Lesson Eight

Data Analysis: Frequency Distribution, Cross-tabulation, and Hypothesis Testing
OBJECTIVES

1. Discuss the nature and scope of data preparation and the data preparation process.
2. Explain questionnaire checking and editing, and treatment of unsatisfactory responses by returning to the field, assigning missing values, and discarding unsatisfactory responses.
3. Describe the guidelines for coding questionnaires, including the coding of structured and unstructured questions.
4. Discuss the data-cleaning process and the methods used to treat missing responses: substitution of a neutral value, imputed response, casewise deletion, and pairwise deletion.
5. State the reasons for and methods of statistically adjusting data: weighting, variable respecification, and scale transformation.
6. Describe the procedure for selecting a data analysis strategy and the factors influencing the process.
7. Classify statistical techniques and give a detailed classification of univariate techniques as well as a classification of multivariate techniques.
8. Explain the use of the Internet and statistical software in data preparation and analysis.
OBJECTIVES

9. Describe the significance of preliminary data analysis and the insights that can be obtained from such an analysis.

10. Discuss data analysis associated with frequencies including measures of location, measures of variability, and measures of shape.


12. Describe data analysis associated with parametric hypothesis testing for one sample, two independent samples, and paired samples.

13. Understand data analysis associated with nonparametric hypothesis testing for one sample, two independent samples, and paired samples.
Chapter Outline

1) Overview
2) The Data Preparation Process
3) Questionnaire Checking
4) Editing
5) Coding
6) Transcribing
7) Data Cleaning
8) Statistically Adjusting the Data
9) Selecting a Data Analysis Strategy
10) A Classification of Statistical Techniques
Chapter Outline

11) Overview
12) Frequency Distribution
13) Statistics Associated with Frequency Distribution
14) Introduction to Hypothesis Testing
15) A General Procedure for Hypothesis Testing
16) Cross-Tabulations
17) Statistics Associated with Cross-Tabulation
18) Cross-Tabulation in Practice
19) Hypothesis Testing Related to Differences
20) Parametric Tests
21) Non-parametric Tests
22) Summary
Data Preparation Process

Fig. 14.1

1. Prepare Preliminary Plan of Data Analysis
2. Check Questionnaire
3. Edit
4. Code
5. Transcribe
6. Clean Data
7. Statistically Adjust the Data
8. Select Data Analysis Strategy
A questionnaire returned from the field may be unacceptable for several reasons.

- Parts of the questionnaire may be incomplete.
- The pattern of responses may indicate that the respondent did not understand or follow the instructions.
- The responses show little variance.
- One or more pages are missing.
- The questionnaire is received after the preestablished cutoff date.
- The questionnaire is answered by someone who does not qualify for participation.
Editing

Treatment of Unsatisfactory Results

- **Returning to the Field** – The questionnaires with unsatisfactory responses may be returned to the field, where the interviewers recontact the respondents.

- **Assigning Missing Values** – If returning the questionnaires to the field is not feasible, the editor may assign missing values to unsatisfactory responses.

- **Discarding Unsatisfactory Respondents** – In this approach, the respondents with unsatisfactory responses are simply discarded.
Coding means assigning a code, usually a number, to each possible response to each question. The code includes an indication of the column position (field) and data record it will occupy.

**Coding Questions**

- **Fixed field codes**, which mean that the number of records for each respondent is the same and the same data appear in the same column(s) for all respondents, are highly desirable.
- If possible, standard codes should be used for missing data. Coding of structured questions is relatively simple, since the response options are predetermined.
- In questions that permit a large number of responses, each possible response option should be assigned a separate column.
Guidelines for Coding Unstructured Questions:

• Category codes should be mutually exclusive and collectively exhaustive.
• Only a few (10% or less) of the responses should fall into the “other” category.
• Category codes should be assigned for critical issues even if no one has mentioned them.
• Data should be coded to retain as much detail as possible.
A **codebook** contains coding instructions and the necessary information about variables in the data set. A codebook generally contains the following information:

- column number
- record number
- variable number
- variable name
- question number
- instructions for coding
The respondent code and the record number appear on each record in the data.

The first record contains the additional codes: project code, interviewer code, date and time codes, and validation code.

It is a good practice to insert blanks between parts.
<table>
<thead>
<tr>
<th>ID</th>
<th>PREFER.</th>
<th>QUALITY</th>
<th>QUANTITY</th>
<th>VALUE</th>
<th>SERVICE</th>
<th>INCOME</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>5</td>
<td>6</td>
<td>5</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>7</td>
<td>6</td>
<td>6</td>
<td>5</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>9</td>
<td>7</td>
<td>6</td>
<td>7</td>
<td>6</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>11</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>12</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>13</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>14</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>15</td>
<td>7</td>
<td>7</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>16</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>17</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>18</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>19</td>
<td>7</td>
<td>5</td>
<td>5</td>
<td>7</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>20</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>
### Table 14.2

<table>
<thead>
<tr>
<th>Name</th>
<th>Type</th>
<th>Width</th>
<th>Decimals</th>
<th>Label</th>
<th>Values</th>
<th>Missing</th>
<th>Columns</th>
<th>Align</th>
<th>Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID</td>
<td>Numeric</td>
<td>8</td>
<td>0</td>
<td>Respondent Number</td>
<td>None</td>
<td>None</td>
<td>8</td>
<td>Right</td>
<td>Scale</td>
</tr>
<tr>
<td>PREFERE</td>
<td>Numeric</td>
<td>8</td>
<td>0</td>
<td>Restaurant Preference</td>
<td>{1, Weak Preference}</td>
<td>None</td>
<td>11</td>
<td>Right</td>
<td>Scale</td>
</tr>
<tr>
<td>QUALITY</td>
<td>Numeric</td>
<td>8</td>
<td>0</td>
<td>Quality of Food</td>
<td>{1, Poor}</td>
<td>None</td>
<td>10</td>
<td>Right</td>
<td>Scale</td>
</tr>
<tr>
<td>QUANTITY</td>
<td>Numeric</td>
<td>8</td>
<td>0</td>
<td>Quantity of Portions</td>
<td>{1, Poor}</td>
<td>None</td>
<td>10</td>
<td>Right</td>
<td>Scale</td>
</tr>
<tr>
<td>VALUE</td>
<td>Numeric</td>
<td>8</td>
<td>0</td>
<td>Overall Value</td>
<td>{1, Poor}</td>
<td>None</td>
<td>10</td>
<td>Right</td>
<td>Scale</td>
</tr>
<tr>
<td>SERVICE</td>
<td>Numeric</td>
<td>8</td>
<td>0</td>
<td>Restaurant service</td>
<td>{1, Poor}</td>
<td>None</td>
<td>10</td>
<td>Right</td>
<td>Scale</td>
</tr>
<tr>
<td>INCOME</td>
<td>Numeric</td>
<td>8</td>
<td>0</td>
<td>Annual Household Income</td>
<td>{1, Less than $20,000}</td>
<td>None</td>
<td>10</td>
<td>Right</td>
<td>Scale</td>
</tr>
</tbody>
</table>

No
### Codebook Excerpt

<table>
<thead>
<tr>
<th>Column Number</th>
<th>Variable Number</th>
<th>Variable Name</th>
<th>Question Number</th>
<th>Coding Instructions</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>ID</td>
<td>1</td>
<td>1 to 20 as coded</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>Preference</td>
<td>1</td>
<td>Input the number circled. 1=Weak Preference 7=Strong Preference</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>Quality</td>
<td>2</td>
<td>Input the number circled. 1=Poor 7=Excellent</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>Quantity</td>
<td>3</td>
<td>Input the number circled. 1=Poor 7=Excellent</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>Value</td>
<td>4</td>
<td>Input the number circled. 1=Poor 7=Excellent</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>Service</td>
<td>5</td>
<td>Input the number circled. 1=Poor 7=Excellent</td>
</tr>
<tr>
<td>Column Number</td>
<td>Variable Number</td>
<td>Variable Name</td>
<td>Question Number</td>
<td>Coding Instructions</td>
</tr>
<tr>
<td>---------------</td>
<td>-----------------</td>
<td>---------------</td>
<td>-----------------</td>
<td>---------------------</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
<td>Income</td>
<td>6</td>
<td>Input the number circled</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1 = Less than $20,000</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2 = $20,000 to 34,999</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3 = $35,000 to 49,999</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4 = $50,000 to 74,999</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5 = $75,000 to 99,999</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>6 = $100,00 or more</td>
</tr>
</tbody>
</table>
Finally, in this part of the questionnaire we would like to ask you some background information for classification purposes.

