Also by Steven Ouellette

Business Performance Excellence (with Dr. Jeffrey Luftig)
Galileo's Telescope

Data-Driven Leadership

Data-Driven Leadership

Empowering Managers with Statistical Insights

Steven Ouellette

Copyright 2023 by Steven Ouellette

Published by Alliance Press

Cover photo by Polina Tankilevitch

All rights reserved. Any unauthorized reprint or use of this material is prohibited. No part of this book may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or by any information storage and retrieval system without the express written permission from the author.

Disclaimer: While all attempts have bene made to verify the accuracy of the information provided in this publication, the author assumes no responsibility for any errors or omissions. Should the reader wish to implement the tools and techniques described herein, it is recommended that they contact a professional consultant with expertise in this area. The author will not be held responsible for attempting to implement this approach without appropriate support.

ISBN: 9798867257507

Table of Contents

INTRODUCTION, OR "WHY DO WE NEED ANOTHER BOOK ON STATISTICS?" ..... **I**

CHAPTER 1 – WHY DO YOU NEED STATISTICS? **1**

What is Data? 1
What is a Statistic? 2
Population vs. Sample 3
Random Sampling 3
Statistics and Parameters 5
Sampling Error 6
The Chain of Inference 6
Summary 8
Asking the Right Questions: Chapter 1 8

CHAPTER 2 – MEASUREMENT LEVELS **9**

Determining Measurement Level 18
Allowed Statistical Functions for Measurement Levels 21
Summary 23
Asking the Right Questions: Chapter 2 23
Answers to Exercise 24

CHAPTER 3 – UNDERSTANDING AND PRESENTING DATA **25**

Changes in the Data Through Time 27
The Shape of the Data 31
The Spread of the Data 34
The Location of the Data 36
Visually Showing the Shape, Spread, and Location of Data 39
What Happens if You Don't Know All Four Aspects? 46
A Statistical Mantra 50
Summary 50
Asking the Right Questions: Chapter 3 50

CHAPTER 4 – PROBABILITY DISTRIBUTIONS **51**

The Binomial Distribution 51
The Poisson Distribution 55
The Normal Distribution 60

THE EXPONENTIAL DISTRIBUTION 65
OTHER DISTRIBUTIONS 68
SUMMARY 69
ASKING THE RIGHT QUESTIONS: CHAPTER 4 69
ANSWERS TO EXERCISES 69

CHAPTER 5 – DISTRIBUTION TESTING 71

TESTING FOR BINOMIAL DISTRIBUTIONS 72
TESTING FOR POISSON DISTRIBUTIONS 72
TESTING FOR DISTRIBUTION NORMALITY 73
TESTING FOR EXPONENTIAL DISTRIBUTION 77
DISTRIBUTION TESTING THOUGHTS 79
SUMMARY 80
ASKING THE RIGHT QUESTIONS: CHAPTER 5 80
ANSWERS TO EXERCISES 80

CHAPTER 6 – ESTIMATION 81

SAMPLING DISTRIBUTIONS AND ESTIMATION 81
THE CENTRAL LIMIT THEOREM 84
ESTIMATING POPULATION PARAMETERS FROM SAMPLE DATA 91
APPLYING THE CONFIDENCE INTERVAL: BEST-CASE / WORST-CASE ANALYSIS 98
SUMMARY 105
ASKING THE RIGHT QUESTIONS: CHAPTER 6 105
ANSWERS TO EXERCISES 105

CHAPTER 7 – FOUNDATIONS OF HYPOTHESIS TESTING 107

AVOIDING "INDUSTRIAL MYTHOLOGY" 107
MAKING A DECISION WITH DATA 109
SUMMARY 121
ASKING THE RIGHT QUESTIONS CHAPTER 7 121

CHAPTER 8 – SAMPLE SIZE 122

EFFECT SIZE 124
CALCULATING POWER 125
CALCULATING SAMPLE SIZE 129
THE IMPORTANCE OF CALCULATING SAMPLE SIZES 130
RETURN ON INVESTMENT AND SAMPLE SIZES 132

