

Learning the Five Number Summary with R: A Comprehensive Guide

Authored by
Mohammed loot

November 3, 2025

RECOMMENDED CITATION

Mohammed loot (2025). *Learning the Five Number Summary with R: A Comprehensive Guide*. PSYCHOLOGICAL STATISTICS. Retrieved from <https://statistics.arabpsychology.com/?p=9169>

The **five number summary** (FNS) is one of the most fundamental tools in **Exploratory Data Analysis** (EDA), providing a concise and robust description of a dataset's distribution. By distilling the data down to just five key values, the FNS offers immediate insight into the central tendency, spread, and potential skewness of a variable without relying on parametric assumptions. This summary is particularly valued because its components are highly resistant to the influence of extreme **outliers**, making it a reliable measure even in noisy data environments.

Crucially, the FNS serves as the statistical backbone for the **boxplot**, a universally recognized graphical tool for visualizing data distribution. For any professional working with statistical computing, mastering the calculation of these values in **R** is an essential skill. The five specific values that define this statistical summary are derived from ordered observations:

The **Minimum**: The absolute smallest observation recorded in the dataset.

The **First Quartile** (Q1): The value representing the 25th percentile, marking the boundary of the lowest quarter of the data.

The **Median** (Q2): The middle value, representing the 50th percentile, which divides the ordered dataset into two equal halves.

The **Third Quartile** (Q3): The value representing the 75th percentile, marking the boundary of the highest quarter of the data.

The **Maximum**: The absolute largest observation recorded in the dataset.

Deconstructing the Five Number Summary Components

Each element of the FNS contributes uniquely to the overall description of the data's shape and spread. Unlike traditional measures like the mean and standard deviation, which are sensitive to every data point, the FNS utilizes percentiles, providing a non-parametric summary that offers a truer sense of the data's concentration away from extreme values.

The two extreme values--the **minimum** and the **maximum**--are primarily useful for defining the overall range of the data, thereby establishing the full scope of observed values. While they give context to the scale, these boundary values should be interpreted cautiously, as they are the most susceptible to being influenced by measurement errors or genuine extreme outliers. Consequently, the three central measures often provide a more reliable picture of the typical behavior within the dataset.

The central measures (Q1, **Median**, Q3) are critical for understanding both the central tendency and the symmetry of the distribution. The median (Q2) functions as the primary measure of central location, dividing the distribution precisely in half. The distance between the first quartile (Q1) and the third quartile (Q3) defines the **Interquartile Range** (IQR). The IQR is statistically robust, measuring the spread of the middle 50% of observations, making it an excellent and stable indicator of statistical dispersion, regardless of skewness or the presence of distant outliers.

The Essential Role of the FNS in Data Diagnostics

The five number summary is essential for rapid data diagnostics, offering a snapshot that facilitates efficient comparison between different variables or groups and guides initial data cleaning efforts. Its structured presentation allows analysts to quickly assess multiple distribution characteristics simultaneously.

Firstly, the FNS immediately establishes the location of the **central tendency** via the median, which remains stable even if the data distribution is heavily skewed. Secondly, it quantifies the **variability** of the central mass through the IQR; a large IQR suggests high dispersion within the core data, while a small IQR indicates tight clustering. Finally, the minimum and maximum values provide the definitive **range**, setting the absolute bounds for the variable under study.

Furthermore, the spatial relationship between the median and the two quartiles is a powerful visual indicator of potential skewness. If the median is positioned closer to Q1, the distribution is characterized as positively skewed (skewed right), suggesting a long tail extending toward higher values. Conversely, if the median is closer to Q3, the distribution is negatively skewed (skewed left). This diagnostic ability makes the FNS a crucial preliminary step for selecting appropriate statistical models. In the R environment, the most straightforward and consistent method for calculating the five number summary is through the dedicated `fivenum()` function, which is part of base R:

fivenum(data)

The `fivenum()` function utilizes a specific computational method, often attributed to Tukey, for determining the quartiles (sometimes called hinges). This method ensures that the calculated values are perfectly consistent with the standard algorithm R uses when generating a boxplot, providing reliability and direct visual confirmation of the summary statistics.

