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

List of Figures

List of Tables

Chapter 1
Introduction to Modern Survey Analytics

Contents

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

W. R. Paczkowski, *Modern Survey Analysis*,
https://doi.org/10.1007/978-3-030-76267-4_1

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.

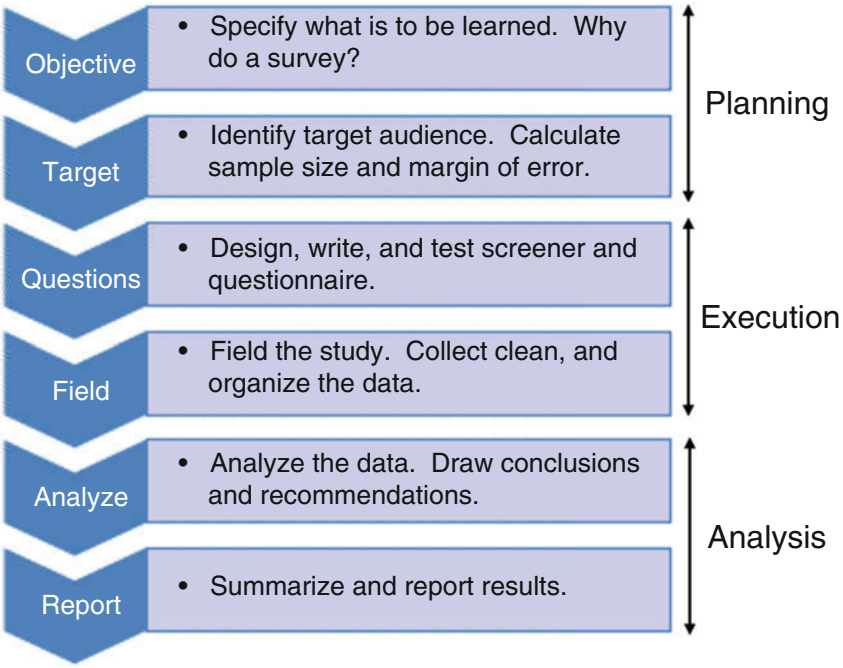

Fig. 1.1 This illustrates a typical survey process. Although the process is shown as a linear one, it certainly could be nonlinear (e.g., flowing back to a previous step to redo something) as well as iterative

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 (*AIO*s).[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.