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

Jelin George (Matriculation: 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: 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 considers future directions in their 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

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. In R, these statistics not only reveal patterns and outliers but also lay the groundwork for informed decision-making.

In this document, we will:

- Define summary statistics and their importance
- Explore key measures and techniques
- Demonstrate practical application with code
- Analyze limitations and consider furture directions

What is Summary Statistics?

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 - 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.

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.

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

Table 1

Comparison of R Summary Packages

Package	Key_Functions	Strengths	Use_Case
Base R	summary(), mean(), sd()	Built-in, quick, basic, not pipeable	General, quick checks
dplyr	summarise(), count()	Tidy, pipeable, flexible, group-wise summaries	Custom summaries in tidyverse workflows
skimr	skim()	Tidy, detailed, sparkline histograms, missing data	Quick, detailed EDA
psych	describe(), describe.by()	Skewness, kurtosis, group summaries	In-depth descriptive stats
summarytools	descr(), dfSummary()	HTML summaries, mini-histograms, missing data info	Publication-ready summaries
Hmisc	describe()	Detailed, percentiles, unique values	Medical, epidemiological data
pastecs	stat.desc()	Wide range of stats, time series support	Broad, one-call summaries
mosaic	favstats(), tally()	Concise, teaching-focused	Teaching, quick summaries
gtsummary	tbl_summary()	Publication-ready tables, auto-detects variable types	Medical, scientific reporting

Key Measures in Summary Statistics

Summary statistics condense complex datasets into a few meaningful numbers, making it easier to understand and communicate data characteristics. The key measures in summary statistics fall into several categories:

Measures of Central Tendency

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

- **Mean**: The arithmetic average of all data points. 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.

```
# Example in R


#| echo: true
#| label: mean-median-mode
data_tbl <- tibble(value = c(2, 4, 4, 4, 5, 7, 9))


data_tbl %>%
  summarise(
    mean = mean(value),
    median = median(value),
    mode = value %>%
      table() %>%
      which.max() %>%
      names() %>%
```

```
      as.numeric()
  )
```

```
# A tibble: 1 x 3
   mean median  mode
  <dbl>  <dbl> <dbl>
1     5      4     4
```

