BEYOND BASIC STATISTICS

Tips, Tricks, and Techniques Every Data Analyst Should Know

Kristin H. Jarman

WILEY

BEYOND BASIC STATISTICS

Tips, Tricks, and Techniques Every Data Analyst Should Know

KRISTIN H. JARMAN

WILEY

Published by John Wiley & Sons, Inc., Hoboken, New Jersey
Published simultaneously in Canada

Library of Congress Cataloging-in-Publication Data:

Jarman, Kristin H.
Beyond basic statistics: tips, tricks, and techniques every data analyst should know / K.H. Jarman.
pages cm
Includes index.
ISBN 978-1-118-85611-6 (pbk.)
1. Mathematical statistics–Popular works. I. Title.
QA276.J37 2015
001.4'22–dc23

2014047952

Set in 10.5/13pt Times by SPi Publisher Services, Pondicherry, India

Printed and bound in Singapore by Markono Print Media Pte Ltd

10 9 8 7 6 5 4 3 2 1

1 2015

BEYOND BASIC STATISTICS

To Anna and Sarah

When something doesn't feel right, it probably isn't.

PREFACE

I've had my share of mistakes: spilled coffee, insensitive remarks, and red socks thrown into a load of white laundry. These are daily occurrences in my life. But it isn't these little, private mishaps that haunt me. It's the big ones, the data analysis disasters, the public humiliations resulting from my own carelessness, mistakes that only reveal themselves when I'm standing in front of a room full of important people, declaring the brilliance of my statistical conclusions to the world.

Fortunately, these humiliations appear much more often in my dreams than they do in real life. When they do happen, however, they hit me when I least expect them, when I'm rushed, or when I'm overconfident in my results. All of them are accidental. I certainly never mean to misinform, but when you analyze as much data as I do, small mistakes are bound to happen every now and then.

This book highlights some of the well-known shortcomings of basic statistics, shortcomings that can, if ignored, lead to false conclusions. It provides tips and tricks to help you spot problem areas in your data analysis and covers techniques to help you overcome them. If, somewhere within the chapters of this book, you find information that prevents you from experiencing your own statistical humiliation, then exposing my own embarrassment will have been worth it.

KRISTIN H. JARMAN

CONTENTS

1

INTRODUCTION: IT SEEMED LIKE THE RIGHT THING TO DO AT THE TIME

As a seasoned statistical scientist, I like to think I'm invincible when it comes to drawing reliable conclusions from data. I'm not, of course. Nobody is. Even the world's best data analysts make mistakes now and then. This is what makes us human.

Just recently, for example, I was humbled by the simplest of all statistical techniques: the confidence interval. I was working with a government panel, helping them to establish criteria for certifying devices that detect certain toxic substances. (Smoke detectors, for example, are certified so you know they're reliable; in other words, they're likely to sound an alarm when there's smoke, and keep quiet when there isn't). The committee members wanted to know how many samples to test in order to reach a certain confidence level on the probability of detection, the probability that, given the toxin is present, the device will actually sound an alarm.

No problem, I thought.

Back in my office, I grabbed a basic statistics book, pulled out the formula for a confidence interval of a proportion (or probability), and went to work. I began calculating the confidence bounds on the probability of detection for different testing scenarios, preparing recommendations as I went along. It wasn't until sometime later I realized all my calculations were wrong.

Beyond Basic Statistics: Tips, Tricks, and Techniques Every Data Analyst Should Know, First Edition. Kristin H. Jarman.

Well, not wrong, the formulas and numbers were correct. But they didn't really fit my problem. When I started the calculations, I'd neglected one small but important detail. The detection probability for the devices being tested is typically very high, say 0.95 or higher. The basic confidence interval for a proportion p uses a normal approximation, which only applies when $Np > 5$ and $N(1-p) > 5$. Since I was limited to relatively small sample sizes of $N = 80$ or less, at best I had $N(1-p) = 80 \times 0.05 = 4$. Not large enough for the standard confidence interval to apply.

This happens more than I care to admit, that I embark on a data analysis using the world's most common statistical techniques, only to realize that my data don't work with the tools I'm using. Maybe the data don't fit the nice, bell-shaped distribution required by most popular methods. Maybe there are extreme values that could skew my results. But whatever the problem, I know that if I don't address it or at least acknowledge the impact it might have on my results, I will be sorry in the end.

