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

Jelin George (Matriculation Number: 400826617)

Hochschule Fresenius - University of Applied Science

Author Note

The authors have no conflicts of interest to disclose.

Correspondence concerning this article should be addressed to Jelin George (Matriculation Number: 400826617), 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. Practical examples and code are provided to demonstrate these concepts in action. The document also discusses the limitations of summary statistics and its application.

Keywords: summary statistics, R programming, data analysis, descriptive statistics, data visualization, exploratory data analysis, limitations

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

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. R offers a rich ecosystem of functions and packages, each with unique features. Base R provides basic summaries, while dplyr allows flexible, tidy group summaries, skimr and summarytools create detailed, readable overviews, often with visual elements. Packages like psych, Hmisc, and pastecs offer more advanced descriptive statistics. For publication-ready tables, gtsummary and table1 are ideal. Additional tools such as janitor, rstatix, and doBy support quick tabulation and custom summaries, giving users a range of options for data analysis.

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 in applying statistics to research and interpreting data effectively.

For a deeper exploration of R packages used for summarizing data, I encourage you to visit the source by (Medcalf, 2018)

Additonally, watch this **tutorial video** on descriptive statistics in R to get you started.

Key Measures in Summary Statistics

Summary statistics simplify complex datasets into a few key numbers, making it easier to understand and communicate the main characteristics of the data.

The primary measures include central tendency (mean, median, and mode), which describe where most values fall, and measures of dispersion, which capture the spread of the data. Measures of shape and distribution, such as skewness and kurtosis, provide insight into the overall pattern and extremities of the data.

Visualization tools like histograms and boxplots help illustrate these patterns and highlight outliers or skewness. Handling missing data is also important, as different types of missingness (MCAR, MAR, MNAR) can bias results if not addressed through omission or imputation.

Frequency tables and cross-tabulations organize and display how often values or categories occur, making it easier to spot patterns and summarize large datasets. Finally, summarizing entire data frames with these statistics offers a comprehensive overview of each variable, revealing trends and important features within 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

The YouTube videos referenced here may assist in further understanding the code chunks presented above (Walker, 2023) (Videos, 2024) (Schork, 2021)

Practical Application

Try this exercise 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.

```
if (!require(pacman)) install.packages("pacman")
pacman::p_load(tidyverse)
```

Load the Star Wars dataset available in the dylyr package. Read more on dylyr package here (Wickham et al., 2023)

```
library(dplyr)
data(starwars)
```

To begin our analysis, we will display the first 10 rows of the starwars dataset. This provides a quick overview of the data structure and its key variables before we proceed with summary statistics.

```
starwars_tbl <- starwars %>%
  slice_head(n = 10)


kable(starwars_tbl, format = "latex", booktabs = TRUE, caption = "Table 1. Star Wars Dat
  kable_styling(latex_options = "striped", full_width = FALSE)
```

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

```
# A tibble: 1 x 1
```

Table 1

Table 1. Star Wars Data

name	height	mass	hair_color	skin_color	eye_color	birth_year	sex	gender
Luke Skywalker	172	77	blond	fair	blue	19.0	male	masculin
C-3PO	167	75	NA	gold	yellow	112.0	none	masculin
R2-D2	96	32	NA	white, blue	red	33.0	none	masculin
Darth Vader	202	136	none	white	yellow	41.9	male	masculin
Leia Organa	150	49	brown	light	brown	19.0	female	feminine
Owen Lars	178	120	brown, grey	light	blue	52.0	male	masculin
Beru Whitesun Lars	165	75	brown	light	blue	47.0	female	feminine
R5-D4	97	32	NA	white, red	red	NA	none	masculin
Biggs Darklighter	183	84	black	light	brown	24.0	male	masculin
Obi-Wan Kenobi	182	77	auburn, white	fair	blue-gray	57.0	male	masculin

```
  mean_height
        <dbl>
1         175.
```

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

```
# A tibble: 1 x 1
  median_height
          <int>
1           180
```

```
# Mode height


starwars %>%
  filter(!is.na(height)) %>%
  count(height, sort = TRUE) %>%
  slice_max(n = 1, order_by = n) %>%
  select(mode_height = height)
```

```
# A tibble: 1 x 1
  mode_height
        <int>
1         183
```

Now, let's apply other summary functions.