Measures of Dispersion

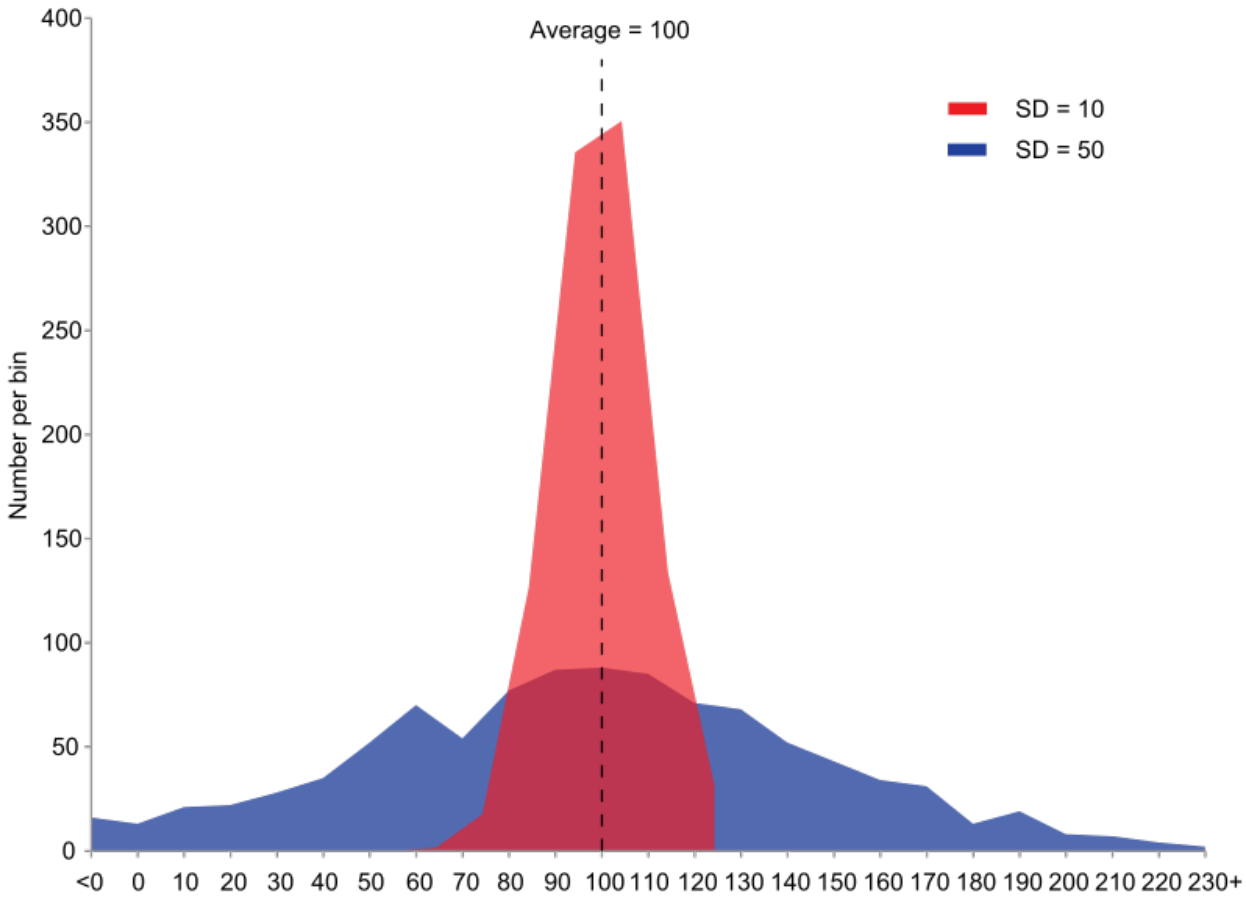
> Dispersion measures describe the spread of data.
>
> *Source: https://en.wikipedia.org/wiki/Statistical_dispersion*

- **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.

```
data_tbl <- tibble(value = c(2, 4, 4, 4, 5, 7, 9))


data_tbl %>%
  summarise(
    min = min(value),
    max = max(value),
    range = max(value) - min(value),
    variance = var(value),
    sd = sd(value),
```

Figure 1

Distributions With Different Dispersion



Note. Example of samples from two populations with the same mean but different dispersion. The blue population is much more dispersed than the red population.

```
    IQR = IQR(value),
    coefficient_of_variation = sd(value) / mean(value)
  )
```

```
# A tibble: 1 x 7
    min   max range variance    sd   IQR coefficient_of_variation
  <dbl> <dbl> <dbl>    <dbl> <dbl> <dbl>                     <dbl>
1     2     9     7     5.33  2.31     2                     0.462
```

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.

```
library(dplyr)


data_tbl <- tibble(value = c(2, 4, 4, 4, 5, 7, 9))


data_tbl %>%
  summarise(
    n = n(),
    mean = mean(value),
    sd = sd(value),
    skewness = sum((value - mean) ^ 3) / n / (sd ^ 3),
    kurtosis = sum((value - mean) ^ 4) / n / (sd ^ 4)
  ) %>%
  select(-mean, -sd, -n) # remove intermediate columns if you only want skewness and kur
```

```
# A tibble: 1 x 2
  skewness kurtosis
     <dbl>    <dbl>
1    0.487     1.79
```

