THE SCIENCE BEHIND THIS BOOK

Every day, our news feeds are full of strategies designed to make our lives easier, make us happier, and help us take over the world. We also hear stories about how teams and organizations use different strategies to transform their technology and win in the market. But how are we to know which actions we take just *happen* to correspond to the changes we observe in our environment and which actions are driving these changes? This is where rigorous primary research comes into play. But what do we mean by "rigorous" and "primary"?

PRIMARY AND SECONDARY RESEARCH

Research falls into two broad classes: primary and secondary research. The key difference between these two types is who collects the data. Secondary research utilizes data that was collected by someone else. Examples of secondary research you are probably familiar with are book reports or research reports we all completed in school or university: we collected existing information, summarized it, and (hopefully) added in our own

insights about what was found. Common examples of this also include case studies and some market research reports. Secondary research reports can be valuable, particularly if the existing data is difficult to find, the summary is particularly insightful, or the reports are delivered at regular intervals. Secondary research is generally faster and less expensive to conduct, but the data may not be well suited to the research team because they are bound by whatever data already exists.

In contrast, primary research involves collecting new data by the research team. An example of primary research includes the United States Census. The research team collects new data every ten years to report on demographic and population statistics for the country. Primary research is valuable because it can report information that is not already known and provide insights that are not available in existing datasets. Primary research gives researchers more power and control over the questions they can address, though it is generally more costly and time intensive to conduct. This book and the State of DevOps Reports are based on primary research.

QUALITATIVE AND QUANTITATIVE RESEARCH

Research can be qualitative or quantitative. Qualitative research is any kind of research whose data isn't in numerical form. This can include interviews, blog posts, Twitter posts, long-form log data, and long-form observations from ethnographers. Many people assume that survey research is qualitative because it doesn't come

from computer systems, but that isn't necessarily true; it depends on the kinds of questions asked in the survey. Qualitative data is very descriptive and can allow for more insights and emergent behavior to be discovered by researchers, particularly in complex or new areas. However, it is often more difficult and costly to analyze; efforts to analyze qualitative data using automated means often codify the data into a numerical format, making it quantitative.

Quantitative research is any kind of research with data that includes numbers. These can include system data (in numerical format) or stock data. System data is any data generated from our tools; one example is log data. It can also include survey data, if the survey asks questions that capture responses in numerical format—preferably on a scale. The research presented in this book is quantitative, because it was collected using a Likert-type survey instrument.

What Is a Likert-Type Scale?

A Likert-type scale records responses and assigns them a number value. For example, "Strongly disagree" would be given a value of 1, neutral a value of 4, and "Strongly agree" a value of 7. This provides a consistent approach to measurement across all research subjects, and provides a numerical base for researchers to use in their analysis.

TYPES OF ANALYSIS

Quantitative research allows us to do statistical data analysis. According to a framework presented by Dr. Jeffrey Leek at Johns Hopkins Bloomberg School of Public Health (Leek 2013), there are six types of data analysis (given below in the order of increasing complexity). This complexity is due to the knowledge required by the data scientist, the costs involved in the analysis, and the time required to perform the analysis. These levels of analysis are:

1. Descriptive
2. Exploratory
3. Inferential predictive
4. Predictive
5. Causal
6. Mechanistic

The analyses presented in this book fall into the first three categories of Dr. Leek's framework. We also describe an additional type of analysis, classification, which doesn't fit cleanly into the above framework.

DESCRIPTIVE ANALYSIS

Descriptive analysis is used in census reports. The data is summarized and reported—that is, described. This type of analysis takes the least amount of effort, and is often done as the first step of data analysis to help the research team understand their dataset (and, by extension, their sample and possibly population of users). In some cases, a report will stop at descriptive analysis, as in the case of population census reports.

What Is a Population and Sample, and Why Are They Important?

