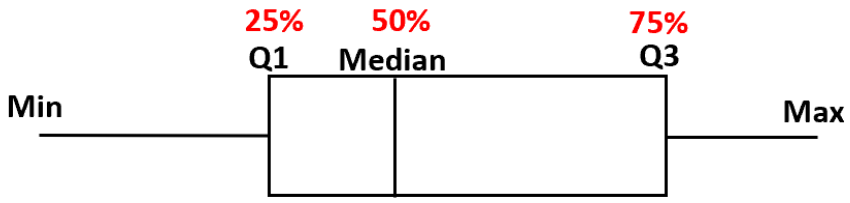
The Relationship Between Quartiles and Percentiles

The heart of interpreting a box plot's percentages lies in the concept of quartiles and their direct correspondence to specific percentile ranks. A quartile is simply a measure that divides the data points into four intervals, or quarters. When we discuss percentages in this context, we are defining what proportion of the data falls below a certain point on the distribution scale. This relationship is non-negotiable and provides the rigid structure necessary for statistical interpretation.

The three quartiles (Q1, Q2, and Q3) map directly to the 25th, 50th, and 75th percentiles, respectively. The **First Quartile (Q1)** is the value below which 25% of the observations fall. This means that if you are analyzing test scores, 25% of students scored at or below the Q1 value. Similarly, the **Median (Q2)**, which cuts the data exactly in half, represents the 50th percentile. This is a measure of central tendency and is often preferred over the mean when dealing with skewed data, as it is less sensitive to extreme values or outliers.

Finally, the **Third Quartile (Q3)** marks the 75th percentile. At this point, 75% of the data points are equal to or less than Q3, and conversely, only 25% of the data lies above it. The distance between these quartiles is critical for measuring variability, as this distance encapsulates the central majority of the data. This visual segmentation immediately shows the density of the data distribution, allowing for rapid comparison of variability across datasets. This conceptual framework is summarized visually below:

The **first quartile** represents the **25th percentile** of all values in the dataset.
The **median** represents the **50th percentile** of all values in the dataset.
The **third quartile** represents the **75th percentile** of all values in the dataset.



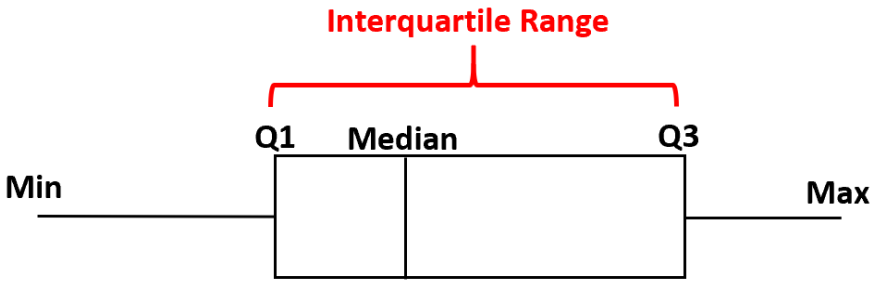
Deconstructing the Box: The Interquartile Range (IQR)

The central box component of the box plot holds a particularly significant statistical measure: the Interquartile Range (IQR). The IQR is defined as the difference between the third quartile (Q3) and the first quartile (Q1), and it is arguably the most informative percentage-based measure derived from the plot. It quantifies the spread of the middle 50% of the data, offering a robust and resistant measure of variability that is unaffected by extreme outliers found in the tails of the distribution.

Statistically, the IQR represents the region where the data is most concentrated. Because 25% of the data lies between the minimum and Q1, 25% between Q1 and the median, 25% between the median and Q3, and 25% between Q3 and the maximum, the combined area of Q1 to Q3 necessarily contains 50% of the observations. A narrow box indicates that the central half of the data is tightly clustered, suggesting low variability, while a wide box signifies that the central 50% is spread out over a larger range of values, indicating higher variability.

Calculating the IQR is straightforward, requiring only the subtraction of the lower boundary of the box from the upper boundary. This simple calculation provides a powerful metric for comparing the dispersion of two or more datasets visualized side-by-side. For instance, in comparing the salaries of employees in two different departments, a department with a smaller IQR suggests more homogeneity in pay among the middle earners, whereas a larger IQR suggests greater disparity. Furthermore, the IQR is critically used in the calculation of fences to identify potential **outliers**, which are points that fall significantly outside this central range.

The **interquartile range** tells us the spread of the **middle 50% of values** in a dataset and is calculated by subtracting the first quartile from the third quartile in a box plot:



Interpreting the Whiskers: Minimum and Maximum Values

While the box captures the middle 50% of the data, the whiskers, extending outward from the box, represent the remaining 50% of the dataset--25% in the lower whisker and 25% in the upper whisker. These whiskers terminate at the minimum and maximum values, defining the overall range of the data distribution (the 0th percentile and the 100th percentile, respectively). In a standard box plot that does not account for outliers, the whiskers stretch from Q1 down to the minimum data point and from Q3 up to the maximum data point.

The length of the whiskers is highly informative regarding the tail distribution. A long whisker indicates that the data points in that quarter are widely dispersed, suggesting that the lowest or highest values are significantly far from the central mass. Conversely, a short whisker implies that the observations are closely clustered near the box boundary. By examining the length and symmetry of both the box and the whiskers, one can quickly infer the overall shape and **skewness** of the distribution.