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


# Tidy summary for numeric variables
starwars_selected %>%
  summarise(
    mean_height = mean(height, na.rm = TRUE),
    sd_height = sd(height, na.rm = TRUE),
    mean_mass = mean(mass, na.rm = TRUE),
    sd_mass = sd(mass, na.rm = TRUE)
  )
```

```
# A tibble: 1 x 4
```

```
  mean_height sd_height mean_mass sd_mass
        <dbl>     <dbl>     <dbl>   <dbl>
1         175.      34.8      97.3    169.
```

```
# Frequency table for gender (tidyverse style)
starwars_selected %>%
  count(gender, name = "frequency")
```

```
# A tibble: 3 x 2
  gender    frequency
  <chr>         <int>
1 feminine         17
2 masculine        66
3 <NA>              4
```

```
# Proportion table for gender (tidyverse style)
starwars_selected %>%
  count(gender, name = "frequency") %>%
  mutate(proportion = frequency / sum(frequency))
```

```
# A tibble: 3 x 3
  gender    frequency proportion
  <chr>         <int>      <dbl>
1 feminine         17     0.195
2 masculine        66     0.759
3 <NA>              4     0.0460
```

```
# Comprehensive tidy summary using skimr
skim(starwars_selected)
```

Table 2

Data summary

Name	starwars_selected
Number of rows	87
Number of columns	5
Column type frequency:	
character	2
numeric	3
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
gender	4	0.95	8	9	0	2	0
species	4	0.95	3	14	0	37	0

Variable type: numeric

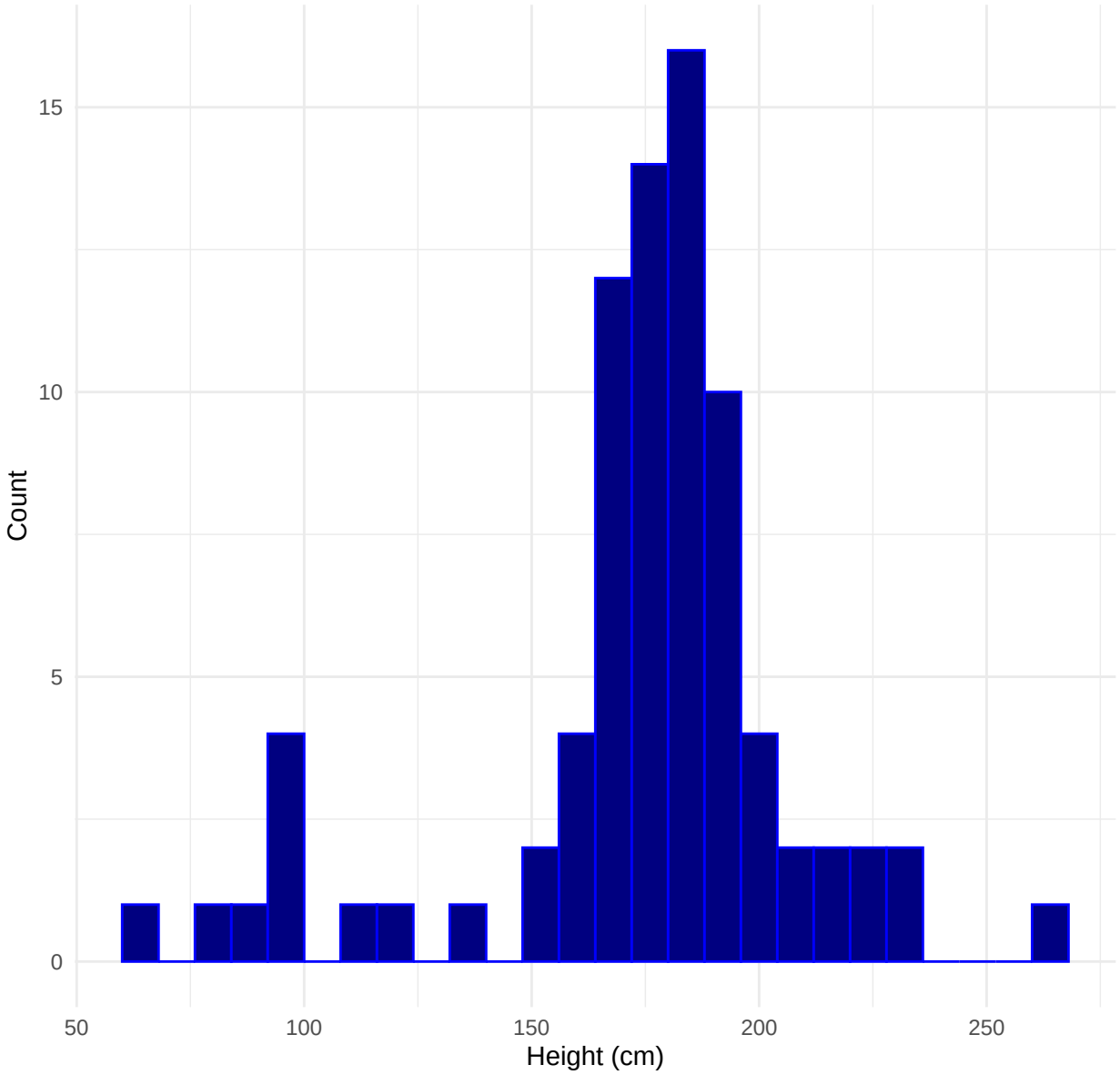
skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
height	6	0.93	174.60	34.77	66	167.0	180	191.0	264	
mass	28	0.68	97.31	169.46	15	55.6	79	84.5	1358	
birth_year	44	0.49	87.57	154.69	8	35.0	52	72.0	896	

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

