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

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 *WSN*s 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