STATISTICAL ANALYSIS

Microsoft Excel® 2016

QUE®

Conrad Carlberg

Statistical Analysis: Microsoft Excel® 2016

Conrad Carlberg

QUE®

800 East 96th Street,
Indianapolis, Indiana 46240 USA

Contents at a Glance

Statistical Analysis: Microsoft Excel® 2016

Copyright © 2018 by Pearson Education, Inc.

All rights reserved. No part of this book shall be reproduced, stored in a retrieval system, or transmitted by any means, electronic, mechanical, photocopying, recording, or otherwise, without written permission from the publisher. No patent liability is assumed with respect to the use of the information contained herein. Although every precaution has been taken in the preparation of this book, the publisher and author assume no responsibility for errors or omissions. Nor is any liability assumed for damages resulting from the use of the information contained herein.

ISBN-13: 978-0-7897-5905-4
ISBN-10: 0-7897-5905-5

Library of Congress Control Number: 2017955944

Printed in the United States of America

1 17

Trademarks

All terms mentioned in this book that are known to be trademarks or service marks have been appropriately capitalized. Que Publishing cannot attest to the accuracy of this information. Use of a term in this book should not be regarded as affecting the validity of any trademark or service mark.

Warning and Disclaimer

Every effort has been made to make this book as complete and as accurate as possible, but no warranty or fitness is implied. The information provided is on an “as is” basis. The author and the publisher shall have neither liability nor responsibility to any person or entity with respect to any loss or damages arising from the information contained in this book.

Special Sales

For information about buying this title in bulk quantities, or for special sales opportunities (which may include electronic versions; custom cover designs; and content particular to your business, training goals, marketing focus, or branding interests), please contact our corporate sales department at corpsales@pearsoned.com or (800) 382-3419.

For government sales inquiries, please contact governmentsales@pearsoned.com.

For questions about sales outside the U.S., please contact intlcs@pearsoned.com.

Editor-in-Chief
Greg Wiegand

Acquisitions Editor
Trina MacDonald

Development Editor
Charlotte Kughen

Managing Editor
Sandra Schroeder

Project Editor
Mandie Frank

Copy Editor
Chuck Hutchinson

Indexer
Erika Millen

Proofreader
Abigail Manheim

Technical Editor
Michael Turner

Editorial Assistant
Courtney Martin

Designer
Chuti Prasertsith

Compositor
codeMantra

Microsoft and/or its respective suppliers make no representations about the suitability of the information contained in the documents and related graphics published as part of the services for any purpose. All such documents and related graphics are provided "as is" without warranty of any kind. Microsoft and/ or its respective suppliers hereby disclaim all warranties and conditions with regard to this information, including all warranties and conditions of merchantability, whether express, implied or statutory, fitness for a particular purpose, title and non-infringement. In no event shall Microsoft and/or its respective sup-pliers be liable for any special, indirect or consequential damages or any damages whatsoever resulting from loss of use, data or profits, whether in an action of contract, negligence or other tortious action, arising out of or in connection with the use or performance of information available from the services.

The documents and related graphics contained herein could include technical inaccuracies or typographical errors. Changes are periodically added to the information herein. Microsoft and/or its respective sup-pliers may make improvements and/or changes in the product(s) and/or the program(s) described herein at any time. Partial screenshots may be viewed in full within the software version specified.

Microsoft® and Windows® are registered trademarks of the Microsoft Corporation in the U.S.A. and other countries. Screenshots and icons reprinted with permission from the Microsoft Corporation. This book is not sponsored or endorsed by or affiliated with the Microsoft Corporation.

Contents

About the Author

Conrad Carlberg started writing about Excel, and its use in quantitative analysis, before workbooks had worksheets. As a graduate student, he had the great good fortune to learn something about statistics from the wonderfully gifted Gene Glass. He remembers much of that and has learned more since. This is a book he has wanted to rewrite for years, and he is grateful for the opportunity.

Dedication

For Toni, who has been putting up with this sort of thing for almost 25 years now, with all my love.

Acknowledgments

I'd like to thank Trina MacDonald, who guided this book's overall progress. Once again, Michael Turner's technical edit was just right. Chuck Hutchinson's deft copy edit not only kept the prose sensible but corrected a number of technical howlers. And in the end, Mandie Frank got the whole thing out the door. My thanks to each of you.

We Want to Hear from You!

As the reader of this book, *you* are our most important critic and commentator. We value your opinion and want to know what we're doing right, what we could do better, what areas you'd like to see us publish in, and any other words of wisdom you're willing to pass our way.

We welcome your comments. You can email or write to let us know what you did or didn't like about this book—as well as what we can do to make our books better.

Please note that we cannot help you with technical problems related to the topic of this book.

When you write, please be sure to include this book's title and author as well as your name and email address. We will carefully review your comments and share them with the author and editors who worked on the book.

Email: feedback@quepublishing.com

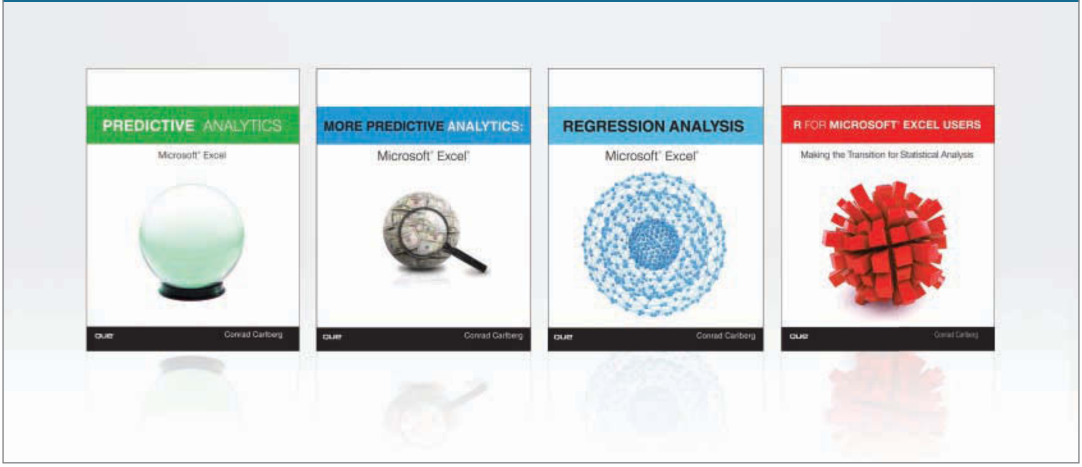
Mail: Que Publishing
ATTN: Reader Feedback
800 East 96th Street
Indianapolis, IN 46240 USA

Reader Services

Register your copy of *Statistical Analysis: Microsoft Excel® 2016* at quepublishing.com for convenient access to downloads, updates, and corrections as they become available. To start the registration process, go to quepublishing.com/register and log in or create an account*. Enter the product ISBN, 9780789759054, and click Submit. Once the process is complete, you will find any available bonus content under Registered Products.

*Be sure to check the box that you would like to hear from us in order to receive exclusive discounts on future editions of this product.

Conrad Carlberg’s Microsoft® Excel® Analytics Series



Visit **informit.com/carlberg** for a complete list of available publications.

Conrad Carlberg, a nationally recognized expert on quantitative analysis and data analysis applications, shows you how to use Excel to perform a wide variety of analyses to solve real-world business problems. Employing a step-by-step tutorial approach, Carlberg delivers clear explanations of proven Excel techniques that can help you increase revenue, reduce costs, and improve productivity. With each book comes an extensive collection of Excel workbooks you can adapt to your own projects. Conrad’s books will show you how to:

- Build powerful, credible, and reliable forecasts
- Use smoothing techniques to build accurate predictions from trended and seasonal baselines
- Employ Excel’s regression-related worksheet functions to model and analyze dependent and independent variables—and benchmark the results against R
- Use decision analytics to evaluate relevant information critical to the business decision-making process

Written using clear language in a straightforward, no-nonsense style, Carlberg makes data analytics easy to learn and incorporate into your business.

Pearson

informIT.com
the trusted technology learning source

O’REILLY®
Safari

INTRODUCTION

IN THIS INTRODUCTION

There was no reason I shouldn't have already written a book about statistical analysis using Excel. But I didn't, although I knew I wanted to. Finally, I talked Pearson into letting me write it for them.

Be careful what you ask for. It's been a struggle, but at last I've got it out of my system, and I want to start by talking here about the reasons for some of the choices I made in writing this book.

Using Excel for Statistical Analysis

The problem is that it's a huge amount of material to cover in a book that's supposed to be only 400 to 500 pages. The text used in the first statistics course I took was about 600 pages, and it was purely statistics, no Excel. I have coauthored a book about Excel (no statistics) that ran to 750 pages. To shoehorn statistics *and* Excel into 520 pages or so takes some picking and choosing.

Furthermore, I did not want this book to be simply an expanded Help document. Instead, I take an approach that seemed to work well in other books I've written. The idea is to identify a topic in statistical analysis; discuss the topic's rationale, its procedures, and associated issues; and illustrate them in the context of Excel worksheets.