It is important to note that many modern statistical packages use a refined method for drawing whiskers, particularly when identifying outliers. In this revised approach, the whiskers extend only to the most extreme data point that is not considered an outlier, typically defined as being within 1.5 times the Interquartile range (IQR) of the box edges. Any data points falling outside this range are plotted individually as separate points, signifying their status as potential anomalies. This modification ensures that the minimum and maximum values displayed in the plot might not always correspond to the absolute minimum and maximum values in the dataset if outliers are present.

Visualizing Data Distribution and Skewness

Beyond simply showing the percentages, the box plot is an excellent tool for visualizing the shape of the data distribution, particularly its symmetry or lack thereof (skewness). The relative position of the median within the box and the comparative lengths of the whiskers provide instant clues about whether the data is symmetrically distributed (like a normal distribution) or heavily weighted to one side.

If the distribution is perfectly symmetrical, the median line (50th percentile) will be located exactly in the center of the box, and the two whiskers will be roughly equal in length. This indicates that the spread of the lower 50% of the data mirrors the spread of the upper 50%. However, most real-world datasets exhibit some degree of **skewness**.

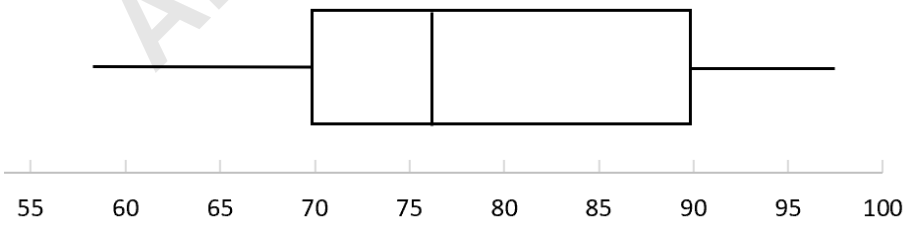
In a **positively skewed** (or right-skewed) distribution, the bulk of the data is concentrated towards the lower values. Visually, this translates to the median being closer to the first quartile (Q1), and the upper whisker being noticeably longer than the lower whisker. This indicates a long tail extending into the higher values. Conversely, a **negatively skewed** (or left-skewed) distribution shows the opposite pattern: the median is closer to the third quartile (Q3), and the lower whisker is longer, indicating a concentration of data at higher values and a tail extending toward the minimum.

Practical Application: Interpreting Percentages in Real-World Data

The utility of understanding box plot percentages is best demonstrated through practical examples, where these visual markers translate directly into actionable insights about a population or process. When interpreting data, the key is to remember that the sections defined by the quartiles contain definite percentages of observations, regardless of the scale of the variable being measured.

The following example illustrates how to utilize the positional information--the 25th, 50th, and 75th percentiles--to answer specific questions regarding data distribution. This analytical approach moves beyond simple descriptive statistics and facilitates inferential reasoning about the population from which the sample data was drawn. For instance, knowing that 75% of a population falls below a certain risk score is far more insightful than simply knowing the average risk score.

The following example shows how to use a box plot to answer questions related to percentages, using a distribution of final exam scores for college students in a certain class:



Interpreting the Box Plot Example Questions

Question 1: What percentage of students scored below a 70?

From the box plot we can observe that the score of 70 aligns precisely with the marker for the **first quartile (Q1)**. As established, Q1 represents the 25th percentile of the dataset.

Thus, **25%** of students scored at or below a 70.

Question 2: What percentage of students scored above a 90?

The score of 90 corresponds exactly to the **third quartile (Q3)** marker on the plot. The third quartile signifies the 75th percentile, meaning 75% of scores are at or below 90. Therefore, to find the percentage scoring *above* 90, we must subtract the 75th percentile from the total distribution (100%).

Calculation: 100% - 75% = **25%** of students scored above a 90.

Question 3: What percentage of students scored between a 70 and a 90?

The scores 70 and 90 represent the first and third quartiles (Q1 and Q3) of the dataset, which correspond with the 25th and 75th percentiles. The range between Q1 and Q3 is, by definition, the **Interquartile Range** (IQR), which contains the middle 50% of the data.

Thus, 75% - 25% = **50%** of students scored between a 70 and 90.

Advantages and Limitations of Using Box Plots

Box plots offer significant advantages in data analysis, primarily due to their ability to present a wealth of statistical information in a visually concise format. They are exceptionally useful when comparing the distributions of multiple datasets simultaneously, as the position and size of the boxes and whiskers immediately highlight differences in central tendency, variability (spread), and skewness between groups. Furthermore, their construction based on the median and quartiles makes them robust against the influence of extreme values, providing a measure of distribution that is often more representative than methods relying solely on the mean and standard deviation, which are sensitive to outliers.

However, despite their strengths, box plots also have limitations that analysts must consider. The primary drawback is that they mask the underlying density of the data distribution within the quartiles. While we know 25% of the data is in each segment, we cannot discern how those data points are distributed within that segment--they could be clustered tightly near the edge, or spread uniformly. For instance, two datasets could have identical box plots but vastly different underlying probability distributions (e.g., bimodal vs. uniform within the quartile segments).

For more granular analysis of the distribution shape, particularly if the underlying process is complex or features multiple modes, a **histogram** or a **density plot** might be more appropriate.

Therefore, experts often recommend using box plots in conjunction with other visualizations. They serve as an excellent summary and comparison tool, but they should be supplemented when a deep understanding of the internal structure of the 25% segments is required. The ability to quickly identify the 0th, 25th, 50th, 75th, and 100th percentile remains their single greatest interpretative strength.

The following tutorials provide additional information about box plots: