The Importance of Summary Statistics and Techniques for Creating Them in R

Jelin George

Abstract

This document presents a concise overview of summary statistics and their importance in R. Summary statistics - such as mean, median, standard deviation, and frequency counts - capture the key features of a dataset, enabling quick exploration and interpretation. R provides powerful functions and visualization tools to efficiently compute and present these statistics, making them essential for simplifying data, identifying patterns, and supporting informed analysis and decision-making." keywords: [summary statistics, R programming, data analysis, data visualisation, data summarisation]

Introduction

Summary statistics are concise numerical measures that capture the essential characteristics of a dataset. They serve as foundational tools in data analysis, providing concise descriptions of large datasets. They help analysts and researchers understand the central tendencies, variability, and overall distribution of data, making complex datasets interpretable and actionable. Without summary statistics, raw data would be overwhelming and difficult to interpret, making it challenging to draw meaningful conclusions or communicate findings effectively.

In R, summary statistics are foundational for data analysis, enabling users to efficiently condense complex data into interpretable values like the mean, median, mode, standard deviation, and quantiles. These statistics provide a snapshot of the data, facilitating initial exploration, quality checks, and communication of results.

In this document, we will:

Define summary statistics and their importance

Explore key measures and techniques

Demonstrate practical application with code

Discuss best practices and common pitfalls

What is Summary Statistics?

Summary statistics are numerical values that describe the main features of a dataset, such as its center and spread. They simplify complex data into easily interpretable numbers, offering quick insights into trends and variability, and serve as a foundation for further analysis. These statistics provide a snapshot of the data, facilitating initial exploration, quality checks, and communication of results.

R offers a rich ecosystem of functions and packages - such as summary(), dplyr::summarise(), and visualization tools like histograms and boxplots - that streamline the computation and presentation of summary statistics for both numeric and categorical data. Their importance lies in simplifying large datasets, revealing patterns and outliers, and laying the groundwork for deeper statistical analyses and informed decision-making.

As the first and often most critical step in any analytical workflow, summary statistics in R empower analysts and researchers to understand, compare, and communicate data-driven insights with clarity and precision.

Summary statistics can be typically divided into:

Descriptive statistics: Summarize the main features of a dataset (e.g., mean, median, mode). This will be our focus here.

Inferential statistics: Make predictions or inferences about a population based on a sample (not the focus here).

Sources: (Lane, 2013) (Oh & Pyrczak, 2023)

Watch this tutorial video on Summary Statistics to get you started.

Key Measures in Summary Statistics

Measures of Central Tendency

Central tendency measures indicate where most values in a dataset fall.

Mean: The arithmetic average. Add up all the values, then divide by how many there are to get the average of all the numbers.

Median: The middle value when data is ordered. If there's an even number, it's the average of the two middle numbers.

Mode: The most frequently occurring value.

{r} # Example in R data <- c(2, 4, 4, 4, 5, 7, 9) mean(data) # Arithmetic mean median(data) # Median Mode <- function(x) { ux <- unique(x) ux[which.max(tabulate(match(x, ux)))] } Mode(data) # Mode

Measures of Dispersion

Dispersion measures describe the spread of data.

Range: Difference between max and min values.

Variance: Average squared deviation from the mean.

Standard Deviation: Square root of variance.

Interquartile Range (IQR): Range between the 25th and 75th percentiles.

Coefficient of Variation: Standard deviation divided by the mean.

{r} range(data) var(data) sd(data) IQR(data) sd(data)/mean(data) # Coefficient of Variation

Measures of Shape and Distribution

Describe the overall pattern and characteristics of how data values are distributed within a dataset.

Skewness: Measures asymmetry of the distribution.

Kurtosis: Measures "tailedness" or peakedness of the distribution.

{r} library(e1071) skewness(data) kurtosis(data)

Visualization tools to help understand distribution of data better.

Histogram: It displays how data values are distributed across different intervals in patterns like bell-shaped (normal), J-shaped, or skewed distributions, as well as spot outlines and the overall spread of the data.

As shown below, the distribution is centered around 4.

{r} # This histogram shows the distribution in the example we set previously.

#| label: fig-histogram #| fig-cap: "Histogram of the data" hist(data, main="Figure1Histogram of Data", col = "grey")

Boxplot: A graphical tool that visually summarizes the distribution, central tendency, spread, and skewness of numerical data using the five-number summary: minimum, first quartile (Q1), median, third quartile (Q3), and maximum.