```
starwars %>%
  ggplot(aes(x = height)) +
  geom_histogram(binwidth = 8, fill = "navy", color = "blue") +
  labs(
    title = "Figure 4. Histogram of Height of Characters in Star Wars",
    x = "Height (cm)",
    y = "Count"
  ) +
  theme_minimal()
```

```
Warning: Removed 6 rows containing non-finite outside the scale range
(`stat_bin()`).
```

Figure 4. Histogram of Height of Characters in Star Wars



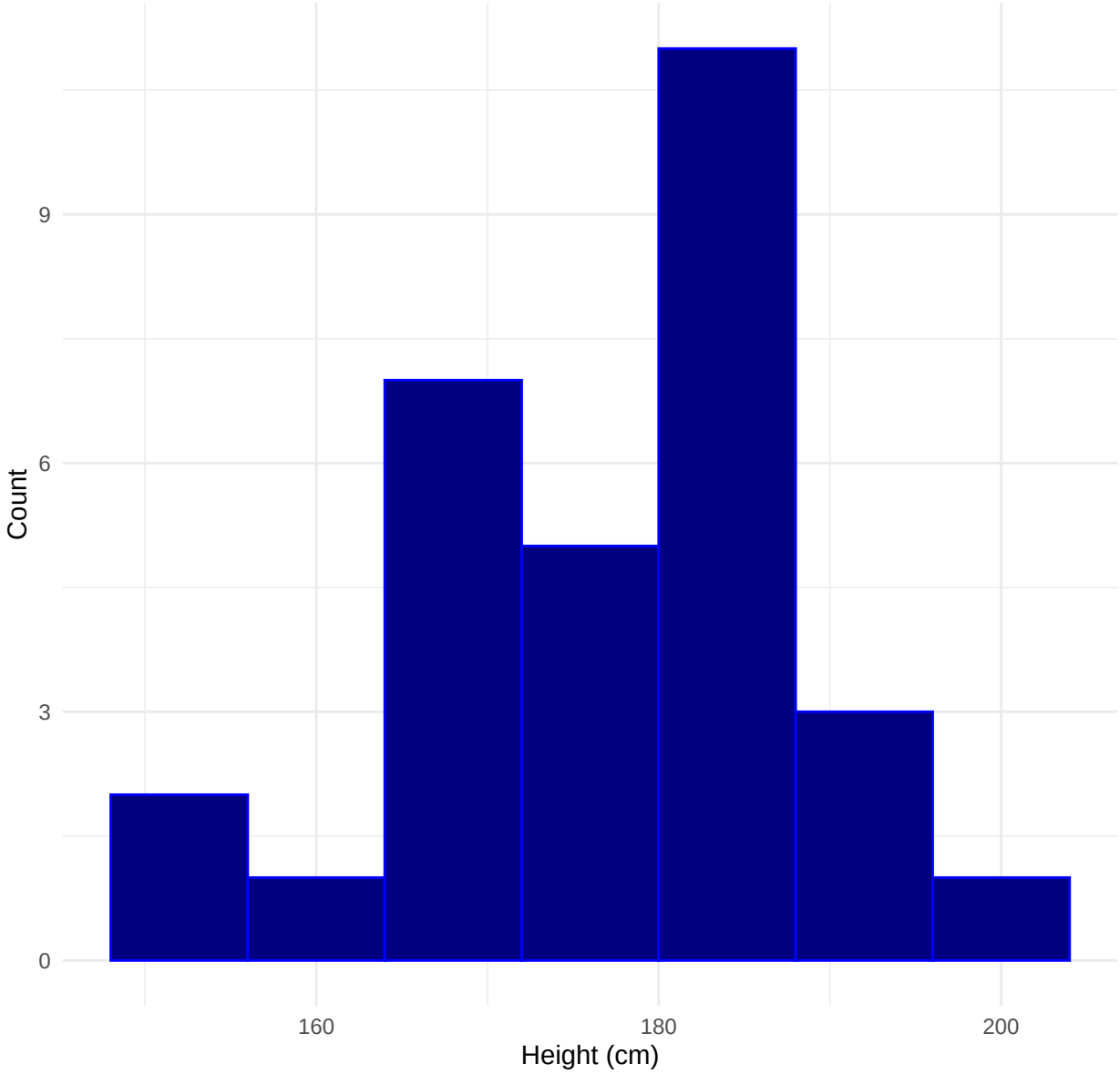
Now, we filter the species to get visual pattern of the height of different *human* characters in Star Wars.

