

Walter R. Paczkowski

Modern Survey Analysis

Using Python for Deeper Insights



Springer

Modern Survey Analysis

Walter R. Paczkowski

Modern Survey Analysis

Using Python for Deeper Insights

 Springer

Walter R. Paczkowski
Data Analytics Corp.
Plainsboro, NJ, USA

ISBN 978-3-030-76266-7 ISBN 978-3-030-76267-4 (eBook)
<https://doi.org/10.1007/978-3-030-76267-4>

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Switzerland AG 2022

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

The historical root for my professional career as a data scientist, including my own consulting company which is focused on data science in general, has been survey analysis, primarily consumer surveys in the marketing domain. My experience has run the gamut from simple consumer attitudes, opinions, and interest (*AIO*) surveys to complex discrete choice, market segmentation, messaging and claims, pricing, and product positioning surveys. And the purpose for these has varied from just informative market scanning to in-depth marketing mix and new product development work. These all have been for companies in a wide variety of industries such as jewelry, pharmaceuticals, household products, education, medical devices, and automotive to mention a few. I learned a lot about survey data: how to collect them, organize them for analysis, and, of course, analyze them for actionable insight and recommendations for my clients. This book is focused on analyzing survey data based on what I learned.

I have two overarching objectives for this book:

1. Show how to extract actionable, insightful, and useful information from survey data
2. Show how to use Python to analyze survey data

Why Surveys?

Why focus on surveys other than the fact that this is my career heritage? The answer is simple. Surveys are a main source of data for key decision makers (*KDMs*), whether in the private or public sector. They need this data for the critical decisions they must make every day, decisions that have short-term and long-term implications and effects. They are not the only and definitely not the least important source. There are four sources that are relied on to some extent, the extent varying by the type of *KDM* and problem. The sources, not in any order, are:

1. Observational
2. Sensors
3. Experimental
4. Surveys

Observational and sensor measurements are historical data—data on what happened. These could be transactional (such as when customers shopped), production, employment, voter registrations and turnout, and the list goes on. Some are endogenous to the business or public agency, meaning they are the result of actions or decisions made by *KDMs* in the daily running of the business or public life. They ultimately have control over how such data are generated (besides random events which no one can control). Other data are exogenous, meaning they are determined or generated by forces outside the control of the *KDMs* and are over and beyond random events. The movement of the economy through a business cycle is a good example. Regardless of the form (endogenous or exogenous), data represent what did happen or is currently happening.

Sensor-generated data are in the observational category. The difference is more degree than kind. Sensor data are generated in real-time and transmitted to a central data collection point, usually over wireless sensor networks (*WSN*). The result is a data flood, a deluge that must be stored and processed almost instantaneously. These data could represent measures in a production process, health measures in a medical facility, automobile performance measures, traffic patterns on major thoroughfares, and so forth. But all this sensor-generated data also represent what did happen or is currently happening. See Paczkowski (2020) for some discussion of sensor data and *WSNs* in the context of new product development.

Experimental data are derived from designed experiments that have very rigid protocols to ensure that every aspect of a problem (i.e., factors or attributes) has equal representation in a study, that is, the experiment. Data are not historical as for observational and sensor data but “what-if” in nature: what-if about future events under controlled conditions. Examples are:

- *What if temperature is set at a high vs. low level?* This is an industrial experiment.
- *What if price is \$X rather than \$Y?* This is a marketing experiment.
- *What if one color is used rather than another?* This is a product development experiment.
- *How would you vote change if candidate XX drops out of the presidential race?* This is a political issue.

Observational and sensor measurements are truly data, that is, they are facts. Some experimental studies, such as those listed above, will tell you about opinions, while others (e.g., the industrial experiments) will not. Generally, none of these will tell you about people’s opinions, plans, attitudes, reasons, understanding, awareness, familiarity, or concerns, all of which are subjective and personal. This list is more emotional, intellectual, and knowledge based. Items on the list are concerned with what people feel, believe, and know rather than on what they did or could do under different conditions. This is where surveys enter the picture. Marketing and public

opinion what-if experiments are embedded in surveys so they are a hybrid of the two forms.

Surveys can be combined with the other three forms. They allow you, for instance, to study artificial, controlled situations as in an industrial experiment. For example, in a pricing study, surveys could reveal preferences for pricing programs, strategies, and willingness to pay without actually changing prices. Conjoint, MaxDiff, and discrete choice studies are examples of experiments conducted within a survey framework. For what follows, I will differentiate between industrial and non-industrial experiments, the latter including marketing and opinion poll experiments embedded in surveys.

Surveys get to an aspect of people's psyche. Behavior can certainly be captured by asking survey respondents what they recently did (e.g., how much did they spend on jewelry this past holiday season) or might do under different conditions (e.g., will they still purchase if the price rises by X%?). These are not as accurate as direct observation, or measured by sensors, or derived from industrial experiments because they rely on what people have to say – and people are not always accurate or truthful in this regard. Even marketing experiments are not as accurate as actual purchase data because people tend to overstate how much they will buy, so such data have to be calibrated to make them more reasonable. Nonetheless, compared to the other three forms of data collection, surveys are the only way to get at what people are thinking.

Why should it matter what people think? This is important because people (as customers, clients, and constituents) make personal decisions, based on what they know or are told, regarding purchases, what to request, what to register for, or who to vote for. These decisions are reflected in actual market behavior (i.e., purchases) or votes cast. Knowing how people think helps explain the observed behavior. Without an explanation, then all you have is observed behavior void of understanding. In short, surveys help to add another dimension to the data collected from the other three data collection methods, especially observed transactional data.

Surveys have limitations, not the least of which are:

1. People's responses are very subjective and open to interpretation.
2. People's memories are dubious, foggy, and unclear.
3. People's predictions of their own behavior (e.g., purchase intent or vote to cast) may not be fulfilled for a host of unknown and unknowable causes.
4. People tend to overstate intentions (e.g., how much they will spend on gifts for the next holiday season).

The other data collection methods also have their shortcomings, so the fact that surveys are not flawless is not a reason not to use them. You just need to know how to use them. This includes how to structure and conduct a survey, how to write a questionnaire, and, of course, how to analyze data. This book focuses on the last way – analyzing survey data for actionable, insightful, and useful information.

Why Python?

The second overarching goal for this book is to describe how Python can be used for survey data analysis. Python has several advantages in this area such as:

- It is free.
- It has a rich array of packages for analyzing data in general.
- It is programmable – every analyst should know some programming – and it is easy to program.