{r} boxplot(data, main="Boxplot of Data", col="grey")

Handling Missing Data

Missing data can bias summary statistics if not handled properly.

Types of missing data: MCAR (Missing Completely at Random), MAR (Missing at Random), MNAR (Missing Not at Random).

Techniques: Omit missing values, impute with mean/median/mode, or use advanced imputation.

{r} data_with_na <- c(2, 4, NA, 4, 5, NA, 9) mean(data_with_na, na.rm=TRUE) # Ignore NAs

Frequency and Cross-Tabulation Techniques

Frequency table: A tool used to organize and display how often each value or category occurs in a dataset. It typically consists of two or more columns: one listing all possible values or categories, and another showing the frequency (count) of each making it easier to see which values are common or rare, summarize large sets of data, and identify patterns.

Relative frequency: Proportion of each category.

Cumulative frequency: Running total of frequencies.

{r} # To create a frequency table category <- c('A', 'B', 'A', 'C', 'B', 'A') table(category)

To calculate the proportion of each unique value in the vector category

prop.table(table(category)) |> round(digits = 2)

{r} # To calculate the total of frequencies category <- c('A', 'B', 'A', 'C', 'B', 'A') cumulative_freq <- cumsum(table(category)) cumulative_freq

Cross-tabulations (contingency tables): Used in statistics to examine and summarise the relationship between two or more categorical variables. In a cross-tabulation, one variable's categories are arranged in the rows and another variable's categories in the columns, with each cell showing the frequency (count) of observations that fall into the corresponding combination of categories.

{r} gender <- c('M', 'F', 'F', 'M', 'F', 'M') table(category, gender)

Row/column proportions: Use prop.table() with margin argument.

{r} gender <- c('M', 'F', 'F', 'M', 'F', 'M') prop.table(table(category, gender)) |> round(digits = 3)

Summarizing Data Frames

We can create comprehensive summaries for entire datasets by summarizing data frames. This involves generating clear overviews of each variable and its values, typically by calculating summary statistics such as the mean, median, minimum, maximum, standard deviation, and counts. These summaries help reveal the structure, trends, and important features of the data.

Let's explore a basic example using the summary() function.

{r} # Install skimr if not already installed # install.packages("skimr")

library(skimr) df <- data.frame(Age = c(21, 22, 22, 23, 24, 25, 25), Gender = c('F', 'M', 'M','F', 'F', 'M', 'M'), Score = c(85, 90, 88, 95, 85, 88, 80)) skim(df)

dplyr::glimpse(df)

summary(df)

We can use the above dplyr::glimpse(df) for a quick structure overview, or summary(df) for base R summaries, but skimr gives the most detailed tidy summary. You can further explore skimr here.

Practical Application

Let us begin with a few fun exercises to understand how to read data and apply summary statistics functions using the Star Wars dataset.

Before we get started, we must install essential packages that might be needed later.

{r} if (!require(pacman)) install.packages("pacman") pacman::p_load(tidyverse)

{r} #| label: setup #| include: false library(conflicted) library(tidyverse) library(flextable) library(ftExtra) library(knitr) library("htmltools") library("ggplot2") library(dplyr) library(DT) conflicts_prefer(dplyr::filter, .quiet = TRUE) conflicts_prefer(flextable::separate_header, .quiet = TRUE)

Now, load the Star Wars dataset available online in the dylyr package. (Wickham et al., 2023)

{r} library(dplyr) data(starwars)

{r} # Summary table for interactive exploration

DT::datatable(starwars, caption = 'Table 1. Star Wars Data')

Let's now streamline the dataset to include only the essential variables before applying the summary functions.

{r} library(knitr) library(kableExtra) starwars %>% select(-films, -starships, -homeworld, -vehicles) %>% #remove the 'film' & 'starships' & 'homeworld' & 'vehicles' column slice(1:10) %>% kable(caption = "Table2. Star Wars Dataset", align = "l") %>% # 'c' centers columns kable_styling(full_width = TRUE, position = "center")

{r} # Mean height starwars %>% summarise(mean_height = mean(height, na.rm = TRUE))

{r} # Median height starwars %>% summarise(median_height = median(height, na.rm = TRUE))

{r} # Mode height get_mode <- function(x) { # Get unique values in x ux <- unique(x)
# Find the value with the highest frequency ux[which.max(tabulate(match(x, ux)))]
}

starwars %>% summarise(mode_height = get_mode(height[!is.na(height)]))

Now, let's apply various summary functions we saw previously.