Summary 138
Asking the Right Questions Chapter 8 138
Answers to Exercises 139

CHAPTER 9 – GUIDING YOU THROUGH HYPOTHESIS TESTING. 143

Testing a Business Question with Data 143
Summary 150
Asking the Right Questions, Chapter 9 150

CHAPTER 10 – ONE-SAMPLE HYPOTHESIS TESTS 151

Selecting the Right One-Sample Test 152
One-Sample Tests for the Average 152
One-Sample x^2 Test for Variance 162
One-Sample Exact Binomial 164
One-Sample Sign Test for Location 166
One-Sample Poisson Exact Test 168
Summary 169
Asking the Right Questions Chapter 10 170
Answers to Exercises 170

CHAPTER 11 – CORRELATION, REGRESSION, AND ASSOCIATION 172

Choosing the Right Measure 172
Correlation of Ratio or Interval Data 173
Introduction to Linear Regression 184
Correlation of Ordinal Data, Spearman's r_S 188
Other Measures of Association 191
Summary 195
Asking the Right Questions Chapter 11 195
Answers to Exercises 195

CHAPTER 12 – TWO-SAMPLE HYPOTHESIS TESTS 198

Independent Versus Dependent Tests 199
Independent Tests 199
Dependent Tests 213
Summary 237
Asking the Right Questions Chapter 12 238
Answers to Exercises 238

CHAPTER 13 – ANALYSIS OF VARIANCE 250

WHY WOULDN'T YOU JUST DO A LOT OF *T*-TESTS? 251
ONE-WAY FIXED FACTOR ANOVA 252
ONE-WAY RANDOM EFFECTS ANOVA 261
ADDITIONAL ANOVA TOPICS 266
SUMMARY 274
ASKING THE RIGHT QUESTIONS CHAPTER 13 274
ANSWERS TO EXERCISES 274

CHAPTER 14 – CLOSING THOUGHTS 280

BIBLIOGRAPHY 282

USEFUL EQUATIONS 284

COMMON STATISTICAL SYMBOLS AND ABBREVIATIONS 287

GLOSSARY 288

ACKNOWLEDGEMENTS 299

ABOUT THE AUTHOR 300

Introduction, or "Why do we need another book on statistics?"

Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write!

(Wilks, 1951) Paraphrasing H.G. Wells

I remember the first time I took a statistics class in college. I was an undergraduate engineer and had heard it was an easy A.

Well, it was, and it wasn't.

It was an easy A in the sense that I could memorize what to do and use it on tests like a moderately trained monkey, but I really didn't see how knowing the probability of drawing an ace from a deck of cards was going to help me in my career. I had statistical tools, but no real understanding of what the point of statistics was.

Flash forward a couple of years. I was working at Alcoa and we had just made a few lots of a new product. A customer called me up and wanted to know if we could narrow the specifications on that product. We had only made ten lots, and none of them failed the new specs. Case closed, right?

On a whim, I dredged up my old statistics knowledge and estimated how much we would scrap with the new specs if the ten samples we had were representative of what future production would be.

It was about 10%.

That would be a *very* expensive amount of metal to scrap. And yeah, we took the business anyway. And it was a disaster.

THAT is when I understood the purpose of statistics, and why you should read this book.

In the real world (and that includes business), we are often required to make decisions based on incomplete knowledge. Statistical techniques give us a shot at making better decisions with data. They allow us to do that even when we don't have all the data we would like to have.

STATISTICS DOESN'T HAVE TO BE HARD TO UNDERSTAND

Statistical techniques help make better business decisions, but many people tell me that they "never got" statistics.

I believe statistics is the most practical math you can learn. In my ideal world, the terminal math class for high school would be statistics, rather than calculus. As much as I love calculus, statistical thinking has far more use for the vast majority of people, including me.

The hard thing about statistics is not the math itself; we are only going to be using some pretty basic algebra. The difficult part is learning a new way of thinking.

We are taught that $2 + 2 = 4$ always, and that $2 < 3 < 4$ always. But in the real world we need to make decisions without such certainty. Statistics help limit risk when we don't know everything. In acknowledging that we don't know everything we leave room for uncertainty. And sometimes that means the seemingly nonsensical $a = b$ and $b = c$ but $a \neq c$ is true.