You could ask “*Why not just use spreadsheets*”? Unfortunately, spreadsheets have major issues, several of which are:

- Data are often spread across several worksheets in a workbook.
- They make it difficult to identify data.
- They lack table operations such as joining, splitting, or stacking.
- They lack programming capabilities except Visual Basic for Applications (VBA), which is not a statistical programming language.
- They lack sophisticated statistical operations beyond arithmetic operations and simple regression analysis (add-on packages help, but they tend to lack depth and rely on the spreadsheet engine.)
- Spreadsheets are notorious for making it difficult to track formulas and catch errors. Each cell could have a separate formula, even cells in the same column for a single variable.
- The formula issue leads to reproducibility problems. The cells in the spreadsheet are linked, even across spreadsheets in the same workbook or across workbooks, often with no clear pattern. Tracing and reproducing an analysis is often difficult or impossible.
- Graphics are limited.

Preliminaries for Getting Started

To successfully read this book, you will need Python and Pandas (and other Python packages) installed on your computer so you can follow the examples. This book is meant to be interactive and not static. A static book is one that you just read and try to absorb its messages. An interactive book is one that you read and then reproduce the examples. The examples are generated in a Jupyter notebook. A Jupyter notebook is the main programming tool of choice by data scientists for organizing, conducting, and documenting their statistical and analytical work. It provides a convenient way to enter programming commands, get the output from those commands, and document what was done or what is concluded from the output. The output from executing a command immediately follows the command so input and output “stay together.” I do everything in Jupyter notebooks.

I provide screenshots of how to run commands and develop analyses along with the resulting output. This way, the Python code and resulting output are presented as a unit. In addition, the code is all well documented with comments so you can easily follow the steps I used to do a task. But of course, you can always go back to the Jupyter notebooks to see the actual code and run them yourself.

I strongly recommend that you have Jupyter installed since Jupyter notebooks will be illustrated in this book. A Jupyter notebook of this book's contents is available. If you do not have Jupyter, Python, and Pandas available, then I recommend that you download and install Anaconda,¹ a freeware package that gives you access to everything you will need. Just select the download appropriate for your operating system. After you install Anaconda, you can use the *Anaconda Navigator* to launch Jupyter.²

A basic, introductory course in statistics is beneficial, primarily for later chapters.

The Book's Structure

This book has seven chapters. Chapter 1 sets the stage with a discussion of the importance of surveys and Python. Chapter 2 focuses on knowing the structure of data, which is really the profile of the survey respondents. Chapter 3 is concerned with shallow data analysis. This is simple statistics and simple visualizations such as bar/pie charts of main survey questions. This is where many analyses of survey data end. Chapter 4 is about deep data analysis that goes beyond the shallow analyses. Chapter 5 extends the deep analysis begun in Chap. 4 by introducing three regression models for deep analysis: *OLS*, logistic regression, and Poisson regression. Chapter 6 covers some specialized survey objectives to illustrate some of the concepts developed in the previous chapters. Chapter 7 changes focus and covers complex sample surveys. Different stages of complex samples are covered. Chapters 8 and 9 cover advanced material: Bayesian statistics applied to survey data analysis. You may be familiar with some Bayesian concepts. If not, then Chap. 8 will help you because it covers the basic concepts leading to Bayes' Rule. I show in this chapter how to estimate Bayesian models using a Python package. I then extend the material in Chap. 8 to more advanced material in Chap. 9. These chapters will provide you with a new perspective on survey data and how to include prior information into your analyses.

Plainsboro, NJ, USA

Walter R. Paczkowski

¹ Download Anaconda from <https://www.anaconda.com/download/>.

² Please note that there is Jupyter and JupyterLab. JupyterLab is the newer development version of Jupyter, so it is not ready for "prime time." I will only use Jupyter which is stable at this time.

Acknowledgments

In my last book, I noted the support and encouragement I received from my wonderful wife, Gail; and my two daughters, Kristin and Melissa. As before, Gail encouraged me to sit down and just write, especially when I did not want to, while my daughters provided the extra set of eyes I needed to make this book perfect. They provided the same support and encouragement for this book, so I owe them a lot, both then and now. I would also like to say something about my two grandsons who, now at 6 and 10, obviously did not contribute to this book but who, I hope, will look at this one in their adult years and say “*Yup. My grandpa wrote this book, too.*”

Contents

1	Introduction to Modern Survey Analytics	1
1.1	Information and Survey Data	3
1.2	Demystifying Surveys	4
1.2.1	Survey Objectives	5
1.2.2	Target Audience and Sample Size	7
1.2.2.1	Key Parameters to Estimate	9
1.2.2.2	Sample Design to Use	9
1.2.2.3	Population Size	10
1.2.2.4	Alpha	10
1.2.2.5	Margin of Error	10
1.2.2.6	Additional Information	10
1.2.3	Screener and Questionnaire Design	12
1.2.4	Fielding the Study	14
1.2.5	Data Analysis	14
1.2.6	Report Writing and Presentation	16
1.3	Sample Representativeness	16
1.3.1	Digression on Indicator Variables	20
1.3.2	Calculating the Population Parameters	21
1.4	Estimating Population Parameters	22
1.5	Case Studies	25
1.5.1	Consumer Study: Yogurt Consumption	25
1.5.2	Public Sector Study: VA Benefits Survey	27
1.5.3	Public Opinion Study: Toronto Casino Opinion Survey	28
1.5.4	Public Opinion Study: San Francisco Airport Customer Satisfaction Survey	30
1.6	Why Use Python for Survey Data Analysis?	30
1.7	Why Use Jupyter for Survey Data Analysis?	32