**PART D**

**Record #**

1. **This questionnaire was answered by**
   1. _____ Primarily the male head of household
   2. _____ Primarily the female head of household
   3. _____ Jointly by the male and female heads of household

2. **Marital Status**
   1. _____ Married
   2. _____ Never Married
   3. _____ Divorced/Separated/Widowed

3. **What is the total number of family members living at home? _____**

4. **Number of children living at home:**
   a. Under six years _____
   b. Over six years _____

5. **Number of children not living at home _____**

6. **Number of years of formal education which you (and your spouse, if applicable) have completed. (please circle)**

<table>
<thead>
<tr>
<th>High School</th>
<th>Undergraduate</th>
<th>Graduate</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. You</td>
<td>8 or less 9 10 11 12</td>
<td>13 14 15 16</td>
</tr>
<tr>
<td>b. Spouse</td>
<td>8 or less 9 10 11 12</td>
<td>13 14 15 16</td>
</tr>
</tbody>
</table>

7. **a. Your age:**
   **b. Age of spouse (if applicable)**

8. **If employed please indicate your household's occupations by checking the appropriate category.**

<table>
<thead>
<tr>
<th>44</th>
<th>45</th>
<th>Male Head</th>
<th>Female Head</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Professional and technical</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.</td>
<td>Managers and administrators</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.</td>
<td>Sales workers</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4.</td>
<td>Clerical and kindred workers</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5.</td>
<td>Craftsman/operative/laborers</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6.</td>
<td>Homemakers</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7.</td>
<td>Others (please specify)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8.</td>
<td>Not applicable</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

9. **Is your place of residence presently owned by household?**
   1. Owned _____
   2. Rented _____

10. **How many years have you been residing in the greater Atlanta area? _____ years.**
Data Transcription

Fig. 14.4

- Raw Data
  - CATI/CAPI
  - Keypunching via CRT Terminal
    - Verification: Correct Keypunching Errors
      - Computer Memory
  - Optical Recognition
  - Digital Tech.
  - Bar Code & Other Technologies
    - Disks
    - Other Storage

Transcribed Data
Data Cleaning Consistency Checks

**Consistency checks** identify data that are out of range, logically inconsistent, or have extreme values.

- Computer packages like SPSS, SAS, EXCEL and MINITAB can be programmed to identify out-of-range values for each variable and print out the respondent code, variable code, variable name, record number, column number, and out-of-range value.

- Extreme values should be closely examined.
Data Cleaning Treatment of Missing Responses

- **Substitute a Neutral Value** – A neutral value, typically the mean response to the variable, is substituted for the missing responses.

- **Substitute an Imputed Response** – The respondents' pattern of responses to other questions are used to impute or calculate a suitable response to the missing questions.

- In **casewise deletion**, cases, or respondents, with any missing responses are discarded from the analysis.

- In **pairwise deletion**, instead of discarding all cases with any missing values, the researcher uses only the cases or respondents with complete responses for each calculation.
Statistically Adjusting the Data Weighting

• In **weighting**, each case or respondent in the database is assigned a weight to reflect its importance relative to other cases or respondents.

• Weighting is most widely used to make the sample data more representative of a target population on specific characteristics.

• Yet another use of weighting is to adjust the sample so that greater importance is attached to respondents with certain characteristics.
## Statistically Adjusting the Data

### Use of Weighting for Representativeness

<table>
<thead>
<tr>
<th>Years of Education</th>
<th>Sample Percentage</th>
<th>Population Percentage</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Elementary School</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0 to 7 years</td>
<td>2.49</td>
<td>4.23</td>
<td>1.70</td>
</tr>
<tr>
<td>8 years</td>
<td>1.26</td>
<td>2.19</td>
<td>1.74</td>
</tr>
<tr>
<td><strong>High School</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 to 3 years</td>
<td>6.39</td>
<td>8.65</td>
<td>1.35</td>
</tr>
<tr>
<td>4 years</td>
<td>25.39</td>
<td>29.24</td>
<td>1.15</td>
</tr>
<tr>
<td><strong>College</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 to 3 years</td>
<td>22.33</td>
<td>29.42</td>
<td>1.32</td>
</tr>
<tr>
<td>4 years</td>
<td>15.02</td>
<td>12.01</td>
<td>0.80</td>
</tr>
<tr>
<td>5 to 6 years</td>
<td>14.94</td>
<td>7.36</td>
<td>0.49</td>
</tr>
<tr>
<td>7 years or more</td>
<td>12.18</td>
<td>6.90</td>
<td>0.57</td>
</tr>
<tr>
<td><strong>Totals</strong></td>
<td><strong>100.00</strong></td>
<td><strong>100.00</strong></td>
<td></td>
</tr>
</tbody>
</table>
Variable respecification involves the transformation of data to create new variables or modify existing variables.

e.g., the researcher may create new variables that are composites of several other variables.

Dummy variables are used for respecifying categorical variables. The general rule is that to respecify a categorical variable with $K$ categories, $K-1$ dummy variables are needed.
Statistically Adjusting the Data – Variable Respecification

<table>
<thead>
<tr>
<th>Product Usage Category</th>
<th>Original Variable Code</th>
<th>Dummy Variable Code</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$X_1$   $X_2$ $X_3$</td>
</tr>
<tr>
<td>Nonusers</td>
<td>1</td>
<td>1          0       0</td>
</tr>
<tr>
<td>Light users</td>
<td>2</td>
<td>0          1       0</td>
</tr>
<tr>
<td>Medium users</td>
<td>3</td>
<td>0          0       1</td>
</tr>
<tr>
<td>Heavy users</td>
<td>4</td>
<td>0          0       0</td>
</tr>
</tbody>
</table>

Note that $X_1 = 1$ for nonusers and 0 for all others. Likewise, $X_2 = 1$ for light users and 0 for all others, and $X_3 = 1$ for medium users and 0 for all others. In analyzing the data, $X_1$, $X_2$, and $X_3$ are used to represent all user/nonuser groups.
Scale transformation involves a manipulation of scale values to ensure comparability with other scales or otherwise make the data suitable for analysis.

A more common transformation procedure is standardization. Standardized scores, $Z_i$, may be obtained as: $Z_i = (X_i - \bar{X})/s_x$
Selecting a Data Analysis Strategy

Fig. 14.5

- Earlier Steps (1, 2, & 3) of the Marketing Research Process
- Known Characteristics of the Data
- Properties of Statistical Techniques
- Background and Philosophy of the Researcher
- Data Analysis Strategy
A Classification of Univariate Techniques

Fig. 14.6

Univariate Techniques

Metric Data
- One Sample
  - Independent
    - Two-Group test
    - Z test
    - One-Way ANOVA
  - Related
    - Paired t test
- Two or More Samples
  - Independent
  - Related

Non-numeric Data
- One Sample
  - Frequency
  - Chi-Square
  - K-S
  - Runs
  - Binomial
- Two or More Samples
  - Independent
  - Related
    - Sign
    - Wilcoxon
    - McNemar
    - Chi-Square
    - Mann-Whitney
    - Median
    - K-S
    - K-W ANOVA
A Classification of Multivariate Techniques

Fig. 14.7

Multivariate Techniques

Dependence Technique

One Dependent Variable
* Cross-Tabulation
* Analysis of Variance and Covariance
* Multiple Regression
* 2-Group Discriminant/Logit
* Conjoint Analysis

More Than One Dependent Variable
* Multivariate Analysis of Variance
* Canonical Correlation
* Multiple Discriminant Analysis
* Structural Equation Modeling and Path Analysis

Interdependence Technique

Variable Interdependence
* Factor Analysis
* Confirmatory Factor Analysis

Interobject Similarity
* Cluster Analysis
* Multidimensional Scaling
The Nielsen Media Research Company, a longtime player in television-related marketing research, has come under fire from the various TV networks for its surveying techniques. Additionally, in another potentially large, new revenue business, Internet surveying, Nielsen is encountering serious questions concerning the validity of its survey results. Due to the tremendous impact of electronic commerce on the business world, advertisers need to know how many people are doing business on the Internet in order to decide if it would be lucrative to place their ads online.

Nielsen performed a survey for CommerceNet, a group of companies that includes Sun Microsystems and American Express, to help determine the number of total users on the Internet.
Statisticians believe the numbers reported by Nielsen may be incorrect in that the weighting used to help match the sample to the population may be flawed. Weighting must be used to prevent research from being skewed toward one demographic segment. Nielsen weighted for gender but not for education which may have skewed the population toward educated adults.
Nielsen then weighted the survey by age and income after they had already weighted it for gender. Statisticians also feel that this is incorrect because weighting must occur simultaneously, not in separate calculations. Nielsen does not believe the concerns about their sample are legitimate and feel that they have not erred in weighting the survey. However, due to the fact that most third parties have not endorsed Nielsen’s methods, the validity of their research remains to be established.
Using the Base module, out-of-range values can be selected using the SELECT IF command. These cases, with the identifying information (subject ID, record number, variable name, and variable value) can then be printed using the LIST or PRINT commands. The Print command will save active cases to an external file. If a formatted list is required, the SUMMARIZE command can be used.

SPSS Data Entry can facilitate data preparation. You can verify respondents have answered completely by setting rules. These rules can be used on existing datasets to validate and check the data, whether or not the questionnaire used to collect the data was constructed in Data Entry. Data Entry allows you to control and check the entry of data through three types of rules: validation, checking, and skip and fill rules.

While the missing values can be treated within the context of the Base module, SPSS Missing Values Analysis can assist in diagnosing missing values and replacing missing values with estimates.