This accounting for uncertainty reflects real life. Every day when we are walking to the bus, talking to others, or making business decisions, we have to humbly accept that while we don't know everything about the world around us, we do know *something*, and we can at least use what we know to make better decisions.

My aim here is to write an introductory book that will help people in business understand, and more importantly use, a tool that will make their decisions and their lives better. In each chapter that follows you will not only learn statistical concepts, but also the questions you should be asking to make the best management decisions.

I WON'T BE DOING ANY ANALYSES! WHY DO I NEED TO LEARN ANY OF THIS?

As a manager or business leader, you will often be presented with reports that contain data analyses. If you don't know the elements of data collection and analysis, you are a hostage to whatever your people put in the report. Learning this material gives you important questions to ask so that you can judge the usefulness of the recommendations to make management decisions.

WHAT ABOUT BIG DATA?

Most of the statistical techniques in this book will be applicable to "little data," or data that we specifically have to go out and gather and is often expensive to come by. These techniques will extract the most information from the smallest samples possible.

A lot of people will say, "My company has gone through 'The Digital Transformation' so we don't need to learn any of this old stuff!"

The quality of digital transformations prior to my arrival at clients is…not so great. I have seen:

- The decisions that need to be made in the business are often not considered when collecting data and moving it to where it would be useful. Data that managers need is often not available and requires special action to collect anyway. You are back to "little data."
- The data from the front line is rolled up to higher levels of management, who should have no role in making decisions for that level. This leads to micromanagement.
- The data are presented in graphics that don't differentiate normal variation from something that managers need to react to, or worse show arbitrary boundaries and danger colors based on business needs without regard to the processes that they measure. This leads to over- or under-reaction by managers.
- The number of metrics presented to management are overwhelming, often outside their control, and muddy the waters, making it more difficult to even know what decisions need to be made.
- Metrics are not translated from business to process and vice versa, leading to confusion at all levels.
- When working on a crisis or when trying to improve a process, you will probably have to gather and make decisions on "little data" anyway.

The tools in this book will help with some of these points. You will learn about properly collecting "little data" and how to analyze it to make better management decisions.

For the larger systemic issues such as creating business success metrics, properly translating them for each manager, and using these metrics to make a strategic plan and manage on a daily basis, see my book *Galileo's Telescope* (Ouellette, 2021).

A Note on the Structure of the Book

Each chapter will cover the topic, most will have exercises so that you can practice the concepts, and all of them will have a summary of the topics and questions you will want to get answers to as you are presented with analyses before you make any decisions. This structure is to help you to achieve statistical literacy in order to leverage data into better business decisions.

A Note on Software

I am specifically writing this book so that it doesn't matter what software you have available to you. There won't be screenshots of a particular package that will be outdated with the next update, and I won't be telling you how to make your software do what we are talking about in this text. Companion documents will be available for some software for this purpose.

That said, it is getting harder for many universities and companies to justify the price tag for proprietary statistical software, since there are now free versions available.

One that is quite popular with universities, companies, and data scientists is called R (r-project.org). R is a powerful programming language specifically designed to work with statistics. It is free, open source, has a huge community, and you can do anything you can imagine with statistics in it.

The downside of R is that it is a programming language. I don't imagine a large part of the target audience for this book is interested in learning how to load data into R, remember and type the right commands, and get a result. (If you are, go for it!)

There is a huge number of "packages" that can be added to R to make things easier for you. The problem is that not all of them perform statistics in the way the businessperson needs them.

That is why I created a graphic user interface for R called ROIstat. It has a familiar menu-driven, point-and-click interface and runs R for you. You don't

need to know anything about R to use it. You can get in, load your data, get your answer, and get on with your life.

You might wonder how much I would charge for this tool that will save you so much time and frustration.

It's free. I'd rather you spend your time learning about how to use statistics to make better decisions than how to program in R.

You can get the app and installation instructions here:

steveouellette.com/book-extras

Let me know what you think!

FINDING THE DATA AND FORMS

All the data and forms mentioned will also be available at:

steveouellette.com/book-extras.

Chapter 1 – Why do You Need Statistics?

By a small sample we may judge of the whole piece.