- 2 First Step: Working with Survey Data** 35
 - 2.1 Best Practices: First Steps to Analysis 36
 - 2.1.1 Installing and Importing Python Packages 36
 - 2.1.2 Organizing Routinely Used Packages, Functions, and Formats 39
 - 2.1.3 Defining Data Paths and File Names 41
 - 2.1.4 Defining Your Functions and Formatting Statements 42
 - 2.1.5 Documenting Your Data with a Dictionary 42
 - 2.2 Importing Your Data with Pandas 43
 - 2.3 Handling Missing Values 48
 - 2.3.1 Identifying Missing Values 49
 - 2.3.2 Reporting Missing Values 49
 - 2.3.3 Reasons for Missing Values 50
 - 2.3.4 Dealing with Missing Values 51
 - 2.3.4.1 Use the *fillna()* Method 51
 - 2.3.4.2 Use the *Interpolation()* Method 51
 - 2.3.4.3 An Even More Sophisticated Method 52
 - 2.4 Handling Special Types of Survey Data 52
 - 2.4.1 CATA Questions 52
 - 2.4.1.1 Multiple Responses 53
 - 2.4.1.2 Multiple Responses by ID 53
 - 2.4.1.3 Multiple Responses Delimited 54
 - 2.4.1.4 Indicator Variable 54
 - 2.4.1.5 Frequencies 54
 - 2.4.2 Categorical Questions 54
 - 2.5 Creating New Variables, Binning, and Rescaling 56
 - 2.5.1 Creating Summary Variables 58
 - 2.5.2 Rescaling 62
 - 2.5.3 Other Forms of Preprocessing 64
 - 2.6 Knowing the Structure of the Data Using Simple Statistics 67
 - 2.6.1 Descriptive Statistics and DataFrame Checks 68
 - 2.6.2 Obtaining Value Counts 69
 - 2.6.3 Styling Your DataFrame Display 69
 - 2.7 Weight Calculations 70
 - 2.7.1 Complex Weight Calculation: Raking 73
 - 2.7.2 Types of Weights 75
 - 2.8 Querying Data 80
- 3 Shallow Survey Analysis** 83
 - 3.1 Frequency Summaries 84
 - 3.1.1 Ordinal-Based Summaries 85
 - 3.1.2 Nominal-Based Summaries 86
 - 3.2 Basic Descriptive Statistics 86
 - 3.3 Cross-Tabulations 89

- 3.4 Data Visualization 94
 - 3.4.1 Visuals Best Practice 95
 - 3.4.2 Data Visualization Background 95
 - 3.4.3 Pie Charts 98
 - 3.4.4 Bar Charts 99
 - 3.4.5 Other Charts and Graphs 101
 - 3.4.5.1 Histograms and Boxplots for Distributions 105
 - 3.4.5.2 Mosaic Charts 105
 - 3.4.5.3 Heatmaps 109
- 3.5 Weighted Summaries: Crosstabs and Descriptive Statistics 111
- 4 Beginning Deep Survey Analysis 113**
 - 4.1 Hypothesis Testing 114
 - 4.1.1 Hypothesis Testing Background 115
 - 4.1.2 Examples of Hypotheses 118
 - 4.1.3 A Formal Framework for Statistical Tests 118
 - 4.1.4 A Less Formal Framework for Statistical Tests 119
 - 4.1.5 Types of Tests to Use 120
 - 4.2 Quantitative Data: Tests of Means 122
 - 4.2.1 Test of One Mean 122
 - 4.2.2 Test of Two Means for Two Populations 126
 - 4.2.2.1 Standard Errors: Independent Populations 126
 - 4.2.2.2 Standard Errors: Dependent Populations 129
 - 4.2.3 Test of More Than Two Means 131
 - 4.3 Categorical Data: Tests of Proportions 142
 - 4.3.1 Single Proportions 143
 - 4.3.2 Comparing Proportions: Two Independent Populations 144
 - 4.3.3 Comparing Proportions: Paired Populations 146
 - 4.3.4 Comparing Multiple Proportions 147
 - 4.4 Advanced Tabulations 153
 - 4.5 Advanced Visualization 158
 - 4.5.1 Extended Visualizations 159
 - 4.5.2 Geographic Maps 162
 - 4.5.3 Dynamic Graphs 165
- Appendix 166
- 5 Advanced Deep Survey Analysis: The Regression Family 177**
 - 5.1 The Regression Family and Link Functions 178
 - 5.2 The Identity Link: Introduction to *OLS* Regression 179
 - 5.2.1 *OLS* Regression Background 180
 - 5.2.2 The Classical Assumptions 180
 - 5.2.3 Example of Application 181
 - 5.2.4 Steps for Estimating an *OLS* Regression 182
 - 5.2.5 Predicting with the *OLS* Model 186

- 5.3 The Logit Link: Introduction to Logistic Regression 187
 - 5.3.1 Logistic Regression Background 189
 - 5.3.2 Example of Application 192
 - 5.3.3 Steps for Estimating a Logistic Regression 194
 - 5.3.4 Predicting with the Logistic Regression Model 200
- 5.4 The Poisson Link: Introduction to Poisson Regression 200
 - 5.4.1 Poisson Regression Background 200
 - 5.4.2 Example of Application 201
 - 5.4.3 Steps for Estimating a Poisson Regression 201
 - 5.4.4 Predicting with the Poisson Regression Model 202
- Appendix 203
- 6 Sample of Specialized Survey Analyses 209**
 - 6.1 Conjoint Analysis 210
 - 6.1.1 Case Study 210
 - 6.1.2 Analysis Steps 210
 - 6.1.3 Creating the Design Matrix 211
 - 6.1.4 Fielding the Conjoint Study 212
 - 6.1.5 Estimating a Conjoint Model 214
 - 6.1.6 Attribute Importance Analysis 215
 - 6.2 Net Promoter Score 217
 - 6.3 Correspondence Analysis 224
 - 6.4 Text Analysis 228
- 7 Complex Surveys 237**
 - 7.1 Complex Sample Survey Estimation Effects 239
 - 7.2 Sample Size Calculation 240
 - 7.3 Parameter Estimation 241
 - 7.4 Tabulation 244
 - 7.4.1 Tabulation 245
 - 7.4.2 CrossTabulation 245
 - 7.5 Hypothesis Testing 246
 - 7.5.1 One-Sample Test: Hypothesized Mean 247
 - 7.5.2 Two-Sample Test: Independence Case 248
 - 7.5.3 Two-Sample Test: Paired Case 248
- 8 Bayesian Survey Analysis: Introduction 251**
 - 8.1 Frequentist vs Bayesian Statistical Approaches 253
 - 8.2 Digression on Bayes’ Rule 259
 - 8.2.1 *Bayes’ Rule* Derivation 259
 - 8.2.2 *Bayes’ Rule* Reexpressions 261
 - 8.2.3 The Prior Distribution 262
 - 8.2.4 The Likelihood Function 263
 - 8.2.5 The Marginal Probability Function 263
 - 8.2.6 The Posterior Distribution 264
 - 8.2.7 Hyperparameters of the Distributions 264