That approach can help you trace the steps that lead from a raw data set to, say, a complete multiple regression analysis. It helps to illuminate that rationale, those procedures, and the associated issues. And it often works the other way, too. Walking through the steps in a worksheet can clarify their rationale.

You shouldn't expect to find discussions of, say, the Weibull function or the lognormal distribution here.

They have their uses, and Excel provides them as statistical functions, but my picking and choosing forced me to ignore them—at my peril, probably—and to use the space saved for material on more bread-and-butter topics such as statistical regression.

About You and About Excel

How much background in statistics do you need to get value from this book? My intention is that you need none. The book starts out with a discussion of different ways to measure things—by categories, such as models of cars, by ranks, such as first place through tenth, by numbers, such as degrees Fahrenheit—and how Excel handles those methods of measurement in its worksheets and its charts.

This book moves on to basic statistics, such as averages and ranges, and only then to intermediate statistical methods such as t-tests, multiple regression, and the analysis of covariance. The material assumes knowledge of nothing more complex than how to calculate an average. You do not need to have taken courses in statistics to use this book. (If you have taken statistics courses, that'll help. But they aren't prerequisites.)

As to Excel itself, it matters little whether you're using Excel 97, Excel 2016, or any version in between. Very little statistical functionality changed between Excel 97 and Excel 2003. The few changes that did occur had to do primarily with how functions behaved when the user stress-tested them using extreme values or in very unlikely situations.

The Ribbon showed up in Excel 2007 and is still with us in Excel 2016. But nearly all statistical analysis in Excel takes place in worksheet functions—very little is menu driven—and there was almost no change to the function list, function names, or their arguments between Excel 97 and Excel 2007. The Ribbon does introduce a few differences, such as how you create a chart. Where necessary, this book discusses the differences in the steps you take using the older menu structure and the steps you take using the Ribbon.

In Excel 2010, several apparently new statistical functions appeared, but the differences were more apparent than real. For example, through Excel 2007, the two functions that calculate standard deviations are STDEV() and STDEVP(). If you are working with a sample of values, you should use STDEV(), but if you happen to be working with a full population, you should use STDEVP().

Both STDEV() and STDEVP() remain in Excel 2016, but they are termed *compatibility functions*. It appears that they might be phased out in some future release. Excel 2010 added what it calls *consistency functions*, two of which are STDEV.S() and STDEV.P(). Note that a period has been added in each function's name. The period is followed by a letter that, for consistency, indicates whether the function should be used with a sample of values (you're working with a statistic) or a population of values (you're working with a parameter).

Other consistency functions were added to Excel 2010, and the functions they are intended to replace are still supported in Excel 2016. There are a few substantive differences between the compatibility version and the consistency version of some functions, and this book discusses those differences and how best to use each version.

Clearing Up the Terms

Terminology poses another problem, both in Excel and in the field of statistics (and, it turns out, in the areas where the two overlap). For example, it's normal to use the word *alpha* in a statistical context to mean the probability that you will decide that there's a true difference between the means of two populations when there really isn't. But Excel extends *alpha* to usages that are related but much less standard, such as the probability of getting some number of heads from flipping a fair coin. It's not wrong to do so. It's just unusual, and therefore it's an unnecessary hurdle to understanding the concepts.

The vocabulary of statistics itself is full of names that mean very different things in slightly different contexts. The word *beta*, for example, can mean the probability of deciding that a true difference does *not* exist, when it does. It can also mean a coefficient in a regression equation (for which Excel's documentation unfortunately uses the letter *m*), and it's also the name of a distribution that is a close relative of the binomial distribution. None of that is due to Excel. It's due to having more concepts than there are letters in the Greek alphabet.

You can see the potential for confusion. It gets worse when you hook Excel's terminology up with that of statistics. For example, in Excel the word *cell* means a rectangle on a worksheet, the intersection of a row and a column. In statistics, particularly the analysis of variance, *cell* usually means a group in a factorial design: If an experiment tests the joint effects of sex and a new medication, one cell might consist of men who receive a placebo, and another might consist of women who receive the medication being assessed. Unfortunately, you can't depend on seeing "cell" where you might expect it: *within cell error* is called *residual error* in the context of regression analysis. (In regression analysis, you often calculate error variance indirectly, by way of subtraction—hence, *residual*).

So this book presents you with some terms you might otherwise find redundant: I use *design cell* for analysis contexts and *worksheet cell* when referring to the worksheet context, where there's any possibility of confusion about which I mean.