{r} # Select relevant variables starwars_selected <- starwars %>% select(height, mass, gender, birth_year, species)

Base R summary for numeric variables

summary(starwars_selected %>% select(height, mass))

Frequency table for gender

table(starwars_selected$gender)

Proportion table for gender

prop.table(table(starwars_selected$gender))

Comprehensive summary using skimr

skim(starwars_selected)

Let's look at a visual pattern of the height of different characters in Star Wars.

{r} library(dplyr) library(ggplot2)

starwars %>% ggplot(aes(x = height)) + geom_histogram(binwidth = 10, fill = "navy", color = "blue") + labs(title = "Fig1. Histogram of Height of Characters in Star Wars", x = "Height (cm)", y = "Count") + theme_minimal()

Now, we filter the species to get visual pattern of the height of different human characters in Star Wars.

{r} starwars %>% filter(species == "Human") %>% ggplot(aes(x = height)) + geom_histogram(binwidth = 10, fill = "navy", color = "blue") + labs(title = "Fig2. Histogram of Height of Human Characters in Star Wars", x = "Height (cm)", y = "Count") + theme_minimal()

Limitations

Summary statistics are essential for providing a quick and accessible overview of a dataset, but they have several important limitations:

No Causality or Explanation: Summary statistics describe what is present in the data but cannot explain why patterns exist or establish causal relationships. For example, knowing the average test score does not reveal the factors that influenced those scores.

Limited to the Sample: These statistics only summarize the data actually measured and cannot be generalized to a broader population without further inferential analysis. They do not account for sampling variability or external validity.

No Predictive Power: Summary statistics cannot be used to make predictions about future observations or unmeasured data; they are purely descriptive.

Loss of Detail and Nuance: By condensing complex data into single values (like the mean or median), summary statistics can obscure important patterns, subgroups, or variability within the data. For instance, two datasets with the same mean can have very different distributions.

Potential for Misleading Conclusions: Relying solely on summary statistics can mask underlying issues such as data bias, or important subgroup differences, leading to incomplete interpretations.

No Insight into Relationships: Summary statistics typically focus on individual variables and do not reveal relationships or associations between multiple variables.

In summary, while summary statistics are valuable for initial data exploration, they should be complemented with more detailed analyses and visualizations to avoid oversimplification and misinterpretation of the data.

Sources: (Wienclaw, 2009) (Wienclaw, 2017)

Future Direction

The future direction of summary statistics is being shaped by advances in data complexity, computational power, and the integration of artificial intelligence.

Several key trends and directions can be identified:

Integration with Advanced Computational Methods As datasets grow larger and more complex, summary statistics will increasingly be complemented by computationally intensive methods such as bootstrapping, simulation-based inference, and machine learning. These approaches allow for more robust and nuanced summaries, especially in high-dimensional or unstructured data settings. (Garden, 2023)

AI-Driven and Automated Summarization Artificial intelligence, including machine learning and natural language processing, is transforming how summary statistics are generated and interpreted. AI-driven summarization tools can quickly distill massive and complex datasets into actionable insights, improving efficiency, accessibility, and consistency. Future developments are expected to include real-time, personalized, and multimodal summarization, making summary statistics more dynamic and tailored to user needs. (Datatas, 2025)

Enhanced Visualization and Exploratory Data Analysis The role of visualization in summary statistics will continue to grow, leveraging advances in computer graphics and interactive tools. This enables more intuitive and exploratory ways to understand data distributions, trends, and anomalies, moving beyond traditional tables and static plots. (Potter et al., 2010)

Addressing Big Data and Complex Structures Summary statistics will need to adapt to the challenges of big data, including handling massive volumes, varied data types, and complex dependencies (such as networks or time-evolving data). This will require new conceptual frameworks and algorithms capable of summarizing information efficiently at scale.

Ongoing Methodological Innovation The field will continue to see progress in robust inference, regularization, causal inference, and adaptive decision analysis, all of which will influence how summary statistics are computed and applied in practice. (Gelman & Vehtari, 2024)

In Summary The future of summary statistics lies in their evolution from simple descriptive tools to components of sophisticated, automated, and interactive analytical systems. They will play a foundational role in making sense of big, complex, and heterogeneous data, driven by advances in computation, AI, and interdisciplinary collaboration.