- 8.3 Computational Method: *MCMC* 265
 - 8.3.1 Digression on Markov Chain Monte Carlo Simulation 265
 - 8.3.2 Sampling from a Markov Chain Monte Carlo Simulation 269
- 8.4 Python Package *pyMC3*: Overview 269
- 8.5 Case Study 270
 - 8.5.1 Basic Data Analysis 272
- 8.6 Benchmark *OLS* Regression Estimation 273
- 8.7 Using *pyMC3* 274
 - 8.7.1 *pyMC3* Bayesian Regression Setup 274
 - 8.7.2 Bayesian Estimation Results 280
 - 8.7.2.1 The *MAP* Estimate 280
 - 8.7.2.2 The Visualization Output 282
- 8.8 Extensions to Other Analyses 289
 - 8.8.1 Sample Mean Analysis 290
 - 8.8.2 Sample Proportion Analysis 290
 - 8.8.3 Contingency Table Analysis 291
 - 8.8.4 Logit Model for Contingency Table 295
 - 8.8.5 Poisson Model for Count Data 297
- 8.9 Appendix 300
 - 8.9.1 Beta Distribution 300
 - 8.9.2 Half-Normal Distribution 300
 - 8.9.3 Bernoulli Distribution 301
- 9 Bayesian Survey Analysis: Multilevel Extension** 303
 - 9.1 Multilevel Modeling: An introduction 304
 - 9.1.1 Omitted Variable Bias 305
 - 9.1.2 Simple Handling of Data Structure 307
 - 9.1.3 Nested Market Structures 307
 - 9.2 Multilevel Modeling: Some Observations 308
 - 9.2.1 Aggregation and Disaggregation Issues 309
 - 9.2.2 Two Fallacies 310
 - 9.2.3 Terminology 311
 - 9.2.4 Ubiquity of Hierarchical Structures 311
 - 9.3 Data Visualization of Multilevel Data 312
 - 9.3.1 Basic Data Visualization and Regression Analysis 313
 - 9.4 Case Study Modeling 318
 - 9.4.1 Pooled Regression Model 318
 - 9.4.2 Unpooled (Dummy Variable) Regression Model 319
 - 9.4.3 Multilevel Regression Model 321
 - 9.5 Multilevel Modeling Using *pyMC3*: Introduction 323
 - 9.5.1 Multilevel Model Notation 324
 - 9.5.2 Multilevel Model Formulation 324
 - 9.5.3 Example Multilevel Estimation Set-up 325
 - 9.5.4 Example Multilevel Estimation Analyses 328
 - 9.6 Multilevel Modeling with Level Explanatory Variables 328

- 9.7 Extensions of Multilevel Models 328
 - 9.7.1 Logistic Regression Model 330
 - 9.7.2 Poisson Model 332
 - 9.7.3 Panel Data 332
- Appendix 333

- References** 337
- Index** 343

List of Figures

Fig. 1.1	The Survey Design Process	5
Fig. 1.2	General Questionnaire Structure	13
Fig. 1.3	Creating an Indicator Function in Python	21
Fig. 1.4	Yogurt Sample Size Calculation	26
Fig. 1.5	Yogurt Consumption Questionnaire Structure.....	27
Fig. 1.6	VA Study Population Control Totals	28
Fig. 1.7	Vets Questionnaire Structure	29
Fig. 1.8	Toronto Casino Questionnaire Structure.....	30
Fig. 1.9	San Francisco International Airport Customer Satisfaction Questionnaire Structure	31
Fig. 1.10	Anaconda Navigator Page	33
Fig. 1.11	Anaconda Environment Page.....	33
Fig. 1.12	Jupyter Dashboard	34
Fig. 2.1	This illustrates the connection between functions and methods for enhanced functionality in Python	39
Fig. 2.2	Python Package Import	40
Fig. 2.3	Use of the %run Magic to Import Packages	40
Fig. 2.4	Illustrative Data and Notebook Path Hierarchy	41
Fig. 2.5	Example of Importing Data	42
Fig. 2.6	Importing a CSV File Into Pandas	45
Fig. 2.7	Importing an Excel Worksheet Into Pandas	46
Fig. 2.8	Importing an SPSS Worksheet Into Pandas	47
Fig. 2.9	Importing an SPSS Using pyReadStat	47
Fig. 2.10	Life Question for <i>pyreadstat</i> Examples.....	48
Fig. 2.11	Retrieving Column Label (Question) for Column Name	48
Fig. 2.12	Retrieving Value Labels (Question Options) for Column Name	49
Fig. 2.13	Categorical Coding of a Likert Scale Variable	55
Fig. 2.14	Value Counts for VA Data without Categorical Declaration	57
Fig. 2.15	Value Counts for VA Data with Categorical Declaration.....	57
Fig. 2.16	Application of CategoricalDtype.....	58
Fig. 2.17	Recoding of Yogurt Satisfaction Data	59

Fig. 2.18 Age Calculation from Vet YOB 60

Fig. 2.19 Military Branch Calculation for the Vet data 61

Fig. 2.20 Simple Weight Calculation in Python 71

Fig. 2.21 Merging Weights into a DataFrame 72

Fig. 2.22 Raking Script 76

Fig. 2.23 Raking with ipfn Function 77

Fig. 2.24 Weights Based om Raking 78

Fig. 2.25 Stacked Weights for Merging 78

Fig. 2.26 Analysis of Stacked Weights 79

Fig. 2.27 Query of Female Voters Only 80

Fig. 2.28 Query of Female Voters Who Are 100% Likely to Vote 81

Fig. 2.29 Query of Female Voters Who Are 100% Likely to Vote or
Extremely Likely to Vote 81