Example 1: Calculating the FNS for a Simple Vector

This foundational example demonstrates the direct application of the `fivenum()` function to a single, unstructured numeric vector. Understanding this basic process is key before moving on to more complex data structures like data frames.

We begin by defining a sample dataset, which consists of thirteen arbitrary numeric data points. We then apply the `fivenum()` function directly to this vector. The output is a structured vector listing the five statistical values in their defined order: Minimum, First Quartile (Q1), Median, Third Quartile (Q3), and Maximum.

#define numeric vector

```
data <- c(4, 6, 6, 7, 8, 9, 12, 13, 14, 15, 15, 18, 22)
```

```
#calculate five number summary of data
```

```
fivenum(data)
```

```
4 7 12 15 22
```

From the printed output, the components defining the distribution are clearly identifiable and interpreted as follows:

The minimum value is: **4**

The first quartile (25th percentile) is: **7**

The median (50th percentile) is: **12**

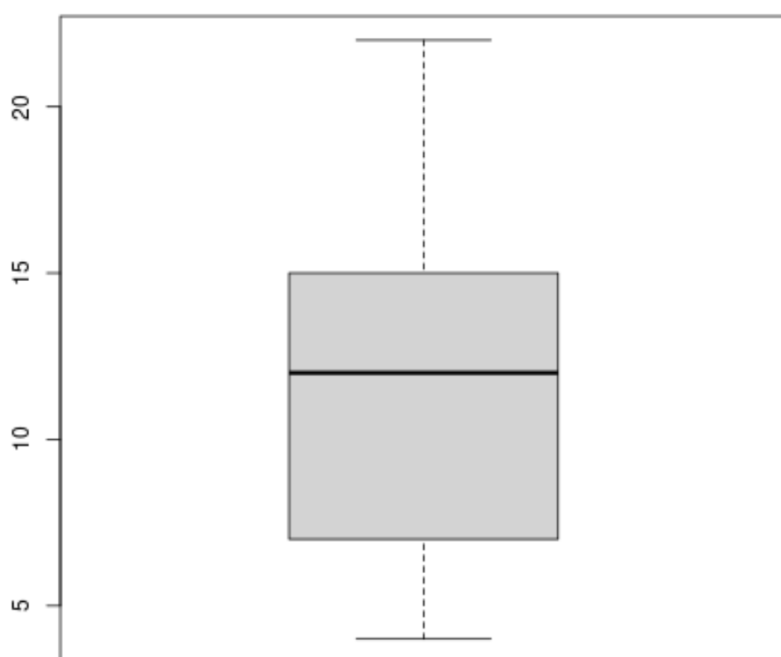
The third quartile (75th percentile) is: **15**

The maximum value is: **22**

To provide immediate graphical context for this summary, we can generate a boxplot using the same vector. The boxplot offers an immediate, visual confirmation of the calculated FNS values, illustrating the central tendency and dispersion graphically:

```
boxplot(data)
```

```
4 7 12 15 22
```



The elements of the boxplot directly correspond to the five number summary, allowing for clear visual interpretation:

The lowest point of the lower whisker corresponds precisely to the minimum value (**4**).

The bottom edge of the central box represents the first quartile (**7**).

The line bisecting the box indicates the median value (**12**).

The top edge of the central box represents the third quartile (**15**).

The highest point of the upper whisker corresponds to the maximum value (**22**).

In this specific visualization, the box (representing the IQR) is relatively symmetric around the median line. This suggests that the middle half of the distribution is quite balanced, indicating that the data is not heavily skewed toward either extreme.