TextSmart by SPSS can help in the coding and analysis of open-ended responses.
SPSS Windows: Creating Overall Evaluation

1. Select TRANSFORM.
2. Click on COMPUTE.
3. Type “overall” in the TARGET VARIABLE box.
4. Click on “quality” and move it to the NUMERIC EXPRESSIONS box.
5. Click on the “+” sign.
6. Click on “quantity” and move it to the NUMERIC EXPRESSIONS box.
7. Click on the “+” sign.
Creating Overall Evaluation

8. Click on “value” and move it to the NUMERIC EXPRESSIONS box.

9. Click on the “+” sign.

10. Click on “service” and move it to the NUMERIC EXPRESSIONS box.

11. Click on TYPE & LABEL under the TARGET VARIABLE box and type “Overall Evaluation.” Click on CONTINUE.

12. Click OK.
SPSS Windows: Recoding Income

1. Select TRANSFORM.

2. Click on RECODE and select INTO DIFFERENT VARIABLES...

3. Click on income and move it to NUMERIC VARIABLE OUTPUT VARIABLE box.

4. Type “rincome” in OUTPUT VARIABLE NAME box.

5. Type “Recode Income” in OUTPUT VARIABLE LABEL box.

6. Click OLD AND NEW VAULES box.

7. Under OLD VALUES on the left, click RANGE. Type 1 and 2 in the range boxes. Under NEW VALUES on the right, click VALUE and type 1 in the value box. Click ADD.
Recoding Income

8. Under OLD VALUES on the left, click VALUE. Type 3 in the value box. Under NEW VALUES on the right, click VALUE and type 2 in the value box. Click ADD.

9. Under OLD VALUES on the left, click VALUE. Type 4 in the value box. Under NEW VALUES on the right, click VALUE and type 3 in the value box. Click ADD.

10. Under OLD VALUES on the left, click RANGE. Type 5 and 6 in the range boxes. Under NEW VALUES on the right, click VALUE and type 4 in the value box. Click ADD.

11. Click CONTINUE.

12. Click CHANGE.

13. Click OK.
1. Select DATA.
2. Click on FILTER AND QUERY.
3. Select the COMPUTED COLUMNS button.
4. Click on NEW.
5. Select BUILD EXPRESSION.
6. Select “QUALITY” and click on ADD TO EXPRESSION.
7. Click on the “+” sign.
8. Select “QUANTITY” and click on ADD TO EXPRESSION.
9. Click on the “+” sign.
10. Select “VALUE” and click on ADD TO EXPRESSION.
11. Click on the “+” sign.
12. Select “SERVICE” and click on ADD TO EXPRESSION.
13. Click OK.
14. Select “CALCULATION1” and click on RENAME.
15. Type OVERALL.
16. Click on CLOSE.
17. Select RUN.
Figure 14.8 A Concept Map for the Data Preparation Process

Preliminary Plan of Data Analysis

Data Preparation Process

involves

Questionnaire Checking

leads to

check for

Completeness

correct

illegible Responses

correct

incomplete Responses

Interviewing Quality

Inconsistent Responses

Ambiguous Responses

is followed by

Editing

Correct

Coding

develop

A Code Book

codes for

Structured Questions

Transcribing

thorough and extensive

transferring

Coded Data into Computers

Data Cleaning

involves

Consistency Checks

may require

Treatment of Missing Responses

Variable Respecification

lead to

New variables

create

Existing Variables

leads to

Selecting a Data Analysis Strategy

using

Univariate techniques

using

Multivariate Techniques
In a **frequency distribution**, one variable is considered at a time.

A frequency distribution for a variable produces a table of frequency counts, percentages, and cumulative percentages for all the values associated with that variable.
# Frequency of Familiarity with the Internet

## Table 15.2

<table>
<thead>
<tr>
<th>Value label</th>
<th>Value</th>
<th>Frequency (n)</th>
<th>Percentage</th>
<th>Valid Percentage</th>
<th>Cumulative Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Not so familiar</td>
<td>1</td>
<td>0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>2</td>
<td>6.7</td>
<td>6.9</td>
<td>6.9</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>6</td>
<td>20.0</td>
<td>20.7</td>
<td>27.6</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>6</td>
<td>20.0</td>
<td>20.7</td>
<td>48.3</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>3</td>
<td>10.0</td>
<td>10.3</td>
<td>58.6</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>8</td>
<td>26.7</td>
<td>27.6</td>
<td>86.2</td>
</tr>
<tr>
<td>Very familiar</td>
<td>7</td>
<td>4</td>
<td>13.3</td>
<td>13.8</td>
<td>100.0</td>
</tr>
<tr>
<td>Missing</td>
<td>9</td>
<td>1</td>
<td>3.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td><strong>30</strong></td>
<td></td>
<td><strong>100.0</strong></td>
<td></td>
<td><strong>100.0</strong></td>
</tr>
</tbody>
</table>
Frequency Histogram

Fig. 15.1
Statistics Associated with Frequency Distribution: Measures of Location

• The **mean**, or average value, is the most commonly used measure of central tendency. The mean, \( \bar{X} \), is given by
\[
\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}
\]

Where,
- \( X_i \) = Observed values of the variable \( X \)
- \( n \) = Number of observations (sample size)

• The **mode** is the value that occurs most frequently. It represents the highest peak of the distribution. The mode is a good measure of location when the variable is inherently categorical or has otherwise been grouped into categories.
The median of a sample is the middle value when the data are arranged in ascending or descending order. If the number of data points is even, the median is usually estimated as the midpoint between the two middle values – by adding the two middle values and dividing their sum by 2. The median is the 50th percentile.
Statistics Associated with Frequency Distribution: Measures of Variability

- The **range** measures the spread of the data. It is simply the difference between the largest and smallest values in the sample.

\[
\text{Range} = X_{\text{largest}} - X_{\text{smallest}}
\]

- The **interquartile range** is the difference between the 75th and 25th percentile. For a set of data points arranged in order of magnitude, the \( p^{\text{th}} \) percentile is the value that has \( p\% \) of the data points below it and \( (100 - p)\% \) above it.
Statistics Associated with Frequency Distribution: Measures of Variability

- The **variance** is the mean squared deviation from the mean. The variance can never be negative.
- The **standard deviation** is the square root of the variance.

\[
s_x = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n - 1}}
\]

- The **coefficient of variation** is the ratio of the standard deviation to the mean expressed as a percentage, and is a unitless measure of relative variability.

\[
CV = \frac{s_x}{\bar{X}}
\]
Skewness. The tendency of the deviations from the mean to be larger in one direction than in the other. It can be thought of as the tendency for one tail of the distribution to be heavier than the other.

Kurtosis is a measure of the relative peakedness or flatness of the curve defined by the frequency distribution. The kurtosis of a normal distribution is zero. If the kurtosis is positive, then the distribution is more peaked than a normal distribution. A negative value means that the distribution is flatter than a normal distribution.
Skewness of a Distribution

Fig. 15.2

Symmetric Distribution

Skewed Distribution

Mean
Median
Mode
(a)

Mean
Median
Mode
(b)
Steps Involved in Hypothesis Testing

Fig. 15.3

1. Formulate \( H_0 \) and \( H_1 \)
2. Select Appropriate Test
3. Choose Level of Significance
4. Collect Data and Calculate Test Statistic
5. Determine Probability Associated with Test Statistic
6. Compare with Level of Significance, \( \alpha \)
7. Determine Critical Value of Test Statistic \( T_{SRC} \)
8. Determine if \( T_{SCAL} \) falls into (Non) Rejection Region
9. Reject or Do not Reject \( H_0 \)
10. Draw Marketing Research Conclusion
A General Procedure for Hypothesis Testing
Step 1: Formulate the Hypothesis

• A **null hypothesis** is a statement of the status quo, one of no difference or no effect. If the null hypothesis is not rejected, no changes will be made.

• An **alternative hypothesis** is one in which some difference or effect is expected. Accepting the alternative hypothesis will lead to changes in opinions or actions.

• The null hypothesis refers to a specified value of the population parameter (e.g., $\mu, \sigma, \pi$), not a sample statistic (e.g., $\bar{X}$).
A General Procedure for Hypothesis Testing
Step 1: Formulate the Hypothesis

• A null hypothesis may be rejected, but it can never be accepted based on a single test. In classical hypothesis testing, there is no way to determine whether the null hypothesis is true.

• In marketing research, the null hypothesis is formulated in such a way that its rejection leads to the acceptance of the desired conclusion. The alternative hypothesis represents the conclusion for which evidence is sought.

\[ H_0: \pi \leq 0.40 \]

\[ H_1: \pi > 0.40 \]
• The test of the null hypothesis is a **one-tailed test**, because the alternative hypothesis is expressed directionally. If that is not the case, then a **two-tailed test** would be required, and the hypotheses would be expressed as:

\[
H_0: \pi = 0.40 \\
H_1: \pi \neq 0.40
\]
A General Procedure for Hypothesis Testing
Step 2: Select an Appropriate Test

- The **test statistic** measures how close the sample has come to the null hypothesis.
- The test statistic often follows a well-known distribution, such as the normal, *t*, or chi-square distribution.
- In our example, the **z statistic**, which follows the standard normal distribution, would be appropriate.

\[
z = \frac{p - \pi}{\sigma_p}
\]

where

\[
\sigma_p = \sqrt{\pi (1 - \pi) \frac{1}{n}}
\]
**A General Procedure for Hypothesis Testing**

**Step 3: Choose a Level of Significance**

**Type I Error**
- **Type I error** occurs when the sample results lead to the rejection of the null hypothesis when it is in fact true.
- The probability of type I error ($\alpha$) is also called the **level of significance**.