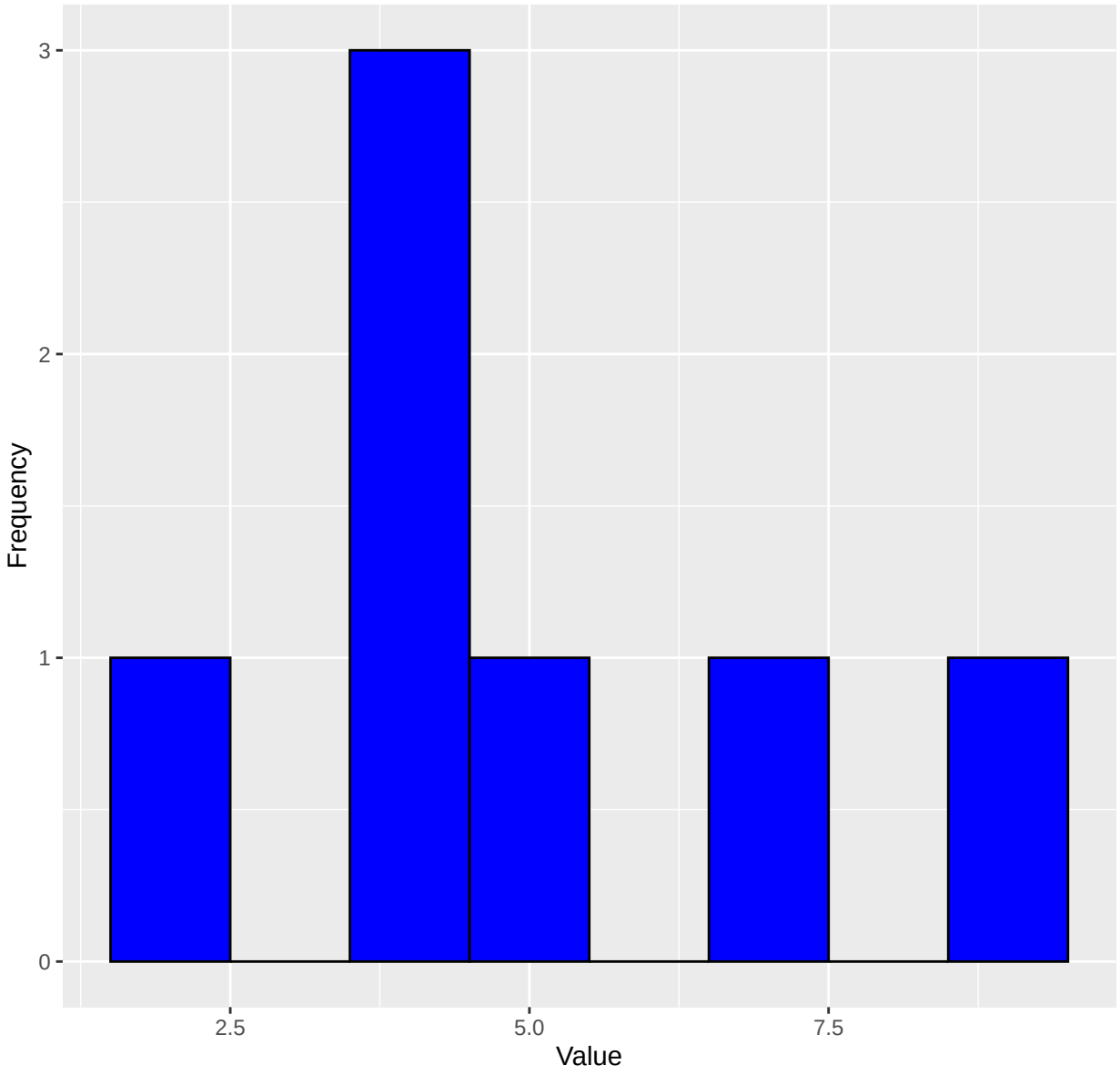
Visualization tools to help understand distribution of data better.

