ML 4202 Chapter 15

Overview of the Data Analysis Procedure:

- Five step procedure
    - Step one: Validation and editing (quality control)
    - Step two: Coding
    - Step three: Data entry
    - Step four: Logical cleaning of data
    - Step five: Tabulation and statistical analysis

Step One: Validation and Editing

- The purpose of the first step is twofold. The researcher wants to make sure that all the interviews actually were conducted as specified (validation) and that the questionnaires have been filled out properly and completely (editing)
- **Validation**: Process of ascertaining that interviews actually were conducted as specified.
    - The goal of validation is solely to detect interviewer fraud or failure to follow key instructions.
    - Percentage ranges from 10-20% (it should be at this amount after all the interviews have been completed). If a particular interviewer surveys 50 people and the research firm normally validates at a 10% rate, 5 respondents surveyed by that interviewer would be recontacted to determine a specific set of questions
- **Editing**: Process of ascertaining that questionnaires were filled out properly or completely.
    - Paper questionnaires usually are edited at least twice before being submitted for data entry. First, they may be edited by the field service firm that conducted the interviews, and then they are edited by the marketing research firm that hired the field service firm to do the interviewing.
    - Editing process for paper surveys involves manually checking for a numbers of problems
        - Whether the interviewer failed to ask certain questions or record answer for certain questions.
        - Whether **skip patterns** were followed. **Skip pattern**: Sequence in which later questions are asked, based on a respondent's answer to an earlier question or questions.
        - Whether the interviewer paraphrased respondents' answers to open-ended questions. Interviewers are trained to record responses verbatim and not to paraphrase or insert their own language.
        - Extremely tedious and time-consuming

Step Two: Coding

- **Coding**: Process of grouping and assigning numeric codes to the various responses to a question.

- Many questions are closed-ended and precoded, meaning that numeric codes have been assigned to the various responses on the questionnaire.
- Open-ended questions create a coding dilemma.

- Coding Process:
  - *List responses*- Coders at the research firm prepare lists of the actual responses given to each open-ended question on a particular survey.
  - *Consolidate responses*-Requires a number of subjective decisions
  - *Set codes*- A numeric code is assigned to each of the categories on the final consolidated list of responses.
  - *Enter codes*- Involves 3 steps
    - Read responses to individual open-ended questions on questionnaires.
    - Match individual responses with the consolidated list of response categories and determine the appropriate numeric code for each response
    - Write the numeric code in the appropriate place on the questionnaire for the response to the particular question or enter the appropriate code in the database electronically.
- Automated Coding Systems:
  - With CATI and Internet surveys, data entry and coding are completely eliminated for closed-ended questions.

Step Three: Data Entry

- **Data entry**: Process of converting information to an electronic format
- **Intelligent data entry**: Form of data entry in which the information being entered into the data entry device is checked for internal logic.
  - They can be programmed to avoid certain types of errors at the point of data entry, such as invalid or wild codes and violation of skip patterns.
- The Data Entry Process:
  - In the upper right-hand corner of the questionnaire, the number 001 is written. The number uniquely identifies the particular questionnaire, which should be the first questionnaire in the stack that the data entry operator is preparing to enter.
  - To the left of the handwritten number 001 is (01-03). This tells the data entry operator that 001 should be entered into fields 01-03.
  - Exhibit 15.1 clearly illustrates the relationship between the layout of the questionnaire, in terms of codes (numbers associated with different answers to questions) and fields (places on the data record where the code is entered) and the layout of the data record.

Scanning:

- The limited use of scanning in marketing research can be attributed to two factors: setup costs and the need to record all responses with a No.2 pencil. Setup costs include the cost of special paper, special ink in the printing process, and very precise placement of the bubbles for recording responses. The break-even point, at which the savings in data entry costs exceeded the setup costs is in the 10,000 to 12,000 survey range. Therefore, for most surveys, scanning was not feasible.
- **Scanning technology**: Form of data entry in which responses on questionnaires are read in automatically by the data entry device.
- Nowadays, it doesn't matter what kind of writing utensil is used (pencil, ballpoint pen, or ink pen) and it also doesn't matter if a full bubble is filled in... simply an X will do

Step Four: Logical Cleaning of Data

- **Logical or machine cleaning of data:** Final computerized error check of data
- **Error checking routines:** Computer programs that accept instructions from the user to check for logical errors in the data

Step Five: Tabulation and Statistical Analysis

- **One-way frequency table**: Table showing the number of respondents choosing each answer to a survey question.
- There are three options for a base:
  - **Total respondents:** If 300 people are interviewed in a particular study and the decision is to use total respondents as the base for calculating percentages, then the percentages in each one-way frequency table will be based on 300 respondents.
  - **Number of people asked the particular question:** Some questions are just specifically aimed towards a group of people for example: 200 people have dogs, so questions 5 and 6 that are just directed to dog owners, and should only be asked for 200 respondents.
  - **Number of people answering the question:** If 200 people were asked a particular question but 28 said I don't know then the response percentage should be based off on 172 people
    - Ordinarily, the number of people who were asked a particular question is used as the base for all percentages throughout the tabulations, but there may be special cases in which other bases are judged appropriate.
- **Cross tabulations**: Examination of the responses to one question relative to the responses to one or more other questions.
  - Includes frequencies and percentages, with the percentages based on column totals.
  - A common way of setting up cross-tabulation tables is to use columns to represent factors such as demographics and lifestyle characteristics, which are expected to be predictors of the state of mind, behavior, or intentions data,