

Describing the Data — Descriptive Statistics and Visualization

The First Look: From Raw Data to Understanding

You have successfully navigated the intricate processes of research design, sampling, and data collection. The interviews are transcribed, the survey responses are compiled, the content has been coded, or the experiment is complete. You are now faced with the tangible result of your efforts: a dataset. In its raw form, this dataset is often an intimidating and uncommunicative entity—a spreadsheet with hundreds or thousands of rows of numbers, a folder filled with dense text files, or a collection of coded observations. It holds the answers to your research questions, but its secrets are locked away in a language of raw information. How do you begin to unlock them?

Before we can leap to the complex work of testing hypotheses or making inferences about a population, we must first engage in the fundamental and indispensable act of **description**. This is the essential first step in data analysis, the process of getting to know our data intimately. It is the work of organizing, summarizing, and simplifying the main features of our dataset to understand its basic characteristics. We must understand the landscape of our own data before we can use it as a map to explore the wider world.

This chapter introduces the two primary toolkits for this descriptive task: **descriptive statistics** and **data visualization**. These are not separate or competing activities; they are deeply intertwined and complementary ways of making sense of information. Descriptive statistics provide the tools to summarize our data with precision and concision, using a few key numbers to represent the central patterns and the spread of our observations. Data visualization, in turn, gives us the power to summarize our data with pictures, transforming those numbers into intuitive and powerful graphical forms that can reveal patterns, trends, and outliers that might otherwise remain hidden. This chapter provides a tool-agnostic guide to the conceptual logic of these methods. We will explore how to find the “center” and describe the “spread” of our data, and we will delve into the core principles of creating visualizations that are not just aesthetically pleasing, but are also clear, honest, and insightful. This is the crucial first look at our data, the foundation upon which all subsequent, more complex analyses will be built.

Descriptive Statistics: Summarizing Data with Numbers

The primary goal of descriptive statistics is to take a large and potentially overwhelming set of observations and distill it down to a few manageable and meaningful summary numbers. These statistics provide a quantitative overview of our sample, allowing us to understand its key features at a glance. The two most fundamental types of descriptive statistics are measures of central tendency, which describe the “typical” value in our data, and measures of dispersion, which describe how spread out our data is.

Measures of Central Tendency: Finding the “Center” of the Data

A measure of central tendency is a single score that best represents the center of a distribution. It is the value that we might consider the most typical or representative of the entire set of scores. There are three primary measures of central tendency, and the choice of which one to use depends on the level of measurement of our variable and the shape of our data’s distribution.

The Mean: The Arithmetic Average

The mean is what most people think of as the “average.” It is calculated by summing all the scores in a dataset and dividing by the total number of scores. The mean is the most common measure of central tendency for interval and ratio-level data because it uses every single data point in its calculation, making it a sensitive and comprehensive summary of the entire dataset. It can be thought of as the “balancing point” of the data.

The great strength of the mean is also its primary weakness: its sensitivity to every score. The mean is highly susceptible to the influence of **outliers**, which are extreme values that lie far from the rest of the data. Consider the final exam scores for a small class of ten students: {85, 88, 82, 90, 84, 86, 91, 83, 89, 12}. The first nine scores are tightly clustered in the 80s, but one student received a very low score of 12. The mean of these scores is 79. This “average” score is not very representative of the typical student’s performance, as it has been pulled down significantly by the single outlier.

The Median: The Middle Point

The median is the value that falls in the exact middle of a distribution when all the scores are arranged in rank order from lowest to highest. It is the 50th percentile, the point that splits the data into two equal halves, with 50% of the scores falling above it and 50% falling below it.

The primary advantage of the median is that it is a **resistant measure**, meaning it is not affected by extreme outliers. In our exam score example {12, 82, 83, 84, **85, 86**, 88, 89, 90, 91}, the median is 85.5 (the average of the two middle scores, 85 and 86). This value is a much more accurate and representative summary of the “typical” student’s performance than the mean of 79. For this reason, the median is the preferred measure of central tendency for data that is measured at the ordinal level, and for interval/ratio data that is highly **skewed** (asymmetrical) or contains significant outliers, such as data on income or housing prices.

The Mode: The Most Frequent Value

The mode is the simplest measure of central tendency. It is the value or category that appears most frequently in a dataset. In the set of exam scores {85, 88, 57, 81, 65, 75, 64, 87, 99, 79, 59, 74, 82, 55, 86, 94, 72, 77, 85}, the mode is 85, because it occurs twice while all other scores occur only once.