- **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 outliers and the overall spread of the data.

```
# This histogram shows the distribution in the example we set previously.


library(ggplot2)
library(tibble)


#| label: fig-histogram
#| fig-cap: "Histogram of the data"


data_tbl <- tibble(value = c(2, 4, 4, 4, 5, 7, 9))


ggplot(data_tbl, aes(x = value)) +
  geom_histogram(binwidth = 1, fill = "blue", color = "black") +
  labs(
    title = "Figure 2. Histogram of Data",
    x = "Value",
    y = "Frequency"
  )
```

Figure 2. Histogram of Data

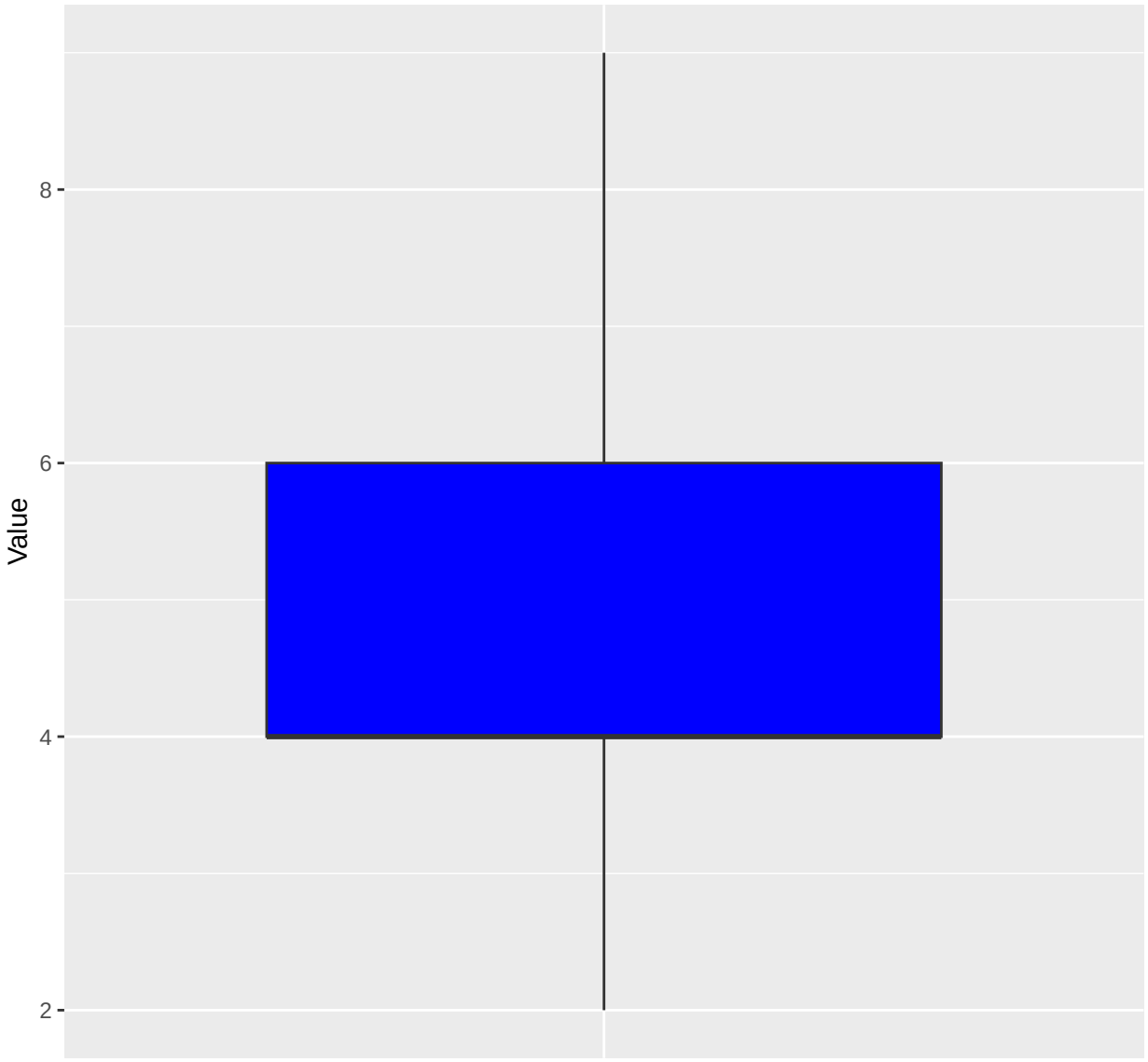


As seen above, the distribution is centered around 4.