**Type II Error**
- **Type II error** occurs when, based on the sample results, the null hypothesis is not rejected when it is in fact false.
- The probability of type II error is denoted by $\beta$.
- Unlike $\alpha$, which is specified by the researcher, the magnitude of $\beta$ depends on the actual value of the population parameter (proportion).
Power of a Test

- The **power of a test** is the probability \((1 - \beta)\) of rejecting the null hypothesis when it is false and should be rejected.
- Although \(\beta\) is unknown, it is related to \(\alpha\). An extremely low value of \(\alpha\) (e.g., \(\alpha = 0.001\)) will result in intolerably high \(\beta\) errors.
- So it is necessary to balance the two types of errors.
Probabilities of Type I & Type II Error

Fig. 15.4

95% of Total Area

\[ \Pi_0 = 0.40 \]
\[ Z \alpha = 1.645 \]

Critical Value of \( z \)

\[ \beta = 0.01 \]

99% of Total Area

\[ \Pi = 0.45 \]

\[ Z \beta = -2.33 \]
Probability of $z$ with a One-Tailed Test

Fig. 15.5

Shaded Area
$= 0.9699$

Unshaded Area
$= 0.0301$

$z = 1.88$
A General Procedure for Hypothesis Testing
Step 4: Collect Data and Calculate Test Statistic

- The required data are collected and the value of the test statistic computed.
- In our example, the value of the sample proportion is \( p = \frac{17}{30} = 0.567 \).
- The value of \( \sigma_p \) can be determined as follows:

\[
\sigma_p = \sqrt{\frac{\pi(1 - \pi)}{n}} = \sqrt{\frac{(0.40)(0.6)}{30}} = 0.089
\]
The test statistic $z$ can be calculated as follows:

$$z = \frac{\hat{p} - \pi}{\sigma_p}$$

$$= \frac{0.567 - 0.40}{0.089}$$

$$= 1.88$$
A General Procedure for Hypothesis Testing
Step 5: Determine the Probability (Critical Value)

• Using standard normal tables (Table 2 of the Statistical Appendix), the probability of obtaining a $z$ value of 1.88 can be calculated (see Figure 15.5).

• The shaded area between $-\infty$ and 1.88 is 0.9699. Therefore, the area to the right of $z = 1.88$ is $1.0000 - 0.9699 = 0.0301$.

• Alternatively, the critical value of $z$, which will give an area to the right side of the critical value of 0.05, is between 1.64 and 1.65 and equals 1.645.

• Note, in determining the critical value of the test statistic, the area to the right of the critical value is either $\alpha$ or $\alpha/2$. It is $\alpha$ for a one-tail test and $\alpha/2$ for a two-tail test.
A General Procedure for Hypothesis Testing Steps 6 & 7: Compare the Probability (Critical Value) & Making the Decision

- If the probability associated with the calculated or observed value of the test statistic (TSCAL) is less than the level of significance (\( \alpha \)), the null hypothesis is rejected.

- The probability associated with the calculated or observed value of the test statistic is 0.0301. This is the probability of getting a \( p \) value of 0.567 when \( \Pi = 0.40 \). This is less than the level of significance of 0.05. Hence, the null hypothesis is rejected.

- Alternatively, if the absolute calculated value of the test statistic (\( |TS_{CAL}| \)) is greater than the absolute critical value of the test statistic (\( |TS_{CR}| \)), the null hypothesis is rejected.
A General Procedure for Hypothesis Testing Steps 6 & 7: Compare the Probability (Critical Value) & Making the Decision

- The calculated value of the test statistic $z = 1.88$ lies in the rejection region, beyond the value of 1.645. Again, the same conclusion to reject the null hypothesis is reached.

- Note that the two ways of testing the null hypothesis are equivalent but mathematically opposite in the direction of comparison.

- If the probability of $TS_{\text{CAL}} < \text{significance level (} \alpha \text{)}$ then reject $H_0$ but if $|TS_{\text{CAL}}| > |TS_{\text{CR}}|$ then reject $H_0$. 
A General Procedure for Hypothesis Testing
Step 8: Marketing Research Conclusion

• The conclusion reached by hypothesis testing must be expressed in terms of the marketing research problem.

• In our example, we conclude that there is evidence that the proportion of Internet users who shop via the Internet is significantly greater than 0.40. Hence, the recommendation to the department store would be to introduce the new Internet shopping service.
A Broad Classification of Hypothesis Tests

Tests of Association

Tests of Differences

Distributions

Means

Proportions

Median/Rankings
Cross-Tabulation

- While a frequency distribution describes one variable at a time, a cross-tabulation describes two or more variables simultaneously.

- Cross-tabulation results in tables that reflect the joint distribution of two or more variables with a limited number of categories or distinct values, e.g., Table 15.3.
## Gender and Internet Usage

Table 15.3

<table>
<thead>
<tr>
<th>Gender</th>
<th>Internet Usage</th>
<th>Male</th>
<th>Female</th>
<th>Row Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Light (1)</td>
<td>5</td>
<td>10</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>Heavy (2)</td>
<td>10</td>
<td>5</td>
<td>15</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td>15</td>
<td>15</td>
<td><strong>15</strong></td>
</tr>
</tbody>
</table>
Two Variables Cross-Tabulation

• Since two variables have been cross-classified, percentages could be computed either columnwise, based on column totals (Table 15.4), or rowwise, based on row totals (Table 15.5).

• The general rule is to compute the percentages in the direction of the independent variable, across the dependent variable. The correct way of calculating percentages is as shown in Table 15.4.
## Internet Usage by Gender

Table 15.4

<table>
<thead>
<tr>
<th>Internet Usage</th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>Light</td>
<td>33.3%</td>
<td>66.7%</td>
</tr>
<tr>
<td>Heavy</td>
<td>66.7%</td>
<td>33.3%</td>
</tr>
<tr>
<td>Column total</td>
<td>100%</td>
<td>100%</td>
</tr>
</tbody>
</table>
## Gender by Internet Usage

Table 15.5

<table>
<thead>
<tr>
<th>Gender</th>
<th>Light</th>
<th>Heavy</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>33.3%</td>
<td>66.7%</td>
<td>100.0%</td>
</tr>
<tr>
<td>Female</td>
<td>66.7%</td>
<td>33.3%</td>
<td>100.0%</td>
</tr>
</tbody>
</table>
Introduction of a Third Variable in Cross-Tabulation

Fig. 15.7

Original Two Variables

Some Association between the Two Variables

Introduce a Third Variable

Refined Association between the Two Variables

No Association between the Two Variables

Introduce a Third Variable

No Change in the Initial Pattern

Some Association between the Two Variables
Three Variables Cross-Tabulation Refine an Initial Relationship

As shown in Figure 15.7, the introduction of a third variable can result in four possibilities:

- As can be seen from Table 15.6, 52% of unmarried respondents fell in the high-purchase category, as opposed to 31% of the married respondents. Before concluding that unmarried respondents purchase more fashion clothing than those who are married, a third variable, the buyer's sex, was introduced into the analysis.

- As shown in Table 15.7, in the case of females, 60% of the unmarried fall in the high-purchase category, as compared to 25% of those who are married. On the other hand, the percentages are much closer for males, with 40% of the unmarried and 35% of the married falling in the high purchase category.

- Hence, the introduction of sex (third variable) has refined the relationship between marital status and purchase of fashion clothing (original variables). Unmarried respondents are more likely to fall in the high purchase category than married ones, and this effect is much more pronounced for females than for males.
## Purchase of Fashion Clothing by Marital Status

Table 15.6

<table>
<thead>
<tr>
<th>Purchase of Fashion Clothing</th>
<th>Current Marital Status</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Married</td>
<td>Unmarried</td>
</tr>
<tr>
<td>High</td>
<td>31%</td>
<td>52%</td>
</tr>
<tr>
<td>Low</td>
<td>69%</td>
<td>48%</td>
</tr>
<tr>
<td>Column</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>Number of respondents</td>
<td>700</td>
<td>300</td>
</tr>
</tbody>
</table>
### Table 15.7

Purchase of Fashion Clothing by Marital Status

<table>
<thead>
<tr>
<th>Purchase of Fashion Clothing</th>
<th>Sex</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Male</td>
<td>Not Married</td>
<td>Married</td>
</tr>
<tr>
<td>High</td>
<td>35%</td>
<td>40%</td>
<td>25%</td>
</tr>
<tr>
<td>Low</td>
<td>65%</td>
<td>60%</td>
<td>75%</td>
</tr>
<tr>
<td>Column totals</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>Number of cases</td>
<td>400</td>
<td>120</td>
<td>300</td>
</tr>
</tbody>
</table>
• Table 15.8 shows that 32% of those with college degrees own an expensive automobile, as compared to 21% of those without college degrees. Realizing that income may also be a factor, the researcher decided to reexamine the relationship between education and ownership of expensive automobiles in light of income level.

• In Table 15.9, the percentages of those with and without college degrees who own expensive automobiles are the same for each of the income groups. When the data for the high income and low income groups are examined separately, the association between education and ownership of expensive automobiles disappears, indicating that the initial relationship observed between these two variables was spurious.
Ownership of Expensive Automobiles by Education Level

Table 15.8

<table>
<thead>
<tr>
<th>Own Expensive Automobile</th>
<th>Education</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>College Degree</td>
</tr>
<tr>
<td>Yes</td>
<td>32%</td>
</tr>
<tr>
<td>No</td>
<td>68%</td>
</tr>
<tr>
<td>Column totals</td>
<td>100%</td>
</tr>
<tr>
<td>Number of cases</td>
<td>250</td>
</tr>
</tbody>
</table>
Ownership of Expensive Automobiles by Education Level and Income Levels

Table 15.9

<table>
<thead>
<tr>
<th>Own Expensive Automobile</th>
<th>Low Income</th>
<th>High Income</th>
</tr>
</thead>
<tbody>
<tr>
<td>College Degree</td>
<td>College Degree</td>
<td>No College Degree</td>
</tr>
<tr>
<td>Yes</td>
<td>20%</td>
<td>20%</td>
</tr>
<tr>
<td>No</td>
<td>80%</td>
<td>80%</td>
</tr>
<tr>
<td>Column totals</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>Number of respondents</td>
<td>100</td>
<td>700</td>
</tr>
</tbody>
</table>
Three Variables Cross-Tabulation Reveal Suppressed Association

- Table 15.10 shows no association between desire to travel abroad and age.