Fig. 3.1 Frequency Summary Table: Ordinal Data 86

Fig. 3.2 Frequency Summary Table: Nominal Data 87

Fig. 3.3 Yogurt Data Subset 88

Fig. 3.4 Yogurt Data Descriptive Statistics 88

Fig. 3.5 Example of Mean Calculation 89

Fig. 3.6 Basic Crosstab 90

Fig. 3.7 Enhanced Cross-tab 91

Fig. 3.8 Enhanced Cross-tab 92

Fig. 3.9 Basic Cross-tab Using the pivot_table Method 94

Fig. 3.10 One-way Table Using the pivot_table Method 94

Fig. 3.11 Matplotlib Figure and Axis Structure 97

Fig. 3.12 Pie Chart for Likelihood to Vote 99

Fig. 3.13 Yogurt Age-Gender Distribution 100

Fig. 3.14 Pie Charts for Yogurt Age-Gender Distribution 101

Fig. 3.15 Alternative Pie Charts for Yogurt Age-Gender Distribution 102

Fig. 3.16 Yogurt Consumers' Gender Distribution 103

Fig. 3.17 Yogurt Consumers' Gender Bar Chart 103

Fig. 3.18 Stacking Data for SBS Bar Chart 104

Fig. 3.19 SBS Bar Chart for the Yogurt Age-Gender Distribution 104

Fig. 3.20 Histogram Example 106

Fig. 3.21 Boxplot Anatomy 107

Fig. 3.22 Histogram Example 107

Fig. 3.23 Mosaic Chart Using Implicit Cross-tab 108

Fig. 3.24 Mosaic Chart Using Explicit Cross-tab 108

Fig. 3.25 Mosaic Chart Using Three Variables 109

Fig. 3.26 Heatmap of the Age-Gender Distribution 110

Fig. 3.27 Check Sum of Weights 111

Fig. 3.28 Calculation of Weighted Descriptive Statistics 112

Fig. 3.29 Weighted Cross-tabs 112

Fig. 4.1 Hypothesis Testing Steps 120

Fig. 4.2 Statistical Test Flowchart 121

Fig. 4.3 Comparison of Normal and Student's t-distribution 123

Fig. 4.4 Unweighted t-Test of Yogurt Price 124

Fig. 4.5 Weighted t-Test of Yogurt Price 124

Fig. 4.6 Unweighted z-Test of Yogurt Price 125

Fig. 4.7 Weighted z-Test of Yogurt Price 125

Fig. 4.8 Unweighted Pooled t-Test Comparing Means 128

Fig. 4.9 Weighted Pooled t-Test Comparing Means 128

Fig. 4.10 Unweighted Pooled z-Test Comparing Means 129

Fig. 4.11 Weighted z-Test of Yogurt Price 129

Fig. 4.12 Paired T-test Example 130

Fig. 4.13 Missing Value Analysis for Vet Age 131

Fig. 4.14 Age Distribution of Vets 132

Fig. 4.15 Mean Age of Vets by Service Branches 133

Fig. 4.16 ANOVA Table of Age of Vets by Service Branches 133

Fig. 4.17 Probability of Incorrect Decision 140

Fig. 4.18 Summary of Tukey’s HSD Test 141

Fig. 4.19 Summary of Plot of Tukey’s HSD Test 142

Fig. 4.20 Heatmap of p-Values of Tukey’s HSD Test 143

Fig. 4.21 Missing Value Report for Question A7 145

Fig. 4.22 Statistical Test Results for Question A7 145

Fig. 4.23 Missing Value Report for Question C1 148

Fig. 4.24 Code for Missing Value Report for Question C1 149

Fig. 4.25 Summary Table and Pie Chart for Missing Value Report
for Question C1 150

Fig. 4.26 Code to create CATA Summary 151

Fig. 4.27 CATA Summary 152

Fig. 4.28 Proportion Summary for the VA CATA Question QC1a 153

Fig. 4.29 Cochran’s Q Test for the VA CATA Question QC1a 154

Fig. 4.30 Marascuillo Procedure for the VA CATA Question QC1a 155

Fig. 4.31 Results Summary of the Marascuillo Procedure for the VA
CATA Question QC1a 156

Fig. 4.32 Abbreviated Results Summary of the Marascuillo
Procedure for the VA CATA Question QC1a 157

Fig. 4.33 Response Distribution for the VA Enrollment Question
QE1: Pre-Cleaning 157

Fig. 4.34 Response Distribution for the VA Enrollment Question
QE1: Post-Cleaning 158