- **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
library(ggplot2)
library(tibble)

data_tbl <- tibble(value = c(2, 4, 4, 4, 5, 7, 9))

ggplot(data_tbl, aes(x = factor(1), y = value)) +
  geom_boxplot(fill = "blue") +
  labs(
    title = "Figure 3. Boxplot of Data",
    x = "",
    y = "Value"
  ) +
  theme(axis.text.x = element_blank(),
        axis.ticks.x = element_blank())
```

Figure 3. Boxplot of Data



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_tbl <- tibble(value = c(2, 4, NA, 4, 5, NA, 9))


data_tbl %>%
  summarise(mean = mean(value, na.rm = TRUE)) # Ignore NAs
```

```
# A tibble: 1 x 1
   mean
  <dbl>
1   4.8
```

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
# Example data
age <- c('Young', 'Old', 'Young', 'Old', 'Young', 'Old')
gender <- c('Male', 'Female', 'Female', 'Male', 'Male', 'Female')
data_tbl <- tibble(age = age, gender = gender)


# Frequency table for combinations of age and gender
data_tbl %>%
  count(age, gender, name = "frequency")
```

```
# A tibble: 4 x 3
```

```
  age   gender frequency
  <chr> <chr>      <int>
1 Old   Female         2
2 Old   Male           1
3 Young Female         1
4 Young Male           2
```

```
# Proportion table for combinations of age and gender
data_tbl %>%
  count(age, gender, name = "frequency") %>%
  mutate(proportion = frequency / sum(frequency))
```

```
# A tibble: 4 x 4
  age   gender frequency proportion
  <chr> <chr>      <int>      <dbl>
1 Old   Female         2      0.333
2 Old   Male           1      0.167
3 Young Female         1      0.167
4 Young Male           2      0.333
```

```
# To calculate the total of frequencies
age <- c('Young', 'Old', 'Young', 'Old', 'Young', 'Old')
gender <- c('Male', 'Female', 'Female', 'Male', 'Male', 'Female')
data_tbl <- tibble(age = age, gender = gender)


data_tbl %>%
  count(age, gender, name = "frequency") %>%
  arrange(age, gender) %>%
  mutate(cumulative_frequency = cumsum(frequency))
```

```
# A tibble: 4 x 4
  age   gender frequency cumulative_frequency
  <chr> <chr>      <int>                <int>
1 Old   Female         2                    2
2 Old   Male           1                    3
3 Young Female         1                    4
4 Young Male           2                    6
```

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.

```
data_tbl %>%
  count(age, gender, name = "frequency") %>%
  pivot_wider(names_from = gender, values_from = frequency, values_fill = 0)
```

```
# A tibble: 2 x 3
  age   Female  Male
  <chr>  <int> <int>
1 Old        2     1
2 Young      1     2
```

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.

