UNIT 12
PREPARATION AND TABULATION OF DATA

This unit covers the essential steps taken after data collection (fieldwork) and before statistical analysis. The goal of data preparation is to ensure the raw data is accurate, consistent, and organized for analysis.

1. DATA PREPARATION PROCESS

Data preparation is a crucial sequence of steps that transforms raw survey responses into a structured data matrix ready for statistical processing.

Step	Focus	Goal
Questionnaire Checking	Initial review of forms.	Ensure forms are complete and usable.
Data Editing	Correction of errors/omissions.	Maximize accuracy and consistency.
Coding	Assigning numerical values.	Convert responses into computer-readable codes.
Data Entry	Inputting coded data.	Create a standardized digital file (database).
Data Cleaning	Final consistency checks.	Identify and treat missing responses, outliers, and errors.

2. QUESTIONNAIRE CHECKING

The initial screening of completed questionnaires to identify those that are unusable or contain critical errors. Questionnaires are typically checked for:

1. **Completeness:** Were all relevant sections filled out?
2. **Legibility:** Are the answers clear and readable (for paper surveys)?
3. **Consistency:** Do the answers make logical sense (e.g., did a respondent answer 'Yes' to owning a boat, but skip all questions about boat usage)?

This is the first quality control step: tossing out surveys that are half-empty, illegible, or obviously filled out incorrectly (like a student who answered every question with the same number).

3. DATA EDITING

Data editing is the process of reviewing the data (either manually or electronically) to ensure that the data is accurate, consistent with the intent of the question, and that the data from all respondents is comparable.

- **Treatment of Unsatisfactory Responses:** If a response is missing or unclear, the editor must decide whether to:
    - Return the questionnaire to the field to re-contact the respondent (ideal but often costly).

- Assign the missing value based on the average answer of other respondents (**imputation**).
- Disregard the case entirely (if many critical questions are missing).

This is the correction phase. If a respondent skipped a question, the editor fills it in using an average value (if possible) or assigns a special code for "missing." If an answer makes no sense, it is flagged or corrected based on context.

4. VARIABLE DEVELOPMENT

Variable development (or re-coding) involves creating new variables by mathematically combining, transforming, or aggregating existing ones. This is typically done to capture a complex marketing concept that wasn't measured by a single question.

Instead of analyzing 10 separate questions about brand image, you combine them into one new, more powerful score called a Brand Image Index. You are simplifying complex data into meaningful measures for analysis.

- *Example:* Creating a "High Loyalty" variable by combining respondents who scored on the "Satisfaction" scale AND answered 'Very Likely' to the "Repurchase Intention" question.

5. CODING

Coding is the assignment of a code, usually a number, to represent a specific response to a question. It is necessary to convert qualitative or textual responses into quantitative, computer-friendly forms.

- **Coding Structured Questions (Closed-ended):** Usually simple (e.g., Gender: 1=Male, 2=Female).
- **Coding Unstructured Questions (Open-ended):** Requires establishing a **code book** (a dictionary of common responses) after reviewing a sample of the answers. Responses are then grouped and assigned numerical codes (e.g., 10=Price, 11=Quality, 12=Customer Service).

A computer can't process the word "Disappointed." Coding turns that word into a number (like −2 or 1) so the software can perform statistics on it.

6. CATEGORIZATION

Categorization involves grouping or binning responses or scale values into a smaller number of distinct categories for simpler analysis and presentation (e.g., creating frequency tables). This is often applied to ratio or interval data.

Taking a precise measure like Age (e.g., 22,23,24,25,…) and grouping it into broader, more manageable Age Groups (e.g., 18-24, 25-34, 35-44). This makes it easier to present findings in charts and reports.

7. DATA ENTRY

The mechanical process of transferring the coded data from questionnaires or coding sheets into a computer file. This is typically done using statistical software (like SPSS, R, or Excel).

- **Data Cleaning:** This is the essential final check immediately following data entry. It involves two main checks:

    1. **Consistency Check:** Identifying data that is logically inconsistent (e.g., a respondent aged 100 claiming to have 10 children under age 5).

    2. **Handling Missing Values:** Determining the best way to deal with blanks (e.g., using a mean replacement, or marking it as '99' to signify 'missing').

Typing all the coded numbers into the spreadsheet. The most important part is the final check (Data Cleaning) to find and fix typos or impossible numbers before the analysis starts.

8. DATA MINING

While Data Mining is technically an analysis technique, it often relies heavily on the quality of data preparation, especially when using internal secondary data (as discussed in Unit 5).

Data Mining is the use of sophisticated analytical tools and techniques (like machine learning and predictive modeling) to discover valid, novel, and ultimately understandable patterns in large, existing databases.

This is advanced, automated pattern recognition. Instead of just counting answers (like in a basic survey), data mining is used to find *hidden relationships* in massive amounts of existing sales data, such as finding which customer traits reliably predict product churn.