(de Cervantes, 1605)

What is Data?

People in the media, online, at work, and in life throw the word "data" around a lot, but we rarely stop to think about what data means, and for our purposes, how it affects what you can learn from it.

It starts with something happening that we are interested in understanding. Let's say you manage a process that makes a crunchy salty snack like corn chips. You might be interested in how fast the product is being made.

You could walk by the corn chip machine and watch it. The good news is that you are directly observing the reality of what you are interested in. The bad news is you really aren't learning anything about the production rate. You might see people running around faster today than yesterday or hear a higher-pitched sound coming from the machine, but you really don't know what, if anything, any of that means.

In order to find meaning in these observations, we need a way to turn them into something we can understand – numbers. This isn't as simple as it might appear. To turn a real event into a number, we need a measurement system. A **measurement system** takes an event and, through the interaction of people, processes, equipment, standards, and the environment, produces a number.

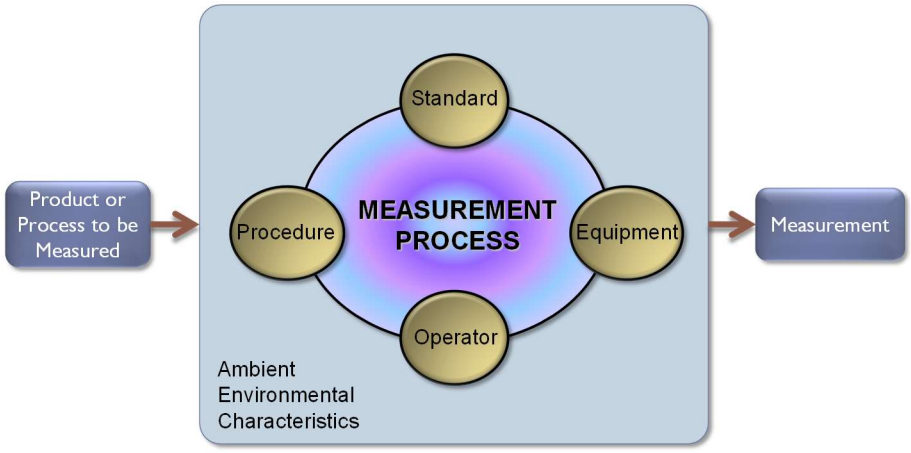


Figure 1-1 - The Components of a Measurement System

When we gather a bunch of these numbers or measurements, we have some data.

With all of those things going on, you might be wondering how we get any useful data, and you would be right to wonder about that. Although measurement system analysis is beyond the scope of this text, I want you to know that if something goes wrong here, *every other thing in this text is irrelevant*.

Before you start committing resources to a problem, be sure that the problem is not just with your measurement system itself. I have seen lots of time and money wasted on solving a problem that was just the result of a bad measurement system. Consult with someone who can help you understand the sources of measurement variability through time.

What is a Statistic?

A statistic is just a number calculated from data. It describes or summarizes something about the data you collected.

There are two types of statistics:

- **Descriptive statistics** summarize a data set that you have.
- **Inferential statistics** use descriptive statistics to predict what is actually happening today and what you might expect in the future.

You can use inferential statistics to improve your chances of making good decisions.

Population vs. Sample

Perhaps you are interested in all the chips you have made, are making, and will make in the future. This is called the "population." The **population** is the group of things that you want to know about that have something in common, for example "chips made on Line 2 today and into the future."

Of course, if we had all those chips, past, present, and future we would have all the information to make perfect decisions. But not only do we not have the future data, we often can't even get all the current or past data. Customers insist on actually having chips to eat before they pay you.

Instead, we usually only have access to a smaller subgroup of the population, called a **sample**. A sample is smaller, and cheaper, and statistics allow us to use the sample to infer something about the population.