This book takes you beyond the basic statistical techniques, showing you how to uncover and deal with those less-than-perfect datasets that occur in the real world. In the following chapters, you'll be introduced to methods for finding outliers, determining if a sample conforms to a normal distribution, and testing hypotheses when your data aren't normal. You'll learn popular strategies for designing experimental studies and performing regression with multiple variables and polynomial functions. And you'll find many tips and tricks for dealing with difficult data.

WHEN GOOD STATISTICS GO BAD: COMMON MISTAKES AND THE IMPACT THEY HAVE

There are many ways good statistics can go wrong and many more ways they can impact a data analyst's life. But in my experience, the vast majority of these mishaps are caused by just a few relatively common mistakes:

- Answering the wrong question
- Gathering the wrong data
- Using the wrong statistical technique
- Misinterpreting the results

Anyone who deals with a lot of data commits at least one of these errors from time to time. In my most recent incident, where I was slapped down by a simple confidence interval, I was clearly applying the wrong technique. Thankfully, this error only cost me a little time and it was easily fixed.

In Chapter 9, I'll share another one of my statistical humiliations, a situation where I misinterpreted the results of an analysis, a mistake that could've ruined my reputation and cost my employer millions of dollars.

This book introduces many statistical techniques designed to keep you from making these four common errors. Chapters 2 and 3 focus on designing studies based on your research goals. Chapters 5–9 introduce statistical techniques that can help you select the right analysis for a particular problem. In all of the chapters, the emphasis lies not on the mathematics of statistics but on how and when to use different techniques so you can avoid making costly mistakes.

STATISTICS 101: CONCEPTS YOU SHOULD KNOW BEFORE READING THIS BOOK

The techniques taught in most introductory statistics classes are built on a relatively small number of concepts, things like the sample mean and the normal distribution. But not-so-basic techniques are built on them, too. Before you dive too deeply into the world of data analysis, it's important to have a working knowledge of a handful of concepts. Here are the ones you'll need to get the most out of this book. For a detailed introduction to these topics, see a basic statistics textbook such as the companion to this book, *The Art of Data Analysis: How to Answer Almost Any Question Using Basic Statistics* by yours truly.

Probability Theory

Statistics and data analysis rely heavily on mathematical probability. Mathematical probability is concerned with describing randomness, and all of the functions and complex formulas you see in a statistics book were derived from this branch of mathematics. To understand the techniques presented in this book, you should be familiar with the following topics from probability.

Random Variables and Probability Distributions A **random variable** represents the outcome of a random experiment. Typically denoted by a capital letter such as X or Y, a random variable is similar to a variable x or y from algebra. Where the variable x or y represents some as yet unsolved value in an algebraic equation, the variable X or Y represents some as-yet-undetermined outcome of a random experiment. For example, on a coin toss, with possible outcomes *heads* and *tails*, you could define a random variable $X=0$ for *tails* and $X=1$ for *heads*. This value of X is undetermined until the experiment is complete.

A **probability distribution** is a mathematical formula for assigning probabilities to the outcomes of a random experiment. Many different probability distributions have been developed over the years, and these can be used to assign probabilities in almost any random experiment you can imagine. Whether or not you'll win the lottery, how many times your new car will break down in the first year, the amount of radioactivity you'll absorb while scooping out your cat's litter box, all of these events have a probability distribution associated with them.

Expected Values and Parameters of a Distribution A random variable is uncertain. You don't know exactly what value it will take until the experiment is over. You can, however, make predictions. The **expected value** is just that: a prediction as to what value a random variable will take on. The two most common expected values are the mean and variance. The mean predicts the value of the random variable, and the variance predicts the likely deviation from the mean. The **parameters** of a distribution are values that specify the exact behavior of a random variable. Every probability distribution has at least one parameter associated with it. The most common parameters are also expected values: in particular, the mean and variance.

Statistics

Statistics is the application of probability to real data. Where probability is concerned with describing the mathematical properties of random variables, statistics is concerned with estimating or predicting mathematical properties from a set of observations. Here are the basic concepts used in this book.

Population vs. Sample In any study, the goal is to learn something about a **population**, the collection of all people, places, or things you are interested in. It's usually too costly or too time-consuming to collect data from the entire population, so you typically must rely on a **sample**, a carefully selected subset of the population.