- When sex was introduced as the third variable, Table 15.11 was obtained. Among men, 60% of those under 45 indicated a desire to travel abroad, as compared to 40% of those 45 or older. The pattern was reversed for women, where 35% of those under 45 indicated a desire to travel abroad as opposed to 65% of those 45 or older.

- Since the association between desire to travel abroad and age runs in the opposite direction for males and females, the relationship between these two variables is masked when the data are aggregated across sex, as in Table 15.10.

- But when the effect of sex is controlled, as in Table 15.11, the suppressed association between desire to travel abroad and age is revealed for the separate categories of males and females.
## Desire to Travel Abroad by Age

Table 15.10

<table>
<thead>
<tr>
<th>Desire to Travel Abroad</th>
<th>Age</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Less than 45</td>
</tr>
<tr>
<td>Yes</td>
<td>50%</td>
</tr>
<tr>
<td>No</td>
<td>50%</td>
</tr>
<tr>
<td>Column totals</td>
<td>100%</td>
</tr>
<tr>
<td>Number of respondents</td>
<td>500</td>
</tr>
</tbody>
</table>
# Desire to Travel Abroad by Age and Gender

## Table 15.11

<table>
<thead>
<tr>
<th>Desire to Travel Abroad</th>
<th>Male</th>
<th></th>
<th>Female</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt; 45</td>
<td>60%</td>
<td>40%</td>
<td>35%</td>
<td>65%</td>
</tr>
<tr>
<td>&gt;=45</td>
<td>40%</td>
<td>60%</td>
<td>65%</td>
<td>35%</td>
</tr>
<tr>
<td><strong>Column totals</strong></td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td><strong>Number of Cases</strong></td>
<td>300</td>
<td>300</td>
<td>200</td>
<td>200</td>
</tr>
</tbody>
</table>
Consider the cross-tabulation of family size and the tendency to eat out frequently in fast-food restaurants, as shown in Table 15.12. No association is observed.

When income was introduced as a third variable in the analysis, Table 15.13 was obtained. Again, no association was observed.
## Eating Frequently in Fast-Food Restaurants by Family Size

### Table 15.12

<table>
<thead>
<tr>
<th>Eat Frequently in Fast-Food Restaurants</th>
<th>Family Size</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Small</td>
</tr>
<tr>
<td>Yes</td>
<td>65%</td>
</tr>
<tr>
<td>No</td>
<td>35%</td>
</tr>
<tr>
<td>Column totals</td>
<td>100%</td>
</tr>
<tr>
<td>Number of cases</td>
<td>500</td>
</tr>
</tbody>
</table>
## Eating Frequently in Fast-Food Restaurants by Family Size and Income

### Table 15.13

<table>
<thead>
<tr>
<th>Eat Frequently in Fast-Food Restaurants</th>
<th>Income</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Low</td>
</tr>
<tr>
<td></td>
<td>Family size</td>
</tr>
<tr>
<td></td>
<td>Small</td>
</tr>
<tr>
<td>Yes</td>
<td>65%</td>
</tr>
<tr>
<td>No</td>
<td>35%</td>
</tr>
<tr>
<td>Column totals</td>
<td>100%</td>
</tr>
<tr>
<td>Number of respondents</td>
<td>250</td>
</tr>
</tbody>
</table>
Statistics Associated with Cross-Tabulation
Chi-Square

- To determine whether a systematic association exists, the probability of obtaining a value of chi-square as large or larger than the one calculated from the cross-tabulation is estimated.

- An important characteristic of the chi-square statistic is the number of degrees of freedom (df) associated with it. That is, df = \((r - 1) \times (c -1)\).

- The null hypothesis \((H_0)\) of no association between the two variables will be rejected only when the calculated value of the test statistic is greater than the critical value of the chi-square distribution with the appropriate degrees of freedom, as shown in Figure 15.8.
Chi-square Distribution

Fig. 15.8

Do Not Reject $H_0$

Reject $H_0$

Critical Value

$\chi^2$
Statistics Associated with Cross-Tabulation Chi-Square

• The **chi-square statistic** ($\chi^2$) is used to test the statistical significance of the observed association in a cross-tabulation.

• The expected frequency for each cell can be calculated by using a simple formula:

$$f_e = \frac{n_r n_c}{n}$$

where

- $n_r$ = total number in the row
- $n_c$ = total number in the column
- $n$ = total sample size
For the data in Table 15.3, the expected frequencies for the cells going from left to right and from top to bottom, are:

\[
\frac{15 \times 15}{30} = 7.50 \quad \frac{15 \times 15}{30} = 7.50
\]

\[
\frac{15 \times 15}{30} = 7.50 \quad \frac{15 \times 15}{30} = 7.50
\]

Then the value of \( \chi^2 \) is calculated as follows:

\[
\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}
\]
For the data in Table 15.3, the value of $\chi^2$ is calculated as:

\[
\begin{align*}
\chi^2 &= \frac{(5 - 7.5)^2}{7.5} + \frac{(10 - 7.5)^2}{7.5} + \frac{(10 - 7.5)^2}{7.5} + \frac{(5 - 7.5)^2}{7.5} \\
&= 0.833 + 0.833 + 0.833 + 0.833 \\
&= 3.333
\end{align*}
\]
The chi-square distribution is a skewed distribution whose shape depends solely on the number of degrees of freedom. As the number of degrees of freedom increases, the chi-square distribution becomes more symmetrical.

Table 3 in the Statistical Appendix contains upper-tail areas of the chi-square distribution for different degrees of freedom. For 1 degree of freedom, the probability of exceeding a chi-square value of 3.841 is 0.05.

For the cross-tabulation given in Table 15.3, there are (2-1) x (2-1) = 1 degree of freedom. The calculated chi-square statistic had a value of 3.333. Since this is less than the critical value of 3.841, the null hypothesis of no association can not be rejected indicating that the association is not statistically significant at the 0.05 level.
Statistics Associated with Cross-Tabulation Phi Coefficient

- The **phi coefficient** ($\phi$) is used as a measure of the strength of association in the special case of a table with two rows and two columns (a 2 x 2 table).
- The phi coefficient is proportional to the square root of the chi-square statistic
  \[ \phi = \sqrt{\frac{\chi^2}{n}} \]
- It takes the value of 0 when there is no association, which would be indicated by a chi-square value of 0 as well. When the variables are perfectly associated, phi assumes the value of 1 and all the observations fall just on the main or minor diagonal.
Statistics Associated with Cross-Tabulation

Contingency Coefficient

- While the phi coefficient is specific to a 2 x 2 table, the **contingency coefficient** ($C$) can be used to assess the strength of association in a table of any size.

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}}$$

- The contingency coefficient varies between 0 and 1.
- The maximum value of the contingency coefficient depends on the size of the table (number of rows and number of columns). For this reason, it should be used only to compare tables of the same size.
Cramer's $V$ is a modified version of the phi correlation coefficient, $\phi$, and is used in tables larger than 2 x 2.

\[
V = \sqrt{\frac{\phi^2}{\min (r-1), (c-1)}}
\]

or

\[
V = \sqrt{\frac{\chi^2/n}{\min (r-1), (c-1)}}
\]
Statistics Associated with Cross-Tabulation
Lambda Coefficient

• **Asymmetric lambda** measures the percentage improvement in predicting the value of the dependent variable, given the value of the independent variable.

• Lambda also varies between 0 and 1. A value of 0 means no improvement in prediction. A value of 1 indicates that the prediction can be made without error. This happens when each independent variable category is associated with a single category of the dependent variable.

• Asymmetric lambda is computed for each of the variables (treating it as the dependent variable).

• A **symmetric lambda** is also computed, which is a kind of average of the two asymmetric values. The symmetric lambda does not make an assumption about which variable is dependent. It measures the overall improvement when prediction is done in both directions.
Other statistics like tau b, tau c, and gamma are available to measure association between two ordinal-level variables. Both tau b and tau c adjust for ties.

- **Tau b** is the most appropriate with square tables in which the number of rows and the number of columns are equal. Its value varies between +1 and -1.

- For a rectangular table in which the number of rows is different than the number of columns, **tau c** should be used.

- **Gamma** does not make an adjustment for either ties or table size. Gamma also varies between +1 and -1 and generally has a higher numerical value than tau b or tau c.
Cross-Tabulation in Practice

While conducting cross-tabulation analysis in practice, it is useful to proceed along the following steps.

1. Test the null hypothesis that there is no association between the variables using the chi-square statistic. If you fail to reject the null hypothesis, then there is no relationship.

2. If \( H_0 \) is rejected, then determine the strength of the association using an appropriate statistic (phi-coefficient, contingency coefficient, Cramer's \( V \), lambda coefficient, or other statistics), as discussed earlier.

3. If \( H_0 \) is rejected, interpret the pattern of the relationship by computing the percentages in the direction of the independent variable, across the dependent variable.

4. If the variables are treated as ordinal rather than nominal, use tau \( b \), tau \( c \), or Gamma as the test statistic. If \( H_0 \) is rejected, then determine the strength of the association using the magnitude, and the direction of the relationship using the sign of the test statistic.
Hypothesis Testing Related to Differences

- **Parametric tests** assume that the variables of interest are measured on at least an interval scale.