When talking about statistics and data analysis, the terms "population" and "sample" have special meanings. The *population* is the entire group of something you are interested in researching; this might be all of the people undergoing technology transformations, everyone who is a Site Reliability Engineer at an organization, or even every line in a log file during a certain time period. A *sample* is a portion of that population that is carefully defined and selected. The sample is the dataset on which researchers perform their analyses. Sampling is used when the entire population is too big or not easily accessible for research. Careful and appropriate sampling methods are important to make sure the conclusions drawn from analyzing the sample are true for the population.

The most common example of descriptive analysis is the government census where population statistics are summarized and reported. Other examples include most vendor and analyst reports that collect data and report summary and aggregate statistics about the state of tool usage in an industry or the level of education and certification among technology professionals. The percentage of firms that have started their Agile or DevOps journeys as reported by Forrester (Klavens et al. 2017), the IDC report on average downtime cost (Elliot 2014), and the O'Reilly Data Science Salary Survey (King and Magoulas 2016) belong in this category.

These reports are very useful as a gauge of the current state of the industry, where reference groups (such as populations or industries) currently are, where they once were, and where the trends are pointing. However, descriptive findings are only as good as the underlying research design and data collection methods. Any reports that aim to represent the underlying population must be sure to sample that population carefully and discuss any limitations. A discussion of these considerations is beyond the scope of this book.

An example of descriptive analysis found in this book is the demographic information about our survey participants and the organizations they work in—what countries they come from, how large their organizations are, the industry vertical they work in, their job titles, and their gender (see Chapter 10).

EXPLORATORY ANALYSIS

Exploratory analysis is the next level of statistical analysis. This is a broad categorization that looks for relationships among the data and may include visualizations to identify patterns in the data. Outliers may also be detected in this step, though the researchers have to be careful to make sure that outliers are, in fact, outliers, and not legitimate members of the group.

Exploratory analyses are a fun and exciting part of the research process. For those who are divergent thinkers, this is often the stage where new ideas, new hypotheses, and new research projects are generated and proposed. Here, we discover how the variables in our data are related and we look for possible new connections

and relationships. However, this should not be the end for a team that wants to make statements about prediction or causation.

Many people have heard the phrase "correlation doesn't imply causation," but what does that mean? The analyses done in the exploratory stage include correlation but not causation. Correlation looks at how closely two variables move together—or don't—but it doesn't tell us if one variable's movement predicts or causes the movement in another variable. Correlation analysis only tells us if two variables move in tandem or in opposition; it doesn't tell us why or what is causing it. Two variables moving together can always be due to a third variable or, sometimes, just chance.

A fantastic and fun set of examples that highlight high correlations due to chance can be found at the website Spurious Correlations.[1] The author Tyler Vigen has calculated examples of highly correlated variables that common sense tells us are not predictive and certainly not causal. For example, he shows (Figure 12.1) that the per capita cheese consumption is highly correlated with the number of people who died by becoming tangled in their bedsheets (with a correlation of 94.71% or $r = 0.9471$; see footnote 2 on correlations in this chapter). Surely cheese consumption doesn't cause strangulation by bedsheets. (And if it does—what kind of cheese?) It would be just as difficult to imagine strangulation by bedsheets causing cheese consumption —unless that is the food of choice at funerals and wakes around the country. (And again: What kind of cheese? That is a morbid marketing opportunity.) And yet, when we go "fishing in the data," our minds fill in the story because our datasets are related and so

often make sense. This is why it is so important to remember that correlation is only the exploratory stage: we can report correlations, and then we move on to more complex analyses.