The mode is the only measure of central tendency that can be used for nominal-level (categorical) data. For example, in a survey of political affiliation, the mode would be the party that was chosen by the most respondents. A dataset can have no mode (if all values occur with equal frequency), one mode (**unimodal**), or multiple modes (**bimodal** or **multimodal**). The presence of two distinct modes in a distribution can be an important finding, as it may suggest that the sample is composed of two different subgroups.

Measures of Dispersion: Describing the “Spread” of the Data

Knowing the center of a distribution is only half the story. Two datasets can have the exact same mean but look completely different. Consider two small classes that both have a mean exam score of 80. In Class A, the scores are {78, 79, 80, 81, 82}. In Class B, the scores are {60, 70, 80, 90, 100}. While their central

tendency is identical, the scores in Class A are tightly clustered around the mean, while the scores in Class B are much more spread out. Measures of dispersion (or variability) are statistics that describe this spread.

The Range: The Simplest Spread

The range is the simplest measure of dispersion, calculated as the difference between the highest and lowest scores in a dataset. In Class A, the range is 4 (82 - 78). In Class B, the range is 40 (100 - 60). The range provides a quick, easy-to-calculate sense of the total spread. However, because it is based on only two data points (the two most extreme scores), it is highly susceptible to outliers and provides a very limited picture of the overall variability.

The Variance and Standard Deviation: The Most Powerful Spread

The variance and standard deviation are the most common and most powerful measures of dispersion. They are used with interval and ratio-level data and are typically reported alongside the mean. Conceptually, the **standard deviation** can be understood as the “average distance of the scores from the mean.” The **variance** is simply the standard deviation squared; it is a crucial statistic for more advanced inferential tests but is less intuitive for descriptive purposes because its units are squared (e.g., “dollars squared”).

A small standard deviation indicates that the data points are tightly clustered around the mean, suggesting a **homogeneous** dataset (like Class A). A large standard deviation indicates that the data points are more spread out, suggesting a **heterogeneous** dataset (like Class B). The standard deviation uses every score in its calculation, making it a sensitive and comprehensive measure of the overall variability in the data.

Data Visualization: Summarizing Data with Pictures

While descriptive statistics provide a precise numerical summary of our data, they can sometimes fail to convey the intuitive, “big picture” understanding that a visual representation can offer. Data visualization is the process of translating numerical data into graphical forms to reveal patterns, trends, and relationships. An effective visualization is not an aesthetic afterthought; it is a crucial part of the analysis and communication process that can communicate a key finding more quickly and powerfully than a paragraph of text.

Core Principles of Effective Visualization

Creating an effective visualization is a craft guided by a set of core principles designed to maximize clarity and minimize distortion. The goal is to create a graphic that is honest, insightful, and easy for your audience to understand.

1. **Show the Data:** The primary goal of any visualization is to present the data clearly. This means focusing on the relevant data points and avoiding unnecessary visual elements—often called “chart junk”—that obscure them. The data itself should be the hero of the graphic.
2. **Reduce the Clutter:** Every element in a chart should serve an informational purpose. Unnecessary elements—such as heavy gridlines, distracting background textures, or misleading 3D effects—should be removed to let the data stand out. As the pioneering designer Edward Tufte advises, maximize the “data-ink ratio.”
3. **Integrate Graphics and Text:** The text in and around a chart is as important as the visual elements. Instead of relying on a separate legend, label data series directly on the chart. Use an “active title” that states the main finding of the chart, like a newspaper headline, rather than a generic description (e.g., “Vaccination Rates Climbed After Campaign Launch” is better than “Figure 1. Vaccination Rates”).

4. **Avoid the “Spaghetti Chart” (Use Small Multiples):** When a single chart becomes too crowded with data (e.g., a line chart with a dozen overlapping lines), it is often better to break it into a series of smaller charts, known as **small multiples** or **panel charts**. These charts all use the same scale and axes but display different subsets of the data, allowing for clear presentation of complex information.
5. **Start with Gray:** This is a powerful practical strategy. Begin designing your chart with all elements in shades of gray. This forces you to make conscious, deliberate decisions about where to use color. Color should be used strategically to highlight the most important information and guide the reader’s attention, not for mere decoration.

A Visual Vocabulary: Choosing the Right Chart for the Job

Different chart types are suited for different analytical tasks. The choice of which chart to use should be driven by the story you want to tell with your data.