For consistency, though, I try always to use *alpha* rather than *Type I error* or *statistical significance*. In general, I use just one term for a given concept throughout. I intend to complain about it when the possibility of confusion exists: When *mean square* doesn't mean *mean square*, you ought to know about it.

Making Things Easier

If you're just starting to study statistical analysis, your timing's much better than mine was. You have avoided some of the obstacles to understanding statistics that once stood in the way. I'll mention those obstacles once or twice more in this book, partly to vent my spleen but also to stress how much better Excel has made things.

Suppose that quite a few years back you were calculating something as basic as the standard deviation of 20 numbers. You had no access to a computer. Or, if there was one around, it was a mainframe or a mini, and whoever owned it had more important uses for it than to support a Psychology 101 assignment.

So you trudged down to the Psych building's basement, where there was a room filled with gray metal desks with adding machines on them. Some of the adding machines might even have been plugged into a source of electricity. You entered your 20 numbers very carefully because the adding machines did not come with Undo buttons or Ctrl+Z. The electricity-enabled machines were in demand because they had a memory function that allowed you to enter a number, square it, and add the result to what was already in the memory.

It could take half an hour to calculate the standard deviation of 20 numbers. It was all incredibly tedious and it distracted you from the main point, which was the concept of a standard deviation and the reason you wanted to quantify it.

Of course, back then our teachers were telling us how lucky we were to have adding machines instead of having to use paper, pencil, and a box of erasers.

Things are different now, and truth be told, they have been changing since the late 1980s when applications such as Lotus 1-2-3 and Microsoft Excel started to find their way onto personal computers' floppy disks. Now, all you have to do is enter the numbers into a worksheet—or maybe not even that, if you downloaded them from a server somewhere. Then, type **=STDEV.S(** and drag across the cells with the numbers before you press Enter. It takes half a minute at most, not half an hour at least.

Many statistics have relatively simple *definitional* formulas. The definitional formula tends to be straightforward and therefore gives you actual insight into what the statistic means. But those same definitional formulas often turn out to be difficult to manage in practice if you're using paper and pencil, or even an adding machine or hand calculator. Rounding errors occur and compound one another.

So statisticians developed *computational* formulas. These are mathematically equivalent to the definitional formulas, but are much better suited to manual calculations. Although it's nice to have computational formulas that ease the arithmetic, those formulas make you take your eye off the ball. You're so involved with accumulating the sum of the squared values that you forget that your purpose is to understand how values vary around their average.

That's one primary reason that an application such as Excel, or an application specifically and solely designed for statistical analysis, is so helpful. It takes the drudgery of the arithmetic off your hands and frees you to think about what the numbers actually mean.

Statistics is conceptual. It's not just arithmetic. And it shouldn't be taught as though it is.

The Wrong Box?

But should you even be using Excel to do statistical calculations? After all, people have been running around, hair afire, about inadequacies in Excel's statistical functions for years. Back when there was a CompuServe, its Excel forum had plenty of complaints about this issue, as did the subsequent Usenet newsgroups. As I write this introduction, I can switch from Word to a browser and see that some people are still complaining on Wikipedia talk pages, and others contribute angry screeds to publications such as *Computational Statistics & Data*

Analysis, which I believe are there as a reminder to us all of the importance of taking a deep breath every so often.

I have sometimes found myself as upset about problems with Excel's statistical functions as anyone. And it's true that Excel has had, and in some cases continues to have, problems with the algorithms it uses to manage certain statistical functions.

But most of the complaints that are voiced fall into one of two categories: those that are based on misunderstandings about either Excel or statistical analysis, and those that are based on complaints that Excel isn't accurate enough.

If you read this book, you'll be able to avoid those misunderstandings. As to complaints about inaccuracies in Excel results, let's look a little more closely at that. The complaints are typically along these lines:

> I enter into an Excel worksheet two different formulas that should return the same result. Simple algebraic rearrangement of the equations proves that. But then I find that Excel calculates two different results.

Well, for the data the user supplied, the results differ at the fifteenth decimal place, so Excel's results disagree with one another by approximately five in 111 trillion.

Or this:

> I tried to get the inverse of the F distribution using the formula FINV(0.025,4198986,1025419), but I got an unexpected result. Is there a bug in FINV?

No. Once upon a time, FINV returned the #NUM! error value for those arguments, but no longer. However, that's not the point. With so many degrees of freedom (over four million and one million, respectively), the person who asked the question was effectively dealing with populations, not samples. To use that sort of inferential technique with so many degrees of freedom is a striking instance of "unclear on the concept."

Would it be better if Excel's math were more accurate—or at least more internally consistent? Sure. But even finger-waggers admit that Excel's statistical functions are acceptable at least, as the following comment shows:

> They can rarely be relied on for more than four figures, and then only for $0.001 < p < 0.999$, plenty good for routine hypothesis testing.

