CHAPTER 2: DESCRIPTIVE STATISTICS

Tabular and Graphical Methods for Summarizing Data

I. The Goal of Summarization

The primary objective of descriptive statistics is to transform raw data into a form that reveals patterns and trends. We distinguish between methods for **Categorical** variables (which focus on proportions) and **Quantitative** variables (which focus on distribution and "shape").

II. Summarizing Categorical Data

Categorical data is summarized by counting how many elements fall into each bin.

- **Frequency Distribution:** A table showing the number (frequency) of items in each non-overlapping category.
- **Relative Frequency:** $RF = f/n$.
- **Bar Chart:** A graphical display where the categories are on one axis and frequency (or relative frequency) is on the other. **Note:** Bars do not touch.
- **Pie Chart:** A circle divided into sectors that represent the relative frequency of each category.

III. Summarizing Quantitative Data

Because quantitative data consists of numbers, we must first group them into **intervals** (or "bins") to make sense of them.

1. Frequency Distributions

To create a frequency distribution for quantitative data, you must define: 1. **Number of bins:** Usually between 5 and 20. 2. **Width of bins:** $\frac{\text{Largest Value} - \text{Smallest Value}}{\text{Number of Bins}}$. 3. **Bin Limits:** Every data point must belong to exactly one bin.

2. Cumulative Distributions

These show the number of items with a value *less than or equal to* the upper limit of each bin.

- **Cumulative Relative Frequency:** The sum of relative frequencies for all bins up to a certain point.

3. Graphical Displays for Numbers

- **Dot Plot:** A horizontal axis shows the range of data; a dot is placed above the axis for each observation.
- **Histogram:** Similar to a bar chart, but used for quantitative data. **Note:** The bars *touch* to show the continuous nature of the data.
- **Stem-and-Leaf Display:** Shows both the rank-order and shape of the data while preserving the actual numerical values.

IV. Two-Variable Summarization

In business, we often want to know how two variables relate (e.g., does advertising spend affect sales?).

1. Crosstabulation

A tabular summary of data for two variables. One variable forms the rows, and the other forms the columns. This helps identify if a relationship exists between them.

2. Scatter Diagram

A graphical display of the relationship between two quantitative variables. One variable is on the x-axis (independent) and one is on the y-axis (dependent).

V. Best Practices for Displays

A "pretty" chart is not always an effective one. Follow these principles:

- **Keep it Simple:** Avoid "chart junk" like 3D effects or distracting backgrounds.
- **Use Clear Labels:** Axes should always be labeled with units (e.g., "Revenue in $").
- **Proportionality:** Ensure the visual size of the bars/sectors corresponds accurately to the data values.

VI. Step-by-Step Example

The Problem: You have five test scores: 72, 75, 78, 85, 92. Create a Stem-and-Leaf display.

The Logic: 1. **Identify the Stems:** The "tens" digit will be our stem (7, 8, 9). 2. **Identify the Leaves:** The "ones" digit will be our leaf. 3. **Construct:**

- 7 | 2, 5, 8
- 8 | 5
- 9 | 2

This shows us that most scores are in the 70s, giving us an instant visual of the "shape" of the class performance.

VII. Practice Set

1. You have a dataset of 200 values ranging from 10 to 90. If you want 8 bins of equal width, what should the width of each bin be?

2. What is the main difference between a Bar Chart and a Histogram?

3. A crosstabulation shows that "Customer Satisfaction" increases as "Response Time" decreases. Is this a positive or negative relationship on a scatter diagram?

4. Calculate the Cumulative Relative Frequency for the following: Bin A (RF=0.20), Bin B (RF=0.35), Bin C (RF=0.15).

VIII. Answer Key

1. $(90 - 10)/8 = \mathbf{10}$. 2. Bar Charts are for **Categorical** data (bars don't touch); Histograms are for **Quantitative** data (bars do touch). 3. **Negative Relationship** (as one goes up, the other goes down). 4. Bin A: 0.20; Bin B: 0.55; Bin C: **0.70**.