Showing a Distribution (for a single variable):

- **Histogram:** This is the classic tool for visualizing a distribution. It is a bar chart that shows the frequency of data points falling into a series of specified intervals, or “bins.” A histogram is excellent for quickly seeing the overall shape of your data—whether it is symmetrical (like a bell curve), skewed, or bimodal.
- **Box-and-Whisker Plot (Boxplot):** This is a compact and powerful summary of a distribution. The “box” shows the middle 50% of the data (the interquartile range), with a line inside marking the median. The “whiskers” extend out to show the range of the data, and individual points are often used to identify potential outliers. Boxplots are especially useful for comparing the distributions of a variable across several different groups.

Comparing Categories:

- **Bar Chart:** This is the workhorse for comparing quantities across discrete categories. The length of the bars is proportional to the value they represent. A crucial rule for bar charts is that the value axis must start at zero to avoid distorting the visual comparison of the bars’ lengths.
- **Dot Plot:** This is an excellent alternative to a bar chart, especially when you have many categories. It uses a simple dot to mark the value for each category, resulting in a cleaner, less ink-heavy graphic.

Showing Change Over Time:

- **Line Chart:** This is the standard for showing trends in a continuous variable over a period of time. The line connects a series of data points, making it easy to see patterns of increase, decrease, and volatility.
- **Slope Chart:** This is a simplified line chart that is perfect for showing the change between just two points in time for multiple categories. It uses a series of lines to connect the starting values on the left to the ending values on the right, clearly showing both the magnitude and direction of change for each category.

Showing a Relationship (between two continuous variables):

- **Scatterplot:** This is the primary tool for visualizing the correlation between two variables. Each case in the dataset is represented by a single dot, plotted according to its values on the horizontal (X) axis and the vertical (Y) axis. The overall pattern of the dots reveals the direction (positive or negative), strength (tightly clustered or widely dispersed), and form (linear or curvilinear) of the relationship.

Showing a Part-to-Whole Relationship:

- **Pie Chart:** While familiar to most audiences, the pie chart is often criticized by data visualization experts because humans are not very good at accurately judging angles and areas. They are best used for a small number of categories (five or fewer) when the goal is to show a simple part-to-whole comparison and the exact values are less important than the general proportions.
- **Stacked Bar Chart or Treemap:** These are often better alternatives to pie charts. A **100% stacked bar chart** can clearly show how the proportional makeup of a whole changes across different categories. A **treemap** uses a series of nested rectangles, where the area of each rectangle is proportional to its value, to show hierarchical part-to-whole relationships.

Tables as Visualizations:

Finally, it is essential to remember that even a simple table is a form of data visualization. A well-designed table can be the most effective way to communicate when the goal is to show precise values. The principles of good design apply here as well: use subtle dividers instead of heavy gridlines, align numbers to the right to make them easy to compare, use white space effectively, and consider adding small visual elements like **heatmaps** (coloring the cells based on their value) or **sparklines** (small, word-sized line charts within a row) to enhance readability and highlight patterns.

Conclusion: The Foundation of Analysis

The process of describing data is the essential first conversation you have with your research findings. It is the foundational stage where you move from a chaotic collection of raw information to a structured and coherent understanding of your sample's basic characteristics. The tools of descriptive statistics—the mean, median, mode, range, and standard deviation—provide the numerical language for this conversation, allowing you to summarize complex patterns with precision. The tools of data visualization provide the graphical language, transforming those numbers into intuitive pictures that can reveal insights and communicate findings with power and clarity.

This descriptive work is not a preliminary chore; it is a fundamental part of the analytical process. It is how we check our assumptions, identify potential problems in our data, and gain the deep familiarity necessary to conduct more advanced analyses responsibly. The insights gained from this first look are the bedrock upon which the inferential claims we will discuss in the next chapter are built.

Journal Prompts

1. Think about a variable you've seen reported often—something like income, grades, or social media followers. Was it reported as an average (mean)? Do you think that number accurately reflected the “typical” case? Based on what you learned in this chapter, would another measure of central tendency (median or mode) have been more appropriate? Why?
2. Describe a time when a graph or chart helped you understand something better than a list of numbers could. What did the visual help reveal? Based on this chapter, which principle of good visualization do you think was at work? If you've seen a bad graph or misleading chart, describe that too—and explain what could have made it clearer.
3. The chapter describes descriptive analysis as a “first conversation” with your data. Why is it essential to fully describe your sample before jumping to conclusions or testing hypotheses? How might skipping this step lead to bad research or misleading claims?