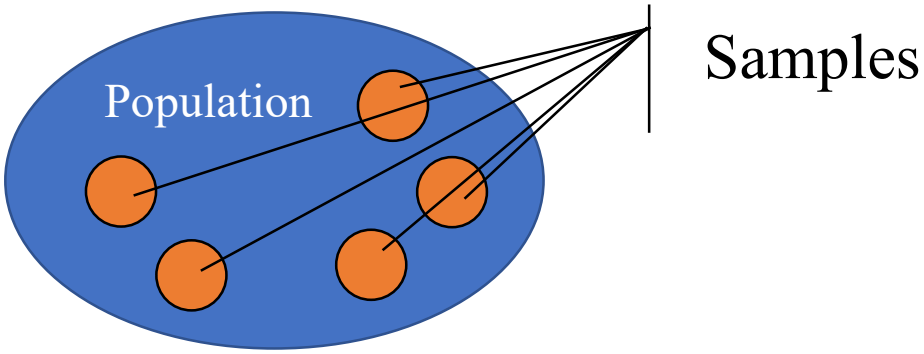


Figure 1-2 - Samples and the Population

Interestingly, you often get better information from a sample than you do if you actually try to measure everything.

In the US, the government is required to count each and every person in the country every 10 years as part of a census. But there are so many ways to miscount people that most of the real policy decisions are made based on estimates from smaller samples that are more accurate than the actual total numbers collected.

Random Sampling

But statistics only work if the sample you are using actually represents the process you want to measure.

You might go out and grab the first bag of chips off of the line, open it up and have a taste. Maybe you find a burnt chip. Does that burnt chip represent what the line usually produces, or might the fact that the line is just starting up have something to do with it? Without more information we don't know.

To get a sample that has the best chance of representing a process, we take a **random sample**. Without a random sample, you run the risk of using data that doesn't really represent the process, and which, you guessed it, can lead to expensive wrong decisions.

A **simple random sample** means that any unit in the population has an equal chance of being selected. There are a LOT of different ways to get representative samples other than simple random sampling, but all of them have randomness as part of the process. Randomizing to get a sample is how we try to control for the things we don't know and don't care to learn at this time. We randomize taking samples through the day because we don't know if time somehow affects the process. We randomize where on the conveyor belt we take a sample because we don't know if that has an effect on the chips.

The point is that you can do all the fanciest statistics correctly and still get the wrong answer if your original data was not sampled randomly.

In the real world, we often have to make compromises when sampling. For example, if we send out a customer survey, the responses we get back are not random, they are from customers who were motivated enough to return the survey, for whatever reason.

We can certainly take the data and analyze it, but you need to understand that you have introduced an unknown, possibly disastrous, bias into the data. What if the only customers who answered were the really grumpy ones with an axe to grind? What if the responses were only the customers who absolutely loved the company and the product? Or, what if it was both, and the uninterested customers in between are not being represented?

In all those cases you might make decisions as to where to spend your resources only to find a much smaller effect than you expected, or none at all. Or maybe made it even made something worse.

Don't give up hope! There is a lot of research about various ways to get random samples that goes beyond what we are going to cover here. But if you are presented a report with data analysis and it is based on samples, one of your questions should be how the researchers ensured that those samples represent the process and how they were randomly selected.

Statistics and Parameters

A **statistic** is a number calculated from a sample, but we are almost never interested in the statistic itself, we are interested in what that statistic tells us about the population we are really interested in. For example, if you were interested in how many of the chips produced were burned in a day, you could take a random sample of chips on the line throughout the day and see what percent of the sample were burnt.

But you don't really care about the sample that you collected; you care about the other chips that went to the customers who might complain about burnt chips. If your sample had 50% burnt chips, that probably tells you something about the chips you sent to the customers.

We use statistics from the sample to infer something about the **parameters** of the population. If we could go back in time and measure all the chips, we would know that the parameter of burnt chips in today's production turned out to be, say, 45%. Our statistic of 50%, was a good estimate but was not exactly the same as if we measured all the chips.

The way I remember this is that they start with the same letter:

- A Sample generates a Statistic.
- A Population has a Parameter.

We make decisions based on what we think is going on with the parameters.

For this book, we will adopt the nomenclature of using Roman letters for samples and Greek letters for parameters. Here are some examples to illustrate what I mean, though we won't get into what they are for a bit.