Fig. 4.35 Pivot Table for the VA Enrollment Question QE1: Post-Cleaning ... 159

Fig. 4.36 Grouped Boxplot of Vets’ Age Distribution 160

Fig. 4.37 3-D Bar Chart of VA Data 161

Fig. 4.38 Faceted Bar Chart of VA Data 163

Fig. 4.39 Geographic Map Data Preparation 164

Fig. 4.40 Geographic Map Code Setup 164

Fig. 4.41 Geographic Map of State of Origin 165

Fig. 4.42 Summary of Static and Dynamic Visualization Functionality 166

Fig. 4.43 Standardized Normal *pdf* 171

Fig. 4.44 Chi Square *pdf* 172

Fig. 4.45 Student’s t *pdf* 173

Fig. 4.46 F Distribution *pdf* 173

Fig. 4.47 Python Code for 3D Bar Chart 174

Fig. 5.1 Regression Model of Yogurt Purchases: Set-up 183

Fig. 5.2 Regression Model of Yogurt Purchases: Results 185

Fig. 5.3 Regression Display Parts 186

Fig. 5.4 OLS Prediction Method 188

Fig. 5.5 OLS Prediction Plot 189

Fig. 5.6 General Logistic Curve 190

Fig. 5.7 SFO Missing Value Report 193

Fig. 5.8 T2B Satisfaction Recoding 194

Fig. 5.9 Gender Distribution Before Recoding 195

Fig. 5.10 Gender Distribution After Recoding 195

Fig. 5.11 SFO Logit Model 196

Fig. 5.12 SFO Crosstab of Gender and Satisfaction 197

Fig. 5.13 Odds Ratio Calculation 199

Fig. 5.14 Odds Ratio Bar Chart 199

Fig. 5.15 Distribution of Yogurt Consumption per Week 202

Fig. 5.16 Poisson Regression Set-up 203

Fig. 5.17 Graphical Depiction of the ANOVA Decomposition 204

Fig. 6.1 Design Generation Set-up 212

Fig. 6.2 Design Matrix in a DataFrame 213

Fig. 6.3 Recoded Design Matrix in a DataFrame 213

Fig. 6.4 Example Conjoint Card 214

Fig. 6.5 Conjoint Estimation 216

Fig. 6.6 Retrieving Estimated Part-Worths 217

Fig. 6.7 Retrieving Estimated Part-Worths 218

Fig. 6.8 SFO Likelihood-to-Recommend Data 219

Fig. 6.9 Recoding of SFO Likelihood-to-Recommend Data 219

Fig. 6.10 NPS Decision Tree Setup 221

Fig. 6.11 NPS Decision Tree 222

Fig. 6.12 Satisfaction and Likelihood-to-Recommend Data Import 222

Fig. 6.13 Satisfaction and Likelihood-to-Recommend Data Recoding 223

Fig. 6.14 Satisfaction and Promoter McNemar Test 223

Fig. 6.15 Venn Diagram of Satisfied and Promoters 224

Fig. 6.16 Cross-tab of Brand by Segment for the Yogurt Survey 226

Fig. 6.17 CA Map of Brand by Segment for the Yogurt Survey 227

Fig. 6.18 CA Summary Table for the Yogurt Survey 228

Fig. 6.19 First Five Records of Toronto Casino Data 229

Fig. 6.20 Toronto Casino Data Missing Value Report 230

Fig. 6.21 Toronto Casino Data Removing White Spaces 231

Fig. 6.22 Toronto Casino Data Removing Punctuation Marks 232

Fig. 6.23 Toronto Casino Data Length Calculation 232

Fig. 6.24 Toronto Casino Data Length Histogram 233

Fig. 6.25 Toronto Casino Data Length Boxplots 233

Fig. 6.26 Toronto Casino Data Verbatim Wordcloud 234

Fig. 7.1 SRS Sample Size Calculation 241

Fig. 7.2 Stratified Sample Size Calculation 242

Fig. 7.3 VA Data Recoding 243

Fig. 7.4 VA Mean Age Calculation 243

Fig. 7.5 VA Mean Age Calculation with Strata 244

Fig. 7.6 Simple Tabulation of a Categorical Variable for Counts 245

Fig. 7.7 Simple Tabulation of a Categorical Variable for Proportions 246

Fig. 7.8 Simple Cross Tabulation of Two Categorical Variables 247

Fig. 7.9 One-Sample Test: Hypothesized Mean 248

Fig. 7.10 Two-Sample Test: Independent Populations 249

Fig. 8.1 Classical Confidence Interval Example 255

Fig. 8.2 Coin toss experiment 257

Fig. 8.3 Informative and Uninformative Priors 262

Fig. 8.4 Example Markov Chain 266

Fig. 8.5 Python Code to Generate a Random Walk 268

Fig. 8.6 Graph of a Random Walk 269

Fig. 8.7 Quantity Histogram 273

Fig. 8.8 Quantity Skewness Test 274

Fig. 8.9 Log Quantity Histogram 275

Fig. 8.10 Set-up to Estimate *OLS* Model 276

Fig. 8.11 Results for the Estimated *OLS* Model 277

Fig. 8.12 Regression using *pyMC3* 279

Fig. 8.13 Example of Skewed Distribution 281

Fig. 8.14 Example *MAP* Estimation 282

Fig. 8.15 Pooled Regression Summary from *pyMC3* Pooled Model 283

Fig. 8.16 Posterior Distribution Summary Charts 284

Fig. 8.17 Examples of Trace Plots 285

Fig. 8.18 Posterior Plots for the Regression Model 286

Fig. 8.19 Posterior Plot for logIncome for the Regression Model 287

Fig. 8.20 Posterior Plot Reference Line at 0 287

Fig. 8.21 Posterior Plot Reference Line at the Median 288

Fig. 8.22 Null Hypothesis and the *HDI* 289

Fig. 8.23 Set-up for Testing the Mean 290

Fig. 8.24 Trace Diagrams for Testing the Mean 291

Fig. 8.25 Posterior Distribution for Testing the Mean 291

Fig. 8.26 Set-up for Testing the Proportion 292

Fig. 8.27 Trace Diagrams for Testing the Proportion 292

Fig. 8.28 Posterior Distribution for Testing the Proportion 293

Fig. 8.29 Z-Test for the Voting Study 293

Fig. 8.30 Set-up for the MCMC Estimation for the Voting Problem 294

Fig. 8.31 MCMC Estimation Results for the Voting Problem 295

Fig. 8.32 Posterior Distributions for the Voting Problem 295

Fig. 8.33 Posterior Distribution for the Differences Between Parties
for the Voting Problem 296

Fig. 8.34 Logit Model for Voting Intentions: Frequentist Approach..... 297

Fig. 8.35 Set-up for Bayesian Logit Estimation 298

Fig. 8.36 Trace Diagrams for the Bayesian Logit Estimation 298

Fig. 8.37 Posterior Distribution for the Odds Ratio of the Bayesian Logit 299

Fig. 8.38 Political Party Odds Ratio Distribution for the Bayesian Logit 299

Fig. 8.39 Beta Distribution..... 301

Fig. 8.40 Normal and Half-Normal Distributions..... 302

Fig. 9.1 Changing Data Levels 309

Fig. 9.2 Multilevel Data Structure: Two Levels 311

Fig. 9.3 Connection of Levels to the Two Main Fallacies..... 312

Fig. 9.4 Pooled Regression with Generated Data..... 314

Fig. 9.5 Graph of Pooled Generated Data..... 315

Fig. 9.6 Pooled Regression with Dummy Variables..... 316

Fig. 9.7 Pooled Regression with Dummy Variables and Interactions 317

Fig. 9.8 Pooled Regression Summary..... 320

Fig. 9.9 Pooled Regression ANOVA Summary..... 321

Fig. 9.10 Pooled Regression with Dummy Variables Summary 322

Fig. 9.11 Set-up for Multilevel Model..... 326

Fig. 9.12 Estimation Results for Multilevel Model 327

Fig. 9.13 Distribution of Check-out Waiting Time by Store Location 329

Fig. 9.14 Relationship Between Check-out Waiting Time and Price 329

Fig. 9.15 Level 2 Regression Set-up..... 330

List of Tables

Table 1.1	Example of an Analysis Plan	15
Table 1.2	Examples of Quantities of Interest	18
Table 1.3	Illustration of Gender Dummy Variables	19
Table 1.4	Example of Population Parameter Calculations	22
Table 1.5	Python Package Categories	32
Table 2.1	Python Packages	37
Table 2.2	Data Dictionary for the VA Data.....	43
Table 2.3	pyreadstat's Returned Attributes	48
Table 2.4	Example Types of Categorical Survey Variables	56
Table 2.5	Pandas Summary Measures.....	62
Table 2.6	Standardization Methods for Response Bias	63
Table 2.7	Example 2 × 2 Table.....	66
Table 2.8	Pandas Data Types	68
Table 2.9	DataFrame Styling Options.....	70
Table 2.10	Population Distributions for Raking Example	74
Table 2.11	Sample Contingency Table for Raking Example	74
Table 3.1	Pandas Statistical Functions	89
Table 3.2	Crosstab Parameters	92
Table 3.3	Pivot_Table Parameters	93
Table 3.4	Matplotlib Annotation Commands	97
Table 3.5	Illustrative Questions for Pie Charts	98
Table 3.6	Pandas Plot Kinds	98
Table 3.7	Weighted Statistics Options	111
Table 4.1	1-Way Table Layout for Service Branches.....	134
Table 4.2	Examples of Effects Coding	136
Table 4.3	General ANOVA Table Structure.....	138
Table 4.4	Stylized Cross-Tab for McNemar Test	146
Table 4.5	CATA Attribution of Response Differences	150
Table 5.1	Link Functions	179
Table 5.2	The General Structure of an ANOVA Table	206
Table 6.1	Parameters Needed for the Watch Case Study	212

Table 8.1 Example Voting Intention Table 257
Table 8.2 Example Prior Distributions 263
Table 8.3 Side Effects by Store Size 271
Table 9.1 Omitted Variable Possibilities 307
Table 9.2 The Problem Cell 307