- **Nonparametric tests** assume that the variables are measured on a nominal or ordinal scale.

- These tests can be further classified based on whether one or two or more samples are involved.

- The samples are **independent** if they are drawn randomly from different populations. For the purpose of analysis, data pertaining to different groups of respondents, e.g., males and females, are generally treated as independent samples.

- The samples are **paired** when the data for the two samples relate to the same group of respondents.
A Classification of Hypothesis Testing Procedures for Examining Differences

Fig. 15.9

Hypothesis Tests

Parametric Tests (Metric Tests)

One Sample
* t test
* Z test

Two or More Samples

Independent Samples
* Two-Group t test
* Z test

Paired Samples
* Paired t test

Non-parametric Tests (Nonmetric Tests)

One Sample
* Chi-Square
* K-S
* Runs
* Binomial

Two or More Samples

Independent Samples
* Chi-Square
* Mann-Whitney
* Median
* K-S

Paired Samples
* Sign
* Wilcoxon
* McNemar
* Chi-Square
The *t statistic* assumes that the variable is normally distributed and the mean is known (or assumed to be known) and the population variance is estimated from the sample.

Assume that the random variable $X$ is normally distributed, with mean and unknown population variance that is estimated by the sample variance $s^2$.

Then, $t = \frac{\bar{X} - \mu}{s_{\bar{X}}}$ is $t$ distributed with $n - 1$ degrees of freedom.

The *t distribution* is similar to the normal distribution in appearance. Both distributions are bell-shaped and symmetric. As the number of degrees of freedom increases, the $t$ distribution approaches the normal distribution.
Hypothesis Testing Using the t Statistic

1. Formulate the null ($H_0$) and the alternative ($H_1$) hypotheses.

2. Select the appropriate formula for the $t$ statistic.

3. Select a significance level, $\alpha$, for testing $H_0$. Typically, the 0.05 level is selected.

4. Take one or two samples and compute the mean and standard deviation for each sample.

5. Calculate the $t$ statistic assuming $H_0$ is true.
6. Calculate the degrees of freedom and estimate the probability of getting a more extreme value of the statistic from Table 4. (Alternatively, calculate the critical value of the t statistic.)

7. If the probability computed in step 5 is smaller than the significance level selected in step 2, reject $H_0$. If the probability is larger, do not reject $H_0$. (Alternatively, if the absolute value of the calculated $t$ statistic in step 4 is larger than the absolute critical value determined in step 5, reject $H_0$.) Failure to reject $H_0$ does not necessarily imply that $H_0$ is true. It only means that the true state is not significantly different than that assumed by $H_0$.

8. Express the conclusion reached by the $t$ test in terms of the marketing research problem.
One Sample : t Test

For the data in Table 15.2, suppose we wanted to test the hypothesis that the mean familiarity rating exceeds 4.0, the neutral value on a 7-point scale. A significance level of $\alpha = 0.05$ is selected. The hypotheses may be formulated as:

$H_0$: $\mu \leq 4.0$

$H_1$: $\mu > 4.0$

$t = \frac{(\bar{X} - \mu)}{s_{\bar{X}}}$

$s_{\bar{X}} = \frac{s}{\sqrt{n}}$

$s_{\bar{X}} = \frac{1.579}{\sqrt{29}}$

$= \frac{1.579}{5.385} = 0.293$

$t = \frac{(4.724-4.0)}{0.293} = \frac{0.724}{0.293} = 2.471$
The degrees of freedom for the $t$ statistic to test the hypothesis about one mean are $n - 1$. In this case, $n - 1 = 29 - 1$ or 28. From Table 4 in the Statistical Appendix, the probability of getting a more extreme value than 2.471 is less than 0.05. (Alternatively, the critical $t$ value for 28 degrees of freedom and a significance level of 0.05 is 1.7011, which is less than the calculated value.) Hence, the null hypothesis is rejected. The familiarity level does exceed 4.0.
One Sample : Z Test

Note that if the population standard deviation was
assumed to be known as 1.5, rather than estimated
from the sample, a z test would be appropriate. In
this case, the value of the z statistic would be:

\[
z = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} \]

where

\[
\sigma_{\bar{X}} = \frac{1.5}{\sqrt{29}} = \frac{1.5}{5.385} = 0.279
\]

and

\[
z = \frac{(4.724 - 4.0)}{0.279} = 0.724/0.279 = 2.595
\]
One Sample : Z Test

- From Table 2 in the Statistical Appendix, the probability of getting a more extreme value of $z$ than 2.595 is less than 0.05. (Alternatively, the critical $z$ value for a one-tailed test and a significance level of 0.05 is 1.645, which is less than the calculated value.) Therefore, the null hypothesis is rejected, reaching the same conclusion arrived at earlier by the $t$ test.

- The procedure for testing a null hypothesis with respect to a proportion was illustrated earlier in this chapter when we introduced hypothesis testing.
Two Independent Samples Means

• In the case of means for two independent samples, the hypotheses take the following form.

\[ H_0 : \mu_1 = \mu_2 \]
\[ H_1 : \mu_1 \neq \mu_2 \]

• The two populations are sampled and the means and variances computed based on samples of sizes \( n_1 \) and \( n_2 \). If both populations are found to have the same variance, a pooled variance estimate is computed from the two sample variances as follows:

\[ s^2 = \frac{\sum_{i=1}^{n_1} (X_{i1} - \bar{X}_1)^2 + \sum_{i=1}^{n_2} (X_{i2} - \bar{X}_2)^2}{n_1 + n_2 - 2} \]

or

\[ s^2 = \frac{(n_1 - 1) s_1^2 + (n_2 - 1) s_2^2}{n_1 + n_2 - 2} \]
Two Independent Samples Means

The standard deviation of the test statistic can be estimated as:

\[ s_{\bar{X}_1 - \bar{X}_2} = \sqrt{s^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)} \]

The appropriate value of \( t \) can be calculated as:

\[ t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{s_{\bar{X}_1 - \bar{X}_2}} \]

The degrees of freedom in this case are \((n_1 + n_2 - 2)\).
An $F$ test of sample variance may be performed if it is not known whether the two populations have equal variance. In this case, the hypotheses are:

$H_0: \sigma_1^2 = \sigma_2^2$

$H_1: \sigma_1^2 \neq \sigma_2^2$
Two Independent Samples F Statistic

The *F statistic* is computed from the sample variances as follows:

\[
F_{(n_1-1),(n_2-1)} = \frac{s_1^2}{s_2^2}
\]

where

- \(n_1\) = size of sample 1
- \(n_2\) = size of sample 2
- \(n_1-1\) = degrees of freedom for sample 1
- \(n_2-1\) = degrees of freedom for sample 2
- \(s_1^2\) = sample variance for sample 1
- \(s_2^2\) = sample variance for sample 2

Using the data of Table 15.1, suppose we wanted to determine whether Internet usage was different for males as compared to females. A two-independent-samples *t* test was conducted. The results are presented in Table 15.14.
# Two Independent-Samples $t$ Tests

## Table 15.14

<table>
<thead>
<tr>
<th></th>
<th>Number of Cases</th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>15</td>
<td>9.333</td>
<td>1.137</td>
</tr>
<tr>
<td>Female</td>
<td>15</td>
<td>3.867</td>
<td>0.435</td>
</tr>
</tbody>
</table>

### $F$ Test for Equality of Variances

<table>
<thead>
<tr>
<th>$F$ value</th>
<th>2-tail probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>15.507</td>
<td>0.000</td>
</tr>
</tbody>
</table>

### $t$ Test

<table>
<thead>
<tr>
<th></th>
<th>Equal Variances Assumed</th>
<th>Equal Variances Not Assumed</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t$ value</td>
<td>Degrees of freedom</td>
<td>2-tail probability</td>
</tr>
<tr>
<td>-4.492</td>
<td>28</td>
<td>0.000</td>
</tr>
</tbody>
</table>
Two Independent Samples Proportions

The case involving proportions for two independent samples is also illustrated using the data of Table 15.1, which gives the number of males and females who use the Internet for shopping. Is the proportion of respondents using the Internet for shopping the same for males and females? The null and alternative hypotheses are:

\[ H_0 : \pi_1 = \pi_2 \]
\[ H_1 : \pi_1 \neq \pi_2 \]

A Z test is used as in testing the proportion for one sample. However, in this case the test statistic is given by:

\[ Z = \frac{P_1 - P_2}{S_{\sqrt{p_1-p_2}}} \]
Two Independent Samples Proportions

In the test statistic, the numerator is the difference between the proportions in the two samples, $P_1$ and $P_2$. The denominator is the standard error of the difference in the two proportions and is given by

\[ S_{p_1-p_2} = \sqrt{P(1-P) \left( \frac{1}{n_1} + \frac{1}{n_2} \right)} \]

where

\[ P = \frac{n_1 P_1 + n_2 P_2}{n_1 + n_2} \]
Two Independent Samples Proportions