```
# Install skimr if not already installed
# install.packages("skimr")


df <- tibble(
  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)
)


# Use glimpse for a tidyverse-style structure overview
glimpse(df)
```

```
Rows: 7
Columns: 3
$ Age    <dbl> 21, 22, 22, 23, 24, 25, 25
$ Gender <chr> "F", "M", "M", "F", "F", "M", "M"
$ Score  <dbl> 85, 90, 88, 95, 85, 88, 80
```

```
# Or use skim for a detailed summary
skim(df)
```

Table 2

Data summary

Name	df

Number of rows	7
Number of columns	3
Column type frequency:	
character	1
numeric	2
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
Gender	0	1	1	1	0	2	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
Age	0	1	23.14	1.57	21	22	23	24.5	25	
Score	0	1	87.29	4.68	80	85	88	89.0	95	

You can further explore skimr here.

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

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.

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

Next, 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,
  position = "center",
  font_size = 8
) %>%
column_spec(1:ncol(starwars_tbl), width = "3cm")
```

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

```
# A tibble: 1 x 1
  mean_height
        <dbl>
1        175.
```

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

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

```r
# 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 various 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
```

```r
# 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
```

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

Table 6

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

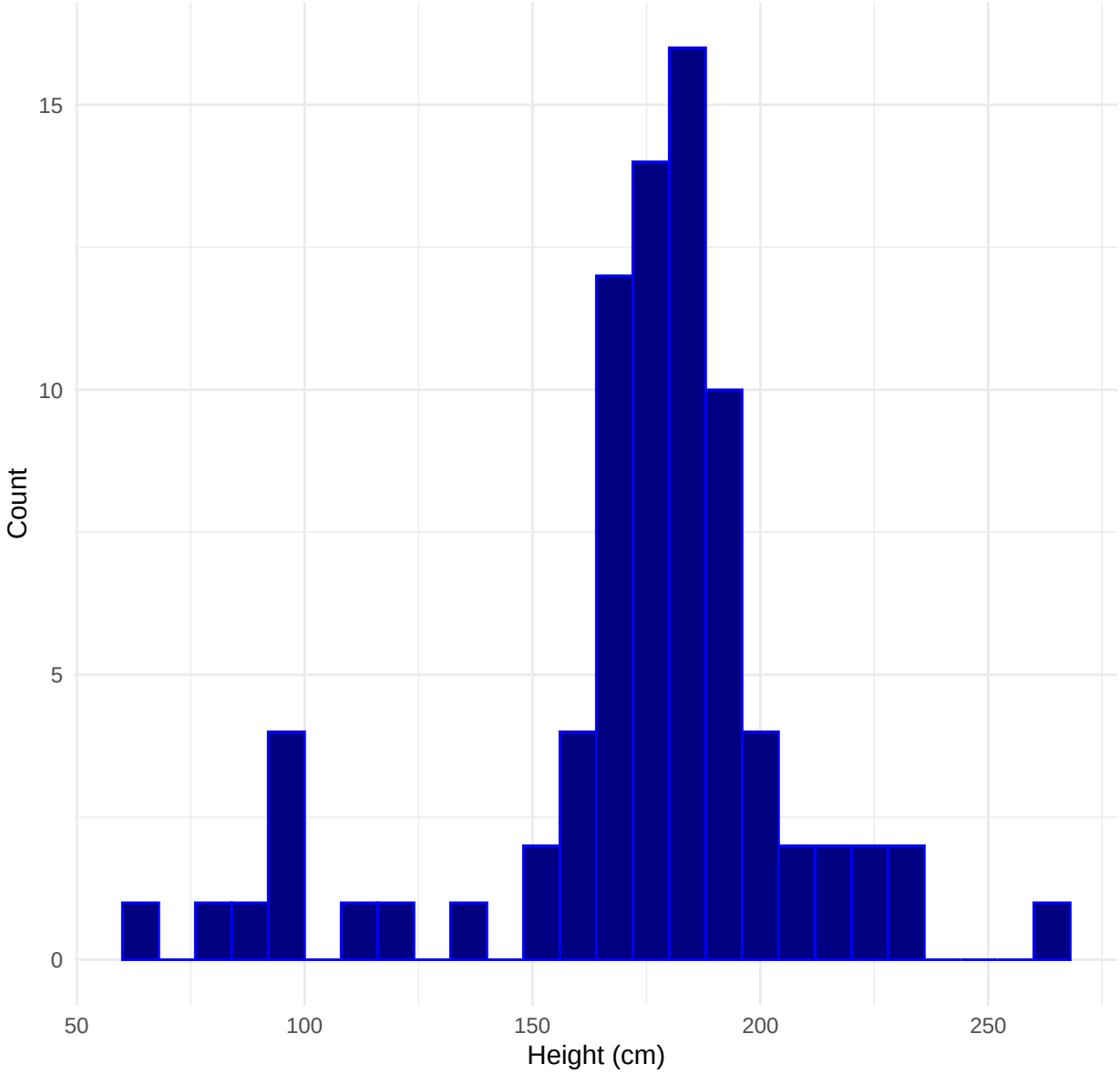
skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
height	6	0.93	174.60	34.77	66	167.0	180	191.0	264	
mass	28	0.68	97.31	169.46	15	55.6	79	84.5	1358	
birth_year	44	0.49	87.57	154.69	8	35.0	52	72.0	896	