Chapter 1

Introduction to Modern Survey Analytics



Contents

1.1	Information and Survey Data	3
1.2	Demystifying Surveys.....	4
1.2.1	Survey Objectives	5
1.2.2	Target Audience and Sample Size.....	7
1.2.3	Screening and Questionnaire Design	12
1.2.4	Fielding the Study	14
1.2.5	Data Analysis	14
1.2.6	Report Writing and Presentation	16
1.3	Sample Representativeness	16
1.3.1	Digression on Indicator Variables	20
1.3.2	Calculating the Population Parameters	21
1.4	Estimating Population Parameters.....	22
1.5	Case Studies	25
1.5.1	Consumer Study: Yogurt Consumption	25
1.5.2	Public Sector Study: VA Benefits Survey	27
1.5.3	Public Opinion Study: Toronto Casino Opinion Survey	28
1.5.4	Public Opinion Study: San Francisco Airport Customer Satisfaction Survey	30
1.6	Why Use Python for Survey Data Analysis?	30
1.7	Why Use Jupyter for Survey Data Analysis?	32

There are two things, it is often said, that you cannot escape: death and taxes. This is too narrow because there is a third: surveys. You are inundated daily by surveys of all kinds that cover both the private and public spheres of your life. In the private sphere, there are product surveys designed to learn what you buy, use, have, would like to have, and uncover what you believe is right and wrong about existing products. They are also used to determine the optimal marketing mix that consists of the right product, placement, promotion, and pricing combination to effectively sell products. They are further used to segment the market recognizing that one marketing mix does not equally apply to all customers. There are surveys used to gauge how well the producers of these products perform in all aspects of making, selling, and supporting their products. And there are surveys internal to

those producers to help business managers determine if their employees are happy with their jobs and if they have any ideas for making processes more efficient or have suggestions and advice regarding new reorganization efforts and management changes.

In the public sphere, there are political surveys—the “polls”—reported daily in the press that tell us how the public views a “hot” issue for an upcoming election, an initiative with public implications, and a policy change that should be undertaken. There are surveys to inform agencies about who is using the public services they offer, why those services are used, how often they are used, and even if the services are known.

Some of these surveys are onetime events meant to provide information and insight for an immediate purpose. They would not be repeated because once conducted, they would have completed their purpose. A private survey to segment the market is done once (or maybe once every, say, 5 years) since the entire business organization is structured around the marketing segments. This includes business unit structure, lines of control and communication, and business or corporate identity. In addition, marketing campaigns are tailored for these segments. All this is founded on surveys.

Other surveys are routinely conducted to keep track of market developments, views, and opinions that require unexpected organizational changes. These are tracking studies meant to show trends in key measures. Any one survey in a tracking study is insightful, but it is the collection over time that is more insightful and the actual reason tracking is done in the first place.

Counting how many surveys are conducted annually, whether onetime or tracking, is next to impossible because many are proprietary. Businesses normally do not reveal their intelligence gathering efforts because doing so then reveals what their management is thinking or concerned about; this is valuable competitive intelligence for its competition.

A further distinction has to be made between a survey *per se* and the number of people who answer the survey. The latter is the number of completions. The online survey provider SurveyMonkey claims that they alone handle 3 million completions a day. That translates to over a billion completes a year!¹ And that is just for one provider. The US government conducts a large number of regular surveys that are used to measure the health of the economy and build valuable data sets for policy makers and business leaders. The US Census Bureau, for example, notes that it “conducts more than 130 surveys each year, including our nation’s largest business survey,” the *Annual Retail Trade (ARTS)*.²

¹ See <https://www.quora.com/How-many-surveys-are-conducted-each-year-in-the-US#XettS>. Last accessed March 29, 20120.

² See the US Census Bureau’s website at <https://www.census.gov/programs-surveys/surveyhelp/list-of-surveys/business-surveys.html>. Last accessed March 29, 2020.

1.1 Information and Survey Data

The reason surveys are an integral part of modern life is simple: They provide information for decision-making, and decisions in modern, high-tech, and interconnected societies are more complex than in previous periods. There is so much more happening in our society than even 20 years ago at the turn of the millennium. We now have sensors in almost all major appliances, in our cars, and at our street corners; we have social media that has created an entangled network where everyone is connected; we have full-fledged computers in our pockets and purses with more power than the best mainframes of 20 years ago;³ and we have the Internet with all its power, drawbacks, and potential benefits as well as dangers. Technology is changing very rapidly following *Moore's Law*: “the observation that the number of transistors in a dense integrated circuit (IC) doubles about every 2 years.”⁴

As a result of this rapid and dynamic change in technology, there has been an equally rapid and dynamic change in our social structure, including what we believe, how we work and are organized, how we relate to each other, how we shop, and what we buy. Decision-makers in the private and public spheres of our society and economy must make decisions regarding what to offer, in terms of products and programs, recognizing that whatever they decide to do may, and probably will, change significantly soon after they make that decision.

To keep pace with this rapidly changing world, they need information, penetrating insight, into what people want, what they believe, how they behave, and how that behavior has and will change. They can certainly get this from databases, so-called Big Data, but this type of data is, by their collection nature, historical. They reflect what did happen, not what will happen or where the world is headed. The only way to gain this information is by asking people about their beliefs, behaviors, intentions, and so on. This is where surveys are important. They are the vehicle, the source of information, for providing decision-makers with information about what drives or motivates people in a rapidly changing world.

Since possibilities are now so much greater, the speed of deployment and coverage of surveys have also become equally greater issues. More surveys have to be conducted more frequently and in more depth to provide information to key decision-makers. These surveys result in an overload, not of information but of data because each one produces a lot of data in a very complex form. The data have to be processed, that is, analyzed, to extract the needed information. Data and information are not the same. Information is hidden inside data; it is latent, needing to be extracted. This extraction is not easy; in fact, it is quite onerous. This is certainly not a problem unique to survey data. The Big Data I referred to above has this same problem, a problem most likely just as large and onerous to solve as for surveys.

³ A McKinsey report in 2012 stated “More and more smartphones are as capable as the computers of yesteryear.” See Bauer et al. (2012).

⁴ See https://en.wikipedia.org/wiki/Moore%27s_law. Last accessed January 8, 2021.

Regardless of their source, data must be analyzed to extract the latent information buried inside them. Extraction methodologies could be shallow or deep. *Shallow Data Analysis* just skims the surface of the data and extracts minimal useful information. The methodologies tend to be simplistic, such as 2×2 tables, volumes of crosstabs (i.e., the “tabs”), and pie and bar charts, which are just one-dimensional views of one, maybe two, variables. More insightful, penetrating information is left latent, untapped. *Deep Data Analysis* digs deeper into the data, searching out relationships and associations across multiple variables and subsets of the data. The methodologies include, but are certainly not limited to, perceptual maps, regression analysis,⁵ multivariate statistical tests, and scientific data visualization beyond simple pie and bar charts, to mention a few. This book’s focus is Deep Data Analysis for extracting actionable, insightful, and useful information latent in survey data. See Paczkowski (2022) about the connection between data and information and the importance of, and discussions about, information extraction methodologies.