Other Sources: (Fan et al., 2014) (Wang et al., 2016)

Conclusion

Summary statistics are fundamental to any data analysis process, serving as the essential first step in understanding and interpreting datasets. In R, summary statistics provide a concise overview of data distributions, central tendencies, and variability, enabling analysts to quickly assess data quality, detect anomalies, and guide subsequent analytical decisions. The flexibility and power of R-through core functions like summary(), mean(), sd(), and packages such as dplyr, to efficiently compute and customize statistical summaries for both ungrouped and grouped data.

By leveraging these tools, R users can automate the calculation of key metrics such as mean, median, standard deviation, range, and percentiles, as well as generate detailed group-wise summaries. This capability not only streamlines exploratory data analysis but also lays a robust foundation for more complex statistical modeling and hypothesis testing. Ultimately, mastering summary statistics in R empowers analysts to derive meaningful insights from data, make informed decisions, and communicate results effectively in any field of research or business analytics. (Lane, 2013)

References

This Section Is an Appendix

Another Appendix

Affadative

I hereby affirm that this submitted paper was authored unaided and solely by me. Additionally, no other sources than those in the reference list were used. Parts of this paper, including tables and figures, that have been taken either verbatim or analogously from other works have in each case been properly cited with regard to their origin and authorship. This paper either in parts or in its entirety, be it in the same or similar form, has not been submitted to any other examination board and has not been published.

I acknowledge that the university may use plagiarism detection software to check my thesis. I agree to cooperate with any investigation of suspected plagiarism and to provide any additional information or evidence requested by the university.

Checklist:

The handout contains 3-5 pages of text.

The submission contains the Quarto file of the handout.

The submission contains the Quarto file of the presentation.

The submission contains the HTML file of the handout.

The submission contains the HTML file of the presentation.

The submission contains the PDF file of the handout.

The submission contains the PDF file of the presentation.

The title page of the presentation and the handout contain personal details (name, email, matriculation number).

The handout contains a abstract.

The presentation and the handout contain a bibliography, created using BibTeX with APA citation style.

Either the handout or the presentation contains R code that proof the expertise in coding.

The handout includes an introduction to guide the reader and a conclusion summarizing the work and discussing potential further investigations and readings, respectively.

All significant resources used in the report and R code development.

The filled out Affidavit.

A concise description of the successful use of Git and GitHub, as detailed here: https://github.com/hubchev/make_a_pull_request.

The link to the presentation and the handout published on GitHub.

Jelin George, 28May2025, Cologne

Datatas. (2025, April 14). *The future of AI-generated data summarization for large reports*. https://datatas.com/the-future-of-ai-generated-data-summarization-for-large-reports/

Fan, J., Han, F., & Liu, H. (2014). Challenges of big data analysis. *National Science Review*, *1*(2), 293–314. https://doi.org/10.1093/nsr/nwt032

Garden, O. (2023, November 27). *The 8 most important statistical ideas: Bootstrapping and simulation-based inference*. https://osc.garden/blog/bootstrapping-and-simulation-based-inference/

Gelman, A., & Vehtari, A. (2024). *What are the most important statistical ideas of the past 50 years?* https://www.stat.columbia.edu/~gelman/research/unpublished/stat50.pdf

Lane, D. M. (2013). Descriptive statistics. In *Introduction to statistics*. Rice University. https://onlinestatbook.com/2/introduction/descriptive.html

Oh, D. M., & Pyrczak, F. (2023). *Making sense of statistics: A conceptual overview*. Routledge.

Potter, K., Kniss, J., Riesenfeld, R., & Johnson, C. R. (2010). Visualizing summary statistics and uncertainty. *Computer Graphics Forum*, *29*(3), 823–832. https://www.sci.utah.edu/~kpotter/publications/potter-2010-VSSU.pdf

Wang, C., Chen, M.-H., Schifano, E., Wu, J., & Yan, J. (2016). Statistical methods and computing for big data. *Statistics and Its Interface*, *9*(4), 399–414. https://doi.org/10.4310/SII.2016.v9.n4.a1

Wickham, H., François, R., Henry, L., Müller, K., & Vaughan, D. (2023). *Dplyr: A grammar of data manipulation*. https://dplyr.tidyverse.org/articles/dplyr.html

Wienclaw, R. A. (2009). The misuse of statistics. *The Research Starters Sociology*, 1–5.

Wienclaw, R. A. (2017). *Hypothesis construction*. Cedarville, OH: Salem Press Encyclopedia.