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



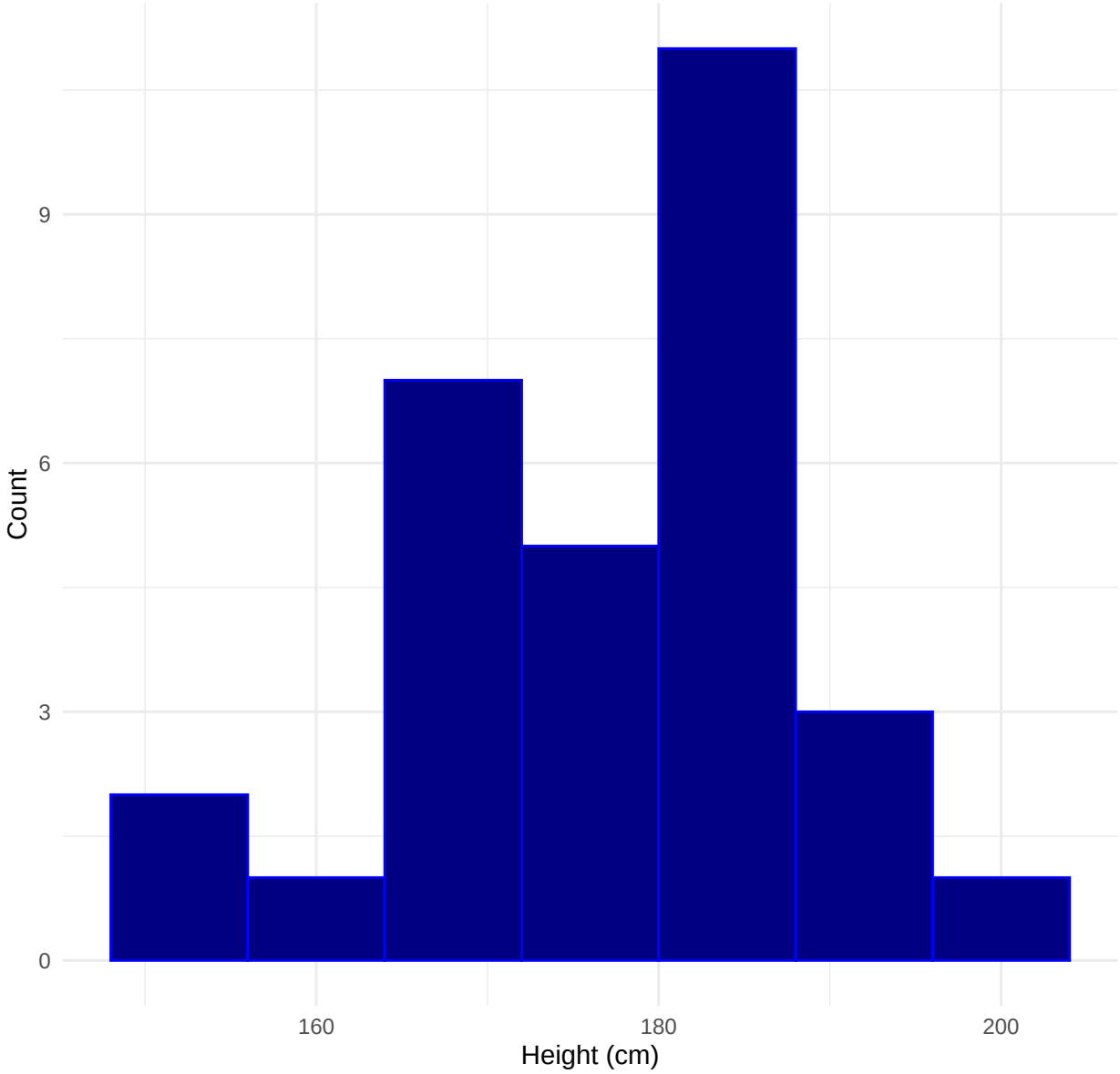
Now, we filter the species to get visual pattern of the height of different *human* characters in Star Wars.

```
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.

Limitations

Summary statistics are essential for providing a quick and accessible overview of a dataset, but they have several important limitations. In the book *Naked Statistics* (Wheelan, 2013), the author highlights key limitations of statistics, warning that statistical measures can be easily misapplied, misinterpreted, or manipulated to mislead people. He explains that while statistics help summarize complex data, this simplification can lead to information loss and oversights, especially when descriptive statistics are mistaken for complete truth. He emphasizes that statistics are only as reliable as the data and methods behind them, and that issues like bias, poor sampling, or careless analysis can produce misleading or false conclusions.

Let's look at some of the limitations in detail:

- 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.

Other sources: (Wienclaw, 2009)

Future Direction

The future of summary statistics is changing as data becomes more complex, computers get faster, and artificial intelligence is used more often. Some important trends are emerging:

- Integration with Advanced Computational Methods: As data gets bigger and more complicated, summary statistics will be used alongside advanced computer methods like bootstrapping, simulations, and machine learning. These tools help create stronger and more detailed summaries, especially when dealing with complex or messy data. (Garden, 2023)

- AI-Driven and Automated Summarization: AI and automation are changing how we create and use summary statistics. Tools that use artificial intelligence can quickly turn large and complex data into useful information, making the process faster and easier. In the future, these tools will likely give real-time and personalized summaries, helping people get the exact information they need. (Datatas, 2025)