A significance level of $\alpha = 0.05$ is selected. Given the data of Table 15.1, the test statistic can be calculated as:

$$P_1 - P_2 = (11/15) - (6/15)$$

$$= 0.733 - 0.400 = 0.333$$

$$P = (15 \times 0.733 + 15 \times 0.4)/(15 + 15) = 0.567$$

$$S_{p_1-p_2} = \sqrt{0.567 \times 0.433 \left[ \frac{1}{15} + \frac{1}{15} \right]} = 0.181$$

$$Z = 0.333/0.181 = 1.84$$
Two Independent Samples Proportions

Given a two-tail test, the area to the right of the critical value is 0.025. Hence, the critical value of the test statistic is 1.96. Since the calculated value is less than the critical value, the null hypothesis cannot be rejected. Thus, the proportion of users (0.733 for males and 0.400 for females) is not significantly different for the two samples. Note that while the difference is substantial, it is not statistically significant due to the small sample sizes (15 in each group).
Paired Samples

The difference in these cases is examined by a **paired samples t test**. To compute $t$ for paired samples, the paired difference variable, denoted by $D$, is formed and its mean and variance calculated. Then the $t$ statistic is computed. The degrees of freedom are $n - 1$, where $n$ is the number of pairs. The relevant formulas are:

\[
H_0: \mu_D = 0
\]

\[
H_1: \mu_D \neq 0
\]

\[
t_{n-1} = \frac{\bar{D} - \mu_D}{s_D} \frac{s_D}{\sqrt{n}}
\]

continued...
Paired Samples

Where:

\[
\overline{D} = \frac{\sum_{i=1}^{n} D_i}{n}
\]

\[
s_D = \sqrt{\frac{\sum_{i=1}^{n} (D_i - \overline{D})^2}{n - 1}}
\]

\[
S_{\overline{D}} = \frac{S_D}{\sqrt{n}}
\]

In the Internet usage example (Table 15.1), a paired \( t \) test could be used to determine if the respondents differed in their attitude toward the Internet and attitude toward technology. The resulting output is shown in Table 15.15.
Paired-Samples *t* Test

Table 15.15

<table>
<thead>
<tr>
<th>Variable</th>
<th>Number of Cases</th>
<th>Mean</th>
<th>Standard Deviation</th>
<th>Standard Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Internet Attitude</td>
<td>30</td>
<td>5.167</td>
<td>1.234</td>
<td>0.225</td>
</tr>
<tr>
<td>Technology Attitude</td>
<td>30</td>
<td>4.100</td>
<td>1.398</td>
<td>0.255</td>
</tr>
</tbody>
</table>

Difference = Internet - Technology

<table>
<thead>
<tr>
<th>Difference</th>
<th>Mean</th>
<th>Standard deviation</th>
<th>Standard error</th>
<th>Correlation</th>
<th>2-tail prob.</th>
<th><em>t</em> value</th>
<th>Degrees of freedom</th>
<th>2-tail probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Difference</td>
<td>1.067</td>
<td>0.828</td>
<td>0.1511</td>
<td>0.809</td>
<td>0.000</td>
<td>7.059</td>
<td>29</td>
<td>0.000</td>
</tr>
</tbody>
</table>
Nonparametric Tests

**Nonparametric tests** are used when the independent variables are nonmetric. Like parametric tests, nonparametric tests are available for testing variables from one sample, two independent samples, or two related samples.
Nonparametric Tests One Sample

- Sometimes the researcher wants to test whether the observations for a particular variable could reasonably have come from a particular distribution, such as the normal, uniform, or Poisson distribution.

- The **Kolmogorov-Smirnov (K-S) one-sample test** is one such goodness-of-fit test. The K-S compares the cumulative distribution function for a variable with a specified distribution. $A_i$ denotes the cumulative relative frequency for each category of the theoretical (assumed) distribution, and $O_i$ the comparable value of the sample frequency. The K-S test is based on the maximum value of the absolute difference between $A_i$ and $O_i$. The test statistic is

$$K = \text{Max} \left| A_i - O_i \right|$$
Nonparametric Tests One Sample

• The decision to reject the null hypothesis is based on the value of $K$. The larger the $K$ is, the more confidence we have that $H_0$ is false. For $\alpha = 0.05$, the critical value of $K$ for large samples (over 35) is given by $1.36/\sqrt{n}$. Alternatively, $K$ can be transformed into a normally distributed $z$ statistic and its associated probability determined.

• In the context of the Internet usage example, suppose we wanted to test whether the distribution of Internet usage was normal. A K-S one-sample test is conducted, yielding the data shown in Table 15.16. Table 15.16 indicates that the probability of observing a $K$ value of 0.222, as determined by the normalized $z$ statistic, is 0.103. Since this is more than the significance level of 0.05, the null hypothesis cannot be rejected, leading to the same conclusion. Hence, the distribution of Internet usage does not deviate significantly from the normal distribution.
# K-S One-Sample Test for Normality of Internet Usage

Table 15.16

<table>
<thead>
<tr>
<th>Test Distribution - Normal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean:</td>
</tr>
<tr>
<td>Standard Deviation:</td>
</tr>
<tr>
<td>Cases:</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Most Extreme Differences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Absolute</td>
</tr>
<tr>
<td>0.222</td>
</tr>
</tbody>
</table>
Nonparametric Tests One Sample

- The **chi-square test** can also be performed on a single variable from one sample. In this context, the chi-square serves as a goodness-of-fit test.

- The **runs test** is a test of randomness for the dichotomous variables. This test is conducted by determining whether the order or sequence in which observations are obtained is random.

- The **binomial test** is also a goodness-of-fit test for dichotomous variables. It tests the goodness of fit of the observed number of observations in each category to the number expected under a specified binomial distribution.
Nonparametric Tests Two Independent Samples

• When the difference in the location of two populations is to be compared based on observations from two independent samples, and the variable is measured on an ordinal scale, the **Mann-Whitney U test** can be used.

• In the Mann-Whitney $U$ test, the two samples are combined and the cases are ranked in order of increasing size.

• The test statistic, $U$, is computed as the number of times a score from sample or group 1 precedes a score from group 2.

• If the samples are from the same population, the distribution of scores from the two groups in the rank list should be random. An extreme value of $U$ would indicate a nonrandom pattern, pointing to the inequality of the two groups.

• For samples of less than 30, the exact significance level for $U$ is computed. For larger samples, $U$ is transformed into a normally distributed $z$ statistic. This $z$ can be corrected for ties within ranks.
We examine again the difference in the Internet usage of males and females. This time, though, the Mann-Whitney U test is used. The results are given in Table 15.17.

One could also use the cross-tabulation procedure to conduct a chi-square test. In this case, we will have a 2 x 2 table. One variable will be used to denote the sample, and will assume the value 1 for sample 1 and the value of 2 for sample 2. The other variable will be the binary variable of interest.

The two-sample median test determines whether the two groups are drawn from populations with the same median. It is not as powerful as the Mann-Whitney U test because it merely uses the location of each observation relative to the median, and not the rank, of each observation.

The Kolmogorov-Smirnov two-sample test examines whether the two distributions are the same. It takes into account any differences between the two distributions, including the median, dispersion, and skewness.
### Mann-Whitney U - Wilcoxon Rank Sum W Test Internet Usage by Gender

**Table 15.17**

<table>
<thead>
<tr>
<th>Sex</th>
<th>Mean Rank</th>
<th>Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>20.93</td>
<td>15</td>
</tr>
<tr>
<td>Female</td>
<td>10.07</td>
<td>15</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>30</td>
</tr>
</tbody>
</table>

Corrected for ties

<table>
<thead>
<tr>
<th>U</th>
<th>W</th>
<th>z</th>
<th>2-tailed $p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>31.000</td>
<td>151.000</td>
<td>-3.406</td>
<td>0.001</td>
</tr>
</tbody>
</table>

**Note**

- $U$ = Mann-Whitney test statistic
- $W$ = Wilcoxon W Statistic
- $z$ = $U$ transformed into normally distributed $z$ statistic
The Wilcoxon matched-pairs signed-ranks test analyzes the differences between the paired observations, taking into account the magnitude of the differences.

It computes the differences between the pairs of variables and ranks the absolute differences.

The next step is to sum the positive and negative ranks. The test statistic, $z$, is computed from the positive and negative rank sums.

Under the null hypothesis of no difference, $z$ is a standard normal variate with mean 0 and variance 1 for large samples.
Nonparametric Tests Paired Samples

- The example considered for the paired t test, whether the respondents differed in terms of attitude toward the Internet and attitude toward technology, is considered again. Suppose we assume that both these variables are measured on ordinal rather than interval scales. Accordingly, we use the Wilcoxon test. The results are shown in Table 15.18.

- The **sign test** is not as powerful as the Wilcoxon matched-pairs signed-ranks test as it only compares the signs of the differences between pairs of variables without taking into account the ranks.