1.2 Demystifying Surveys

It helps to clarify exactly what is a survey. This may seem odd considering that so many are conducted each year and that you are either responsible for one in your organization, receive reports or summaries of surveys, or have been asked to take part in a survey. So the chances are you had some contact with a survey that might lead you to believe you know what they are. Many people confuse a survey with something associated with it: the questionnaire. They often treat these terms—survey and questionnaire—as synonymous and interchangeable. Even professionals responsible for all aspects of a survey do this. But they are different.

A survey is a process that consists of six hierarchically linked parts:

1. Objective statement
2. Target audience identification
3. Questionnaire development and testing
4. Fielding of the survey
5. Data analysis
6. Results reporting

You can think of these collectively as a *survey design*. This sequence, of course, is only theoretical since many outside forces intervene in an actual application. I show an example survey design in Fig. 1.1 that highlights these six components. A careful study of them suggests that they could be further grouped into three overarching categories:

1. Planning
2. Execution
3. Analysis

⁵ Regression analysis includes a family of methods. See Paczkowski (2022) for a discussion.

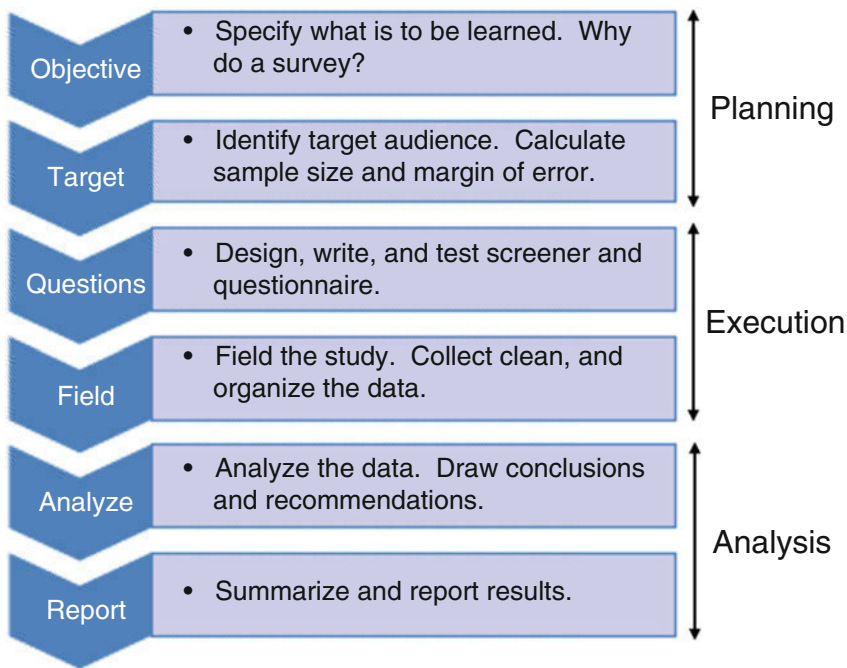


Fig. 1.1 This illustrates a typical survey process. Although the process is shown as a linear one, it certainly could be nonlinear (e.g., flowing back to a previous step to redo something) as well as iterative

which I also show in Fig. 1.1. The report stage is part of the analysis category because writing a report is often itself an analytical process. Writing, in general, is a creative process during which questions are asked that were not previously thought about but that become obvious and, therefore, which need to be addressed (and hopefully the data are available to answer them). I will expound on these six components in the next six subsections.

1.2.1 Survey Objectives

The possible objectives for a survey are enormous to say the least. They can, however, be bucketed into four categories:

1. Fact finding
2. Trend analysis
3. Pattern identification
4. Intentions

Fact finding runs the gamut from current viewpoints to behavioral habits to awareness to familiarity. Current viewpoints include beliefs and opinions such as

satisfaction, political affiliation, and socioeconomic judgments. Current behavioral habits include, as examples: where someone currently shops, the amount purchased on the last shopping visit, the frequency of employment changes, the services either currently or previously used, organizational or professional affiliations, the number of patients seen in a typical week, voted in the last election, proportion of patients in a medical practice who receive a particular medication, and so on. Demographic questions are included in this category because they are facts that aid profiling respondents and identifying more facts about behaviors and items from the other categories. For example, the gender of respondents is a fact that can be used to subdivide professional affiliation.

Also in this behavioral habits category are questions about attitudes, interests, and opinions (*AIOs*).⁶ These questions are usually simple binary (or trinary) questions or Likert Scale questions. Binary questions require a *Yes/No* answer, while trinary ones require a *Yes/No/Maybe* or *Yes/No/Don't Know* response. Likert Scale questions typically (but not always) have five points spanning a negative to positive sentiment. Common examples are *Disagree-Agree* and *Dislike-Like*. See Vyncke (2002) for some discussion of these types of questions in communications research.

Awareness and familiarity are sometimes confused, but they differ. Awareness is just knowing that something exists, but the level of knowledge or experience with that item is flimsy at best. For example, someone could be aware of home medical services offered by Medicare but has never talked to a Medicare representative about care services, read any literature about them, and never spoke to a healthcare provider about what could be useful for his/her situation. Familiarity, on the other hand, is deeper knowledge or experience. The depth is not an issue in most instances because this is probably difficult to assess. Nonetheless, the knowledge level is more extensive so someone could reasonably comment about the item or issue. For the Medicare example, someone might be familiar with home healthcare provisioning for an elderly parent after talking to a Medicare representative or elder-care provider such as an attorney, medical practitioner, and assisted living coordinator.

Trend analysis shows how facts are changing over time. This could take two forms: within a survey or between surveys. The first is based on a series of questions about, for example, amount purchased in each of the last few weeks or how much was spent on, say, jewelry the previous year's holiday season and the recent holiday season. This type of tracking is useful for determining how survey respondents have changed over time and if that change has larger implications for the organization sponsoring the survey (i.e., the client). The between-surveys analysis involves asking the same fact-type questions each time a survey is conducted and then analyzing how the responses have changed over time. These are used to identify new trends useful for the organization or help spot problems that need attention. For example, hospitals track patient satisfaction for various parts of the hospital (e.g., emergency room, front desk assistance, and patient care) on, say, a monthly basis and post the monthly mean satisfaction scores for staff and patients to see.

⁶ *AIO* is also used for *Activities/Interests/Opinions*. See Vyncke (2002) for this use.