- Enhanced Visualization and Exploratory Data Analysis: As computer graphics and interactive tools improve, visualizations will become even more important for summary statistics. This will make it easier to explore and understand data, spot patterns, and find unusual values, going beyond just using tables or simple charts. (Potter et al., 2010)

- Addressing Big Data and Complex Structures: Summary statistics will have to change to handle big data, which includes huge amounts of information, different kinds of data, and complex patterns. New ways of thinking and new tools will be needed to quickly and effectively summarize this information.(Fan et al., 2014)

In summary, summary statistics are becoming more advanced and will be part of smarter, more interactive analysis tools. As data gets bigger and more complex, and as technology and teamwork improve, summary statistics will remain important for understanding information.

Conclusion

Summary statistics are a key starting point for any data analysis, helping us quickly understand and interpret data. In R, these statistics—like the mean, median, and standard deviation—give a clear overview of the data and help spot problems or unusual values. R makes it easy to calculate these numbers for all the data or for different groups, using built-in functions and packages like dplyr.

As (Lane, 2013) explains, using R to automate summary statistics saves time and helps organize your analysis. This basic step is important before moving on to more advanced methods. By mastering summary statistics in R, you can find useful patterns, make better decisions, and clearly share your results in research or business.

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/

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

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.

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

Table 5

Table 1. Star Wars Data

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