- In the special case of a binary variable where the researcher wishes to test differences in proportions, the McNemar test can be used. Alternatively, the chi-square test can also be used for binary variables.
### Wilcoxon Matched-Pairs Signed-Rank Test

Internet with Technology

<table>
<thead>
<tr>
<th>(Technology - Internet)</th>
<th>Cases</th>
<th>Mean rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>-Ranks</td>
<td>23</td>
<td>12.72</td>
</tr>
<tr>
<td>+Ranks</td>
<td>1</td>
<td>7.50</td>
</tr>
<tr>
<td>Ties</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>30</strong></td>
<td></td>
</tr>
</tbody>
</table>

\[ z = -4.207 \]

2-tailed \( p = 0.0000 \)
## A Summary of Hypothesis Tests Related to Differences

Table 15.19

<table>
<thead>
<tr>
<th>Sample</th>
<th>Application</th>
<th>Level of Scaling</th>
<th>Test/Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>One Sample</td>
<td>Proportion</td>
<td>Metric</td>
<td>Z test</td>
</tr>
<tr>
<td>One Sample</td>
<td>Distributions</td>
<td>Nonmetric</td>
<td>K-S and chi-square for goodness of fit</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Runs test for randomness</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Binomial test for goodness of fit for dichotomous variables</td>
</tr>
<tr>
<td>One Sample</td>
<td>Means</td>
<td>Metric</td>
<td>( t ) test, if variance is unknown</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>( z ) test, if variance is known</td>
</tr>
</tbody>
</table>
## A Summary of Hypothesis Tests Related to Differences

### Table 15.19, cont.

<table>
<thead>
<tr>
<th>Two Independent Samples</th>
<th>Two independent samples</th>
<th>Distributions</th>
<th>Nonmetric</th>
<th>K-S two-sample test for examining the equivalence of two distributions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Two independent samples</td>
<td>Means</td>
<td>Metric</td>
<td></td>
<td>Two-group $t$ test</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$F$ test for equality of variances</td>
</tr>
<tr>
<td>Two independent samples</td>
<td>Proportions</td>
<td>Metric</td>
<td></td>
<td>$z$ test</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Nonmetric</td>
<td></td>
<td>Chi-square test</td>
</tr>
<tr>
<td>Two independent samples</td>
<td>Rankings/Medians</td>
<td>Nonmetric</td>
<td></td>
<td>Mann-Whitney U test is more powerful than the median test</td>
</tr>
</tbody>
</table>
## A Summary of Hypothesis Tests Related to Differences

### Table 15.19, cont.

<table>
<thead>
<tr>
<th>Paired Samples</th>
<th>Means</th>
<th>Metric</th>
<th>Paired $t$ test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paired samples</td>
<td>Proportions</td>
<td>Nonmetric</td>
<td>McNemar test for binary variables</td>
</tr>
<tr>
<td></td>
<td>Rankings/Medians</td>
<td>Nonmetric</td>
<td>Chi-square test</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Wilcoxon matched-pair: ranked-signs</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>test is more powerful than the sign</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>test</td>
</tr>
</tbody>
</table>
SPSS Windows

• The main program in SPSS is FREQUENCIES. It produces a table of frequency counts, percentages, and cumulative percentages for the values of each variable. It gives all of the associated statistics.

• If the data are interval scaled and only the summary statistics are desired, the DESCRIPTIVES procedure can be used.

• The EXPLORE procedure produces summary statistics and graphical displays, either for all of the cases or separately for groups of cases. Mean, median, variance, standard deviation, minimum, maximum, and range are some of the statistics that can be calculated.
To select these procedures click:

**Analyze>Descriptive Statistics>Frequencies**
**Analyze>Descriptive Statistics>Descriptives**
**Analyze>Descriptive Statistics>Explore**

The major cross-tabulation program is CROSSTABS. This program will display the cross-classification tables and provide cell counts, row and column percentages, the chi-square test for significance, and all the measures of the strength of the association that have been discussed.

To select these procedures, click:

**Analyze>Descriptive Statistics>Crosstabs**
The major program for conducting parametric tests in SPSS is COMPARE MEANS. This program can be used to conduct $t$ tests on one sample or independent or paired samples. To select these procedures using SPSS for Windows, click:

Analyze>Compare Means>Means ...
Analyze>Compare Means>One-Sample T Test ...
Analyze>Compare Means>Independent-Samples T Test ...
Analyze>Compare Means>Paired-Samples T Test ...
The nonparametric tests discussed in this chapter can be conducted using NONPARAMETRIC TESTS.

To select these procedures using SPSS for Windows, click:

- Analyze>Nonparametric Tests>Chi-Square ...
- Analyze>Nonparametric Tests>Binomial ...
- Analyze>Nonparametric Tests>Runs ...
- Analyze>Nonparametric Tests>1-Sample K-S ...
- Analyze>Nonparametric Tests>2 Independent Samples ...
- Analyze>Nonparametric Tests>2 Related Samples ...
1. Select ANALYZE on the SPSS menu bar.
2. Click DESCRIPTIVE STATISTICS and select FREQUENCIES.
3. Move the variable “Familiarity [familiar]” to the VARIABLE(s) box.
4. Click STATISTICS.
5. Select MEAN, MEDIAN, MODE, STD. DEVIATION, VARIANCE, and RANGE.
6. Click CONTINUE.
7. Click CHARTS.
8. Click HISTOGRAMS, then click CONTINUE.
9. Click OK.
1. Select ANALYZE on the SPSS menu bar.
2. Click on DESCRIPTIVE STATISTICS and select CROSSTABS.
3. Move the variable “Internet Usage Group [iusagegr]” to the ROW(S) box.
4. Move the variable “Sex[sex]” to the COLUMN(S) box.
5. Click on CELLS.
6. Select OBSERVED under COUNTS and COLUMN under PERCENTAGES.
7. Click CONTINUE.
8. Click STATISTICS.
9. Click on CHI-SQUARE, PHI AND CRAMER’S V.
10. Click CONTINUE.
11. Click OK.
SPSS Windows: One Sample $t$ Test

1. Select ANALYZE from the SPSS menu bar.
2. Click COMPARE MEANS and then ONE SAMPLE T TEST.
3. Move “Familiarity [familiar]” into the TEST VARIABLE(S) box.
4. Type “4” in the TEST VALUE box.
5. Click OK.
1. Select ANALYZE from the SPSS menu bar.
2. Click COMPARE MEANS and then INDEPENDENT SAMPLES T TEST.
3. Move “Internet Usage Hrs/Week [iusage]” in to the TEST VARIABLE(S) box.
4. Move “Sex[sex]” to GROUPING VARIABLE box.
5. Click DEFINE GROUPS.
6. Type “1” in GROUP 1 box and “2” in GROUP 2 box.
7. Click CONTINUE.
8. Click OK.
1. Select ANALYZE from the SPSS menu bar.

2. Click COMPARE MEANS and then PAIRED SAMPLES T TEST.

3. Select “Attitude toward Internet [iattitude]” and then select “Attitude toward technology [tattitude].” Move these variables in to the PAIRED VARIABLE(S) box.

4. Click OK.
Figure 15.10 A Concept Map for Frequency Distribution

Frequency Analysis

- The Frequency of Each Value
  - divide by the total number of responses
  - Percentage for Each Value
    - adjust for missing values
      - Valid Percentage for Each Value
        - % of responses ≤ each value
          - Cumulative Percentages for Each value
            - plot
              - Frequency Histogram

Descriptive Statistics

- measures of Location
  - Mean
  - Mode
  - Median

- measures of Variability
  - Range
  - Variance
  - Standard Deviation
Figure 15.11 A Concept Map for Cross-Tabulation

Construct the Cross-Tabulation Table

Null Hypothesis, $H_0$: There is no Association

- $H_0$ is not rejected: There is no relationship
- $H_0$ is rejected: There is an Association

There is an Association

determine

Strength of Association

interpret

Pattern of Relationship

draw

Marketing conclusion

Chi-Square Statistic, $\chi^2$

use appropriate statistic

Phi Coefficient, $\phi$

Contingency Coefficient, $C$

Cramer's $V$
Figure 15.12 A Concept Map for Conducting t-Tests

- **Hypothesis Testing**
  - formulate
  - select

- **H₀ and H₁**
  - guided by:
  - H₀: Reflects Status Quo Whereas H₁: is One in Which an Effect is Expected

- **Appropriate t-Test**
  - choose
  - consider:
    - Number of Samples (One or Two) and Whether the Samples are Independent or Paired
    - Type-I Error, α and Type-II Error, β
  - commonly:
    - α = 0.05

- **Level of Significance, α**
  - collect data
  - determine

- **Calculate the Appropriate t Statistic**
  - determine

- **Probability Associated with t Statistic (TS_cal)**
  - compare with:
    - Level of Significance, α
    - Calculated Value of t Statistic (TS_cal)
  - results in:
    - Reject or Do Not Reject H₀
    - Marketing Research Conclusion

- **Critical Value of t Statistic (TS_cr)**
  - compare with (using absolute values):
    - Probability of TS_cal < α
    - |TS_cal| > |TS_cr|
  - results in:
    - Recommend Status Quo
    - Recommend Action Implied by H₁
OBJECTIVES

1. Discuss the nature and scope of data preparation and the data preparation process.
2. Explain questionnaire checking and editing, and treatment of unsatisfactory responses by returning to the field, assigning missing values, and discarding unsatisfactory responses.
3. Describe the guidelines for coding questionnaires, including the coding of structured and unstructured questions.
4. Discuss the data-cleaning process and the methods used to treat missing responses: substitution of a neutral value, imputed response, casewise deletion, and pairwise deletion.
5. State the reasons for and methods of statistically adjusting data: weighting, variable respecification, and scale transformation.
6. Describe the procedure for selecting a data analysis strategy and the factors influencing the process.
7. Classify statistical techniques and give a detailed classification of univariate techniques as well as a classification of multivariate techniques.
8. Explain the use of the Internet and statistical software in data preparation and analysis.
OBJECTIVES

9. Describe the significance of preliminary data analysis and the insights that can be obtained from such an analysis.

10. Discuss data analysis associated with frequencies including measures of location, measures of variability, and measures of shape.


12. Describe data analysis associated with parametric hypothesis testing for one sample, two independent samples, and paired samples.

13. Understand data analysis associated with nonparametric hypothesis testing for one sample, two independent samples, and paired samples.