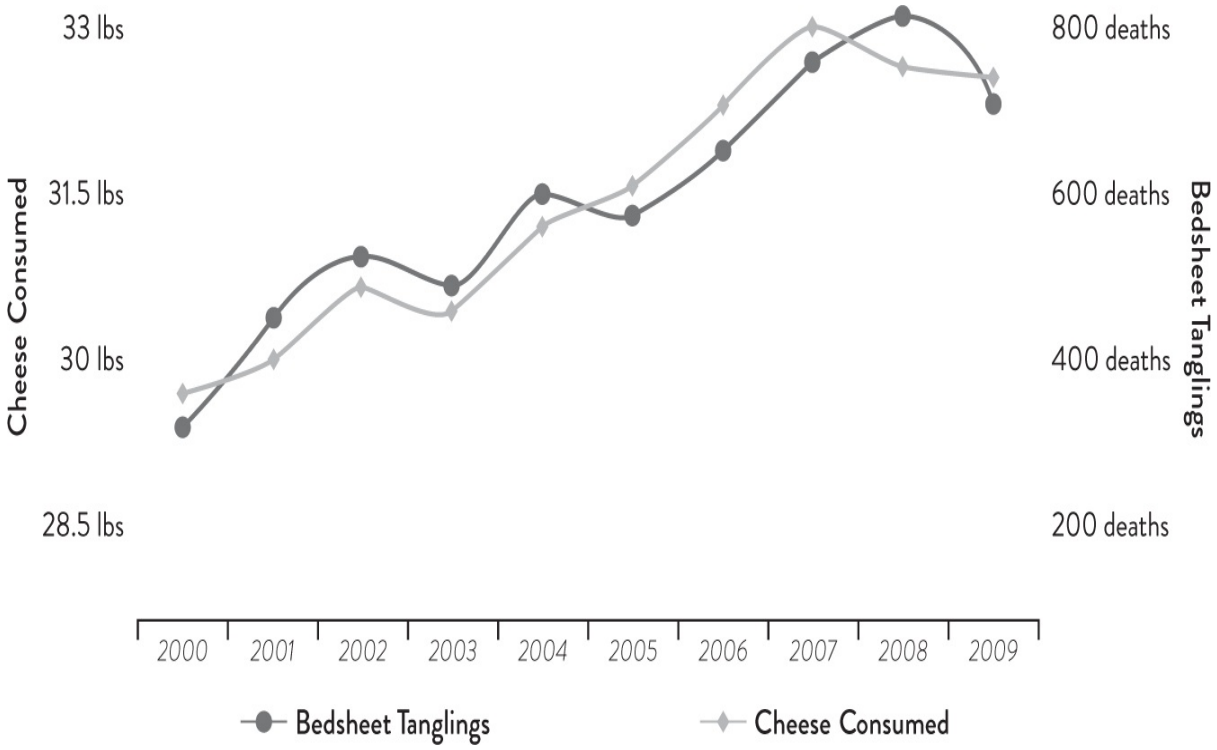

Figure 12.1: Spurious Correlation: Per Capita Cheese Consumption and Strangulation by Bedsheets

There are several examples of correlations that are reported in our research and in this book, because we know the importance and value of understanding how things in our environment interrelate. In all cases, we reported Pearson correlations,[2] which is the correlation type most often used in business contexts today.

INFERENTIAL PREDICTIVE ANALYSIS

The third level of analysis, inferential, is one of the most common types conducted in business and technology research today. It is

also called inferential predictive, and it helps us understand impacts of HR policies, organizational behavior and motivation, and how technology impacts outcomes like user satisfaction, team efficiency, and organizational performance. Inferential design is used when purely experimental design is not possible and field experiments are preferred—for example, in business, when data collection happens in complex organizations, not in sterile lab environments, and companies won't sacrifice profits to fit into control groups defined by the research team.

To avoid problems with "fishing for data" and finding spurious correlations, hypotheses are theory driven. This type of analysis is the first step in the scientific method. Many of us are familiar with the scientific method: state a hypothesis and then test it. In this level of analysis, the hypothesis must be based on a well-developed and well-supported theory.

Whenever we talk about *impacting* or *driving* results in this book, our research design utilized this third type of analysis. While some suggest that using a theory-based design opens us up to confirmation bias, this is how science is done. Well, wait—almost. Science isn't done by simply confirming what the research team is looking for. Science *is* done by stating hypotheses, designing research to test those hypotheses, collecting data, and then testing the stated hypotheses. The more evidence we find to support a hypothesis, the more confidence we have for it. This process also helps to avoid the dangers that come from fishing for data—finding the spurious correlations that might randomly exist but have no real reason or explanation beyond chance.