Table 1-1 - Symbols for Common Statistics and Parameters

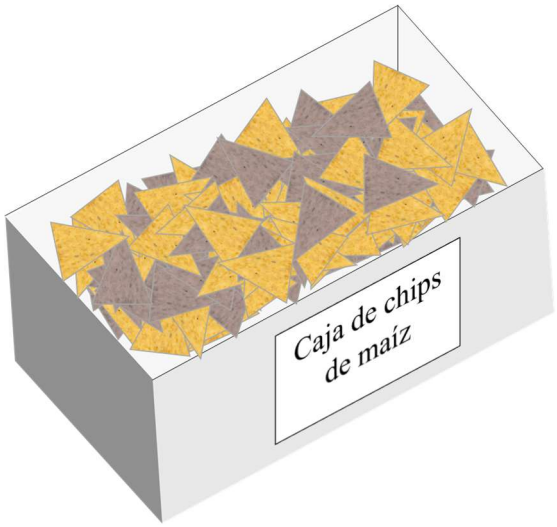
Characteristic	Statistic from a Sample	Parameter of a Population
Individual measure	X	Not applicable
Average (denoted by a bar above the statistic)	$\bar{X}$	μ
Standard deviation	s	σ
Proportion	p	π
Count	c	λ
Skewness	g_3	γ_3
Kurtosis	g_4	γ_4

Sampling Error

Did you notice above when the sample was 50% burnt chips but I said that the population actually had 45% burnt chips? That happens all the time – statistics are usually not exactly equal to the parameter.

But why not?

Let's say I have 1000 chips, of which 45% are burnt. That means I have 450 burnt chips and 550 unburnt chips. Imagine those are mixed up in a bin.



I could take a random sample of ten chips and find anything from none of them being burnt to all of them being burnt. I am not guaranteed to get the same proportion of burnt chips in my sample as are in the whole bin. It just depends on which ones were randomly selected.

The fact that my sample statistic differs from my population parameter is called **sampling error**. To my mind, it is unfortunate that the word "error" is in that phrase, since it implies that you did something wrong. In fact, in real world experimentation we work very hard to reduce all the other ways of getting the wrong answer since the best case is when we have only pure sampling error.

Of course, you might guess that the chances for getting a sample of zero or ten burnt chips in a sample burnt are not as high as getting a few burnt, and you would be right. The amount of sampling error is potentially knowable. That is the essence of inferential statistics.

The Chain of Inference

While we really care about the population, in the real world it is expensive, often impossible, and completely unnecessary to measure the population. So, we will end up inferring something about the population from a sample. "There ain't no such thing as a free lunch" though. As you would guess, as you move away from measuring everything you start to incur some risks.

Making a good decision from a sample is like a chain that needs strong links to hold.

For example, if you are interested in trying new settings to reduce the number of burnt chips, the population is "all chips that I produce from today on." But we are doing the experiment today and can't test the future, so we only have today's chip production available for the test. That is the **research population**. And of all the chips produced today we are only going to examine a few to count the burnt and unburnt chips. That is the sample.

Figure 1-3 - The Inferential Chain

If we don't do random sampling, we don't know that the sample represents today's production, so that makes the link between them (referred to as **internal validity**) weaker. If today is particularly humid, today's production might not really represent production into the future, and that would make that link (referred to as **external validity**) weaker.

In this case, by the time we try to use today's non-random sample to infer something about the population of the future, we are relying on a weak set of links in the **chain of inference**. So, we could calculate all the statistics right only to have results that don't predict what happens tomorrow. And then we might decide to make a change that doesn't reduce, or might even increase, the amount of burnt chips going forward.

To avoid this pitfall, when someone presents the results of an experiment, make sure you understand how they kept the chain of inference strong.

Summary

Measurement systems turn an event you are interested in understanding into numbers called data. Data can be summarized using a variety of statistics, and using statistics we try to infer something about the larger population, even the ones we didn't measure. To make the best decisions we need to use good sampling techniques so that the data is related to the population. Ensuring the inferential chain is strong means that we have a better chance of making the right decisions.

Asking the Right Questions: Chapter 1

- What is the population here?
- How were the samples collected, and are they representative of the population?
- If the samples run the risk of not representing the population, what was done to control for this risk?
- How much sampling error is present?
- How did you preserve the chain of inference?
- What are your threats to internal and external validity? How are you mitigating them?