```
starwars %>%
  filter(species == "Human") %>%
  ggplot(aes(x = height)) +
  geom_histogram(binwidth = 8, fill = "navy", color = "blue") +
```

```
  labs(
    title = "Figure 5. Histogram of Height of Human Characters in Star Wars",
    x = "Height (cm)",
    y = "Count"
  ) +
  theme_minimal()
```

```
Warning: Removed 5 rows containing non-finite outside the scale range
(`stat_bin()`).
```

Figure 5. Histogram of Height of Human Characters in Star Wars



Furthermore, this YouTube video Return of the Star Wars dataset may be an interesting resource to help you better understand the dataset.

Limitation

Summary statistics provide a quick and accessible overview of a dataset, but they come with important limitations.

As highlighted in Naked Statistics (Wheelan, 2013), these measures can be misapplied, misinterpreted, or even manipulated, leading to misleading conclusions. Summary statistics only

describe what is present in the data—they do not explain underlying causes, cannot be generalized beyond the sample without further analysis, and offer no predictive power. By condensing complex data into single values, they may obscure important patterns or differences within the data. Additionally, summary statistics do not reveal relationships between variables and can mask issues like bias or subgroup variation.

Therefore, while useful for initial exploration, summary statistics should always be complemented by more detailed analyses and visualizations to avoid oversimplification and misinterpretation (Wienclaw, 2009).

Conclusion

Summary statistics are a vital first step in data analysis, offering a fast and accessible way to understand and interpret datasets. In R, calculating measures like the mean, median, and standard deviation is straightforward, whether for an entire dataset or for specific groups, thanks to built-in functions and powerful packages such as dplyr. Automating these summaries in R, as noted by (Lane, 2013), streamlines your workflow and helps organize your analysis.

However, it is important to recognize the limitations of summary statistics. While they provide useful snapshots, they do not explain underlying causes, predict future outcomes, or reveal relationships between variables. Summary statistics can also obscure important patterns or differences within subgroups and may mask issues like bias or sampling problems, as highlighted in Naked Statistics (Wheelan, 2013). Relying solely on these measures can therefore lead to oversimplification or misinterpretation of your data (Wienclaw, 2009).

For these reasons, summary statistics should be viewed as an essential starting point, but always complemented with more detailed analyses and visualizations to gain a deeper, more accurate understanding of your data. By mastering summary statistics in R—and remaining aware of their limitations—you can uncover valuable insights, make more informed decisions, and communicate your findings clearly in research or business contexts.

References

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

Medcalf, A. (2018). *My favourite r package for: Summarising data*. https://dabblingwithdata.amedcalf.com/2018/01/02/my-favourite-r-package-for-summarising-data/

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

Schork, J. (2021). *How to calculate summary statistics for the columns of a data frame in r (example code)*. YouTube; Statistics Globe. https://www.youtube.com/watch?v=FMRkUqy1Sjw

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

Wheelan, C. (2013). *Naked statistics: Stripping the dread from the data*. W. W. Norton & Company.

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.

Affidative

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, 2025,May 28, Cologne