Parameter vs. Estimate A parameter is a value that characterizes a probability distribution, or a population. An estimate is a value calculated from a dataset that estimates the corresponding population parameter. For example, think of the population mean and sample mean, or average. The population mean is a parameter, the true (often unknown) center of the population. The sample mean is an estimate, an educated guess as to what the population mean might be.

Discrete vs. Continuous Data Any data collection exercise produces one or more outcomes, and these outcomes—called observations, measurements, or data—can be either discrete or continuous. **Discrete observations** are whole numbers, counts, or categories, in other words, anything that can be easily listed. For example, the outcome of one roll of a six-sided die is discrete. **Continuous observations**, on the other hand, cannot be listed. Real numbers are continuous. If you choose any two real numbers, no matter which two you choose, there's always some number in between them. Different statistical techniques are often applied to discrete and continuous data.

Descriptive Statistics **Descriptive statistics** are estimates for the center location, shape, texture, and other properties of a population. Descriptive statistics are the foundation of data analysis. They're used to describe a sample, construct margins of error, compare two datasets, find relationships between variables, and just about anything else you might want to do with your data. The two most common descriptive statistics are the sample mean (average) and standard deviation.

The **average**, or **sample mean**, describes center location of a sample. Calculated as the sum of all your data values divided by the number of data values in the dataset, the average is the arithmetic center of a set of observations. The standard deviation measures the spread of a set of observations. The **standard deviation** is the average deviation, or variation, of all the values around the center location.

Sample Statistics and Sample Distributions A **sample statistic** is calculated from a dataset. It's a value with certain statistical properties that can be used to construct confidence intervals and perform hypothesis tests. A z-statistic is an example of a sample statistic. A **sample distribution** is a probability distribution for a sample statistic. Critical thresholds and p-values used in confidence intervals and hypothesis tests are calculated from sample distributions. Examples of such distributions include the z-distribution and the t-distribution.

Confidence Intervals A **confidence interval**, or **margin of error**, is a measure of confidence in a descriptive statistic, most commonly the sample mean. Confidence intervals are typically reported as a mean value plus or minus some margin of error, say 8 ± 2 or as a corresponding range, such as (6, 10).

Hypothesis Tests A **hypothesis test** uses data to compare competing claims about a population in order to determine which claim is most likely. There are typically two hypotheses being compared: H_0 and H_A. H_0 is called the **null**

hypothesis. It's the fall-back position. It's what you're automatically assuming to be true. H_A is the **alternative hypothesis**. This is the claim you accept as true only if you have enough evidence in the data to reject H_0.

Hypothesis tests are performed by comparing a test statistic to a critical threshold. The **test statistic** is a sample statistic, a value calculated from the data. This value carries evidence for or against H_0. The **critical threshold** is a value calculated from a sample distribution and the **significance level**, or probability of falsely rejecting H_0. You compare the test statistic to this threshold in order to decide whether to accept that H_0 is true, or reject H_0 in favor of H_A.

Alternatively, you can use the test statistic to calculate a ***p*-value**, a probability for the evidence under the null hypothesis, and compare it to the significance level of the test. If the *p*-value is smaller than the significance level, then H_0 is rejected.

In general, hypothesis tests are either one-sided or two-sided. A **one-sided hypothesis test** looks for deviations from the null hypothesis in one direction only, for example, when testing if the mean of a population is zero or *greater than* zero. A **two-sided hypothesis test** looks for deviations in both directions, as in testing whether the mean of a population is zero or *not equal* to zero. One-sided and two-sided hypothesis tests often have the same test statistic, but to achieve the same significance level, they typically end up using use different critical thresholds.

Linear Regression **Linear regression** is a common modeling technique for predicting the value of a dependent variable *Y* from a set of independent *X* variables. In linear regression, a line is used to describe the relationship between the *X*s and *Y*. **Simple linear regression** is linear regression with a single *X* and *Y* variable.

TIPS, TRICKS, AND TECHNIQUES: A ROADMAP OF WHAT FOLLOWS

Each chapter in this book begins by asking a specific question and reviewing the basic statistics approach to answering it. Common problems that can derail the basic approach are presented, followed by a discussion of methods for overcoming them. Along the way, tips and tricks are introduced, taking you beyond the techniques themselves into the real-world application of them. In most cases, the chapter wraps up with a case study that pulls the different concepts together and answers the question posed at the beginning.