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

Jelin George

Hochschule Fresenius - University of Applied Science

Author Note

Jelin George https://orcid.org/0000-0000-0000-0001

The authors have no conflicts of interest to disclose.

Correspondence concerning this article should be addressed to Jelin George, Email: george.jelin@stud-hs.fresenius.de

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

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

Table of contents

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

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:

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

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

I would like to highlight a book, *Making sense of statistics: A conceptual overview*, (Oh & Pyrczak, 2023) which offers a clear and accessible introduction to key statistical concepts for beginners. The book focuses on building conceptual understanding of both descriptive and inferential statistics, using simple explanations, practical examples, and step-by-step guidance. It is designed to help students from any discipline gain confidence in applying statistics to research and interpreting data effectively.

Key Measures in Summary Statistics

- **Measures of Central Tendency**: Central tendency measures indicate where most values in a dataset fall.
- **Measures of Dispersion**: Dispersion measures describe the spread of data.
- **Measures of Shape and Distribution**: Describe the overall pattern and characteristics of how data values are distributed within a dataset.
- **Visualization tools**: Histogram and boxplot are tools to help understand distribution of data better.
- **Frequency table**: It shows how often each value or category appears in a dataset, making it easy to spot common or rare values and summarize the data.
- **Cross-tabulations** (contingency tables): It shows how two or more categorical variables are related by displaying the count of observations for each combination of categories.

Watch this **tutorial video** on descriptive statistics in R to get you started. Additionally, the YouTube videos listed here may be helpful for understanding the code chunks (Walker, 2023) (Videos, 2024).

Summarizing Data Frames

We can summarize entire datasets by calculating statistics like mean, median, minimum, maximum, standard deviation, and counts for each variable. This provides a clear overview of the data's structure and key features.

Practical application: Apply summary statistics functions to the Star Wars dataset. Use functions such as summary(), mean(), median(), sd(), and others to explore key variables like height, mass, and age. This will help you quickly understand the central tendencies, variability, and distribution patterns within the Star Wars data frame. Watch this YouTube video Return of the Star Wars dataset to understand the details of the dataset.

Furthermore, read Modern Statistics with R to understand essential tools and techniques in

contemporary statistical data analysis, using the R programming language. The book features numerous examples and over 200 exercises with worked solutions. The online version is freely available and regularly updated, with downloadable datasets for hands-on learning.

Limitations

The future of summary statistics is shaped by growing data complexity, computational advances, and artificial intelligence. As noted in (Garden, 2023), summary statistics are increasingly combined with methods like bootstrapping and machine learning to handle complex data. AI is automating and personalizing summaries (Datatas, 2025), while enhanced visual tools make data exploration more intuitive (Potter et al., 2010). (Fan et al., 2014) highlight the need for new frameworks to address big data challenges, and (Gelman & Vehtari, 2024) emphasize ongoing methodological innovation.

In sum, summary statistics are evolving into dynamic, automated tools essential for understanding large and complex datasets.

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.

According to (Lane, 2013), using R's tools to automate calculations of summary statistics-such as mean, median, standard deviation, range, and percentiles-enables users to efficiently produce both overall and group-wise summaries. This not only streamlines exploratory data analysis but also establishes a strong basis for advanced statistical modeling and hypothesis testing. As Lane emphasizes, mastering summary statistics in R allows analysts to extract meaningful insights, make better decisions, and clearly communicate results across various fields of research and business analytics.

References

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

Videos, D. E. R. (2024). *Gentle r #4: Basic summary statistics in r with r studio [video]*. YouTube. https://www.youtube.com/watch?v=8XFmPP93w_Y

Walker, L. (2023). *Easy summary tables in r with gtsummary [video]*. YouTube. https://www.youtube.com/watch?v=gohF7pp2XCg

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