Examples of hypotheses tested with inferential analysis in our

project include continuous delivery and architecture practices driving software delivery performance, software delivery positively affecting organizational performance, and organizational culture having a positive impact on both software delivery and organizational performance. In these cases, the statistical methods used were either multiple linear regression or partial least squares regression. These methods are described in more detail in Appendix C.

PREDICTIVE, CAUSAL, AND MECHANISTIC ANALYSIS

The final levels of analysis were not included in our research, because we did not have the data necessary for this kind of work. We will briefly summarize them here for the sake of completeness and to appease your curiosity.

- Predictive analysis is used to predict, or forecast, future events based on previous events. Common examples include cost or utilities predictions in business. Prediction is very hard, particularly as you try to look farther away into the future. This analysis generally requires historical data.
- Causal analysis is considered the gold standard, but is more difficult than predictive analysis and is the most difficult analysis to conduct for most business and technology situations. This type of analysis generally requires randomized studies. A common type of casual analysis done in business is A/B testing in prototyping or websites,

when randomized data can be collected and analyzed.

- Mechanistic analysis requires the most effort of all methods and is rarely seen in business. In this analysis, practitioners calculate the exact changes to make to variables to cause *exact* behaviors that will be observed under certain conditions. This is seen most often in the physical sciences or in engineering, and is not suitable for complex systems.

CLASSIFICATION ANALYSIS

Another type of analysis is classification, or clustering, analysis. Depending on the context, research design, and the analysis methods used, classification may be considered an exploratory, predictive, or even causal analysis. We use classification in this book when we talk about our high-, medium-, and low-performance software delivery teams. This may be familiar to you in other contexts when you hear about customer profiles or market basket analysis. At a high level, the process works like this: classification variables are entered into the clustering algorithm and significant groups are identified.

In our research, we applied this statistical method using the tempo and stability variables to help us understand and identify if there were differences in how teams were developing and delivering software, and what those differences looked like. Here is what we did: we put our four technology performance variables — deployment frequency, lead time for changes, mean time to repair, and change fail rate—into the clustering algorithm, and

looked to see what groups emerged. We see distinct, statistically significant differences, where high performers do significantly better on all four measures, low performers perform significantly worse on all four measures, and medium performers are significantly better than low performers but significantly worse than high performers. For more detail, see Chapter 2.

What Is Clustering?

For those armchair (or professional) statisticians who are interested, we used hierarchical clustering. We chose this over k-means clustering for a few reasons. First, we didn't have any theoretical or other ideas about how many groups to expect prior to the analysis. Second, hierarchical clustering allowed us to investigate parent-child relationships in the emerging clusters, giving us greater interpretability. Finally, we didn't have a huge dataset, so computational power and speed wasn't a concern.

THE RESEARCH IN THIS BOOK

The research presented in this book covers a four-year time period, and was conducted by the authors. Because it is primary research, it is uniquely suited to address the research questions we had in mind—specifically, what capabilities drive software delivery performance and organizational performance? This project was based on quantitative survey data, allowing us to do statistical

analyses to test our hypotheses and uncover insights into the factors that drive software delivery performance.

In the next chapters, we discuss the steps we took to ensure the data we collected from our surveys was good and reliable. Then, we look into why surveys may be a preferred source of data for measurement—both in a research project like ours and in your own systems.

---

[1] http://www.tylervigen.com/spurious-correlations.

[2] Pearson correlations measure the strength of a linear relationship between two variables, called Pearson's r. It is often referred to as just correlation and takes a value between -1 and 1. If two variables have a perfect linear correlation, that is they move together exactly, $r = 1$. If they move in exactly opposite directions, $r = -1$. If they are not correlated at all, $r = 0$.