Example 2: Analyzing a Specific Column within an R Data Frame

In typical data science workflows, observations are organized into structured tables, commonly referred to as R [data frames](#). When the objective is to calculate the five number summary for a single variable housed within this structure, it is necessary to correctly reference the specific column of interest. This is achieved in R using the dollar sign (\$) operator, which isolates the vector associated with that column name.

For this example, we construct a small hypothetical data frame containing basketball statistics for various teams. Our focus is solely on the distribution of the `points` scored. We apply the `fivenum()` function to `df$points`, ensuring that only the numerical values from that single variable are passed to the function for summarization.

#create data frame

```
df <- data.frame(team=c('A', 'B', 'C', 'D', 'E', 'F', 'G', 'H'),
points=c(99, 90, 86, 88, 95, 87, 85, 89),
assists=c(33, 28, 31, 39, 34, 30, 29, 25),
rebounds=c(30, 28, 24, 24, 28, 30, 31, 35))
```

#calculate five number summary of points column

```
fivenum(df$points)
```

```
85.0 86.5 88.5 92.5 99.0
```

The resulting summary provides immediate and actionable insights into the team performance data. We observe that the range of scores spans from a minimum of **85.0** to a maximum of **99.0**. The Interquartile Range (calculated as Q3 minus Q1: $92.5 - 86.5$) equals 6.0, which is a relatively small spread. This tight IQR indicates that the middle 50% of the team scores are closely clustered

around the median performance of **88.5**, suggesting high consistency among the central group of teams.

Example 3: Streamlining Analysis with Batch Processing

During the exploratory phase of a project, data analysts frequently need to calculate descriptive statistics for numerous numeric variables simultaneously. Relying on repeated manual calls to `fivenum()` for each column becomes tedious, time-consuming, and increases the likelihood of human error.

To dramatically streamline this process, R offers powerful vectorized functions. We can leverage the `sapply()` function, which is designed to apply a specified function--in this case, `fivenum`--across every element of a list or, more practically, across a selected subset of columns within a data frame. This method returns the results organized efficiently into a matrix.

Using the same basketball statistics data frame, we select the three quantitative columns (`points`, `assists`, and `rebounds`) and pass this subset to `sapply()` along with the `fivenum` argument. The result is a concise summary matrix, allowing for direct comparison of the three distributions:

#create data frame

```
df <- data.frame(team=c('A', 'B', 'C', 'D', 'E', 'F', 'G', 'H'),  
points=c(99, 90, 86, 88, 95, 87, 85, 89),  
assists=c(33, 28, 31, 39, 34, 30, 29, 25),  
rebounds=c(30, 28, 24, 24, 28, 30, 31, 35))
```

```
#calculate five number summary of points, assists, and rebounds column  
sapply(df, fivenum)
```

```
points assists rebounds  
85.0 25.0 24.0  
86.5 28.5 26.0  
88.5 30.5 29.0  
92.5 33.5 30.5  
99.0 39.0 35.0
```

This resulting matrix is perfectly structured: Row 1 invariably contains the Minimum value for each variable, Row 3 provides all Median values, and Row 5 contains the Maximum values. This vectorized technique significantly enhances efficiency and readability when conducting comprehensive descriptive analysis on data frames containing many variables.

Expanding Descriptive Analysis Beyond `fivenum()`

While the `fivenum()` function offers a precise and boxplot-consistent summary, it represents just one facet of R's powerful descriptive statistics capabilities. The calculation of the five number summary often serves as a necessary prelude to deeper statistical investigations.

For analysts seeking a broader descriptive overview, the base R function `summary()` is an excellent alternative. It automatically provides the FNS components but also includes the mean and a count of any missing values (NAs), offering a more complete initial picture of the data's central location and quality. Furthermore, the `quantile()` function provides granular control over percentile calculation methods, allowing users to customize how quartiles and other percentiles are derived, which can be crucial when dealing with complex or non-standard distributions. Exploring these related functions will provide a much deeper understanding of the descriptive statistics ecosystem available within the R environment.