Now look. Chapter 8, "Telling the Truth with Statistics," goes further into this issue, but the point deserves a better soapbox, closer to the start of the book. Regardless of the accuracy of a statement such as "They can rarely be relied on for more than four figures," it's pointless to make it. It's irrelevant whether a finding is "statistically significant" at the 0.001 level instead of the 0.005 level, and to worry about whether Excel can successfully distinguish between the two findings is to miss the context.

There are many possible explanations for a research outcome other than the one you're seeking: a real and replicable treatment effect. Random chance is only one of these. It's one that gets a lot of attention because we attach the word *significance* to our tests to rule out chance, but it's not more important than other possible explanations you should be concerned about when you design your study. It's the design of your study, and how well you implement it, that allows you to rule out alternative explanations such as selection bias and statistical regression. Those explanations—selection bias and regression—are just two examples of possible alternative explanations for an apparent treatment effect: explanations that might make a treatment look like it had an effect when it actually didn't.

Even the strongest design doesn't enable you to rule out a chance outcome. But if the design of your study is sound, and you obtained what looks like a meaningful result, you'll want to control chance's role as an alternative explanation of the result. So, you certainly want to run your data through the appropriate statistical test, which *does* help you control the effect of chance.

If you get a result that doesn't clearly rule out chance—or rule it in—you're much better off to run the experiment again than to take a position based on a borderline outcome. At the very least, it's a better use of your time and resources than to worry in print about whether Excel's F tests are accurate to the fifth decimal place.

Wagging the Dog

And ask yourself this: Once you reach the point of planning the statistical test, are you going to reject your findings if they might come about by chance five times in 1,000? Is that too loose a criterion? What about just one time in 1,000? How many angels are on that pinhead anyway?

If you're concerned that Excel won't return the correct distinction between one and five chances in 1,000 that the result of your study is due to chance, you allow what's really an irrelevancy to dictate how, and using what calibrations, you're going to conduct your statistical analysis. It's pointless to worry about whether a test is accurate to one point in a thousand or two in a thousand. Your decision rules for risking a chance finding should be based on more substantive grounds.

Chapter 10, "Testing Differences Between Means: Further Issues," goes into the matter in greater detail, but a quick summary of the issue is that you should let the risk of making the wrong decision be guided by the costs of a bad decision and the benefits of a good one—not by which criterion appears to be the more selective.

What's in This Book

You'll find that there are two broad types of statistics. I'm not talking about that scurrilous line about lies, damned lies and statistics—both its source and its applicability are disputed. I'm talking about *descriptive* statistics and *inferential* statistics.

No matter if you've never studied statistics before this, you're already familiar with concepts such as averages and ranges. These are descriptive statistics. They describe identified groups: The average age of the members is 42 years; the range of the weights is 105 pounds; the median price of the houses is $370,000. A variety of other sorts of descriptive statistics exists, such as standard deviations, correlations, and skewness. The first six chapters of this book take a fairly close look at descriptive statistics, and you might find that they have some aspects that you haven't considered before.

Descriptive statistics provides you with insight into the characteristics of a restricted set of beings or objects. They can be interesting and useful, and they have some properties that aren't at all well known. But you don't get a better understanding of the world from descriptive statistics. For that, it helps to have a handle on inferential statistics. That sort of analysis is based on descriptive statistics, but you are asking and perhaps answering broader questions. Questions such as this:

> The average systolic blood pressure in this sample of patients is 135. How large a margin of error must I report so that if I took another 99 samples, 95 of the 100 would capture the true population mean within margins calculated similarly?

Inferential statistics enables you to make inferences about a population based on samples from that population. As such, inferential statistics broadens the horizons considerably.

Therefore, I prepared new material on inferential statistics for the 2013 edition and 2016 editions of *Statistical Analysis: Microsoft Excel*. Chapter 13, "Experimental Design and ANOVA," explores the effects of fixed versus random factors on the nature of your F-tests. It also examines crossed and nested factors in factorial designs, and how a factor's status in a factorial design affects the mean square you should use in the F ratio's denominator. Chapter 13 also discusses how to adjust the analysis to accommodate randomized block designs such as repeated measures.

In recent years, Excel has added some charts that are particularly useful in statistical analysis. There are enough such charts now that two new ones deserve and own chapter in this edition, Chapter 5, "Charting Statistics."

You have to take on some assumptions about your samples, and about the populations that your samples represent, to make the sort of generalization that inferential statistics support. From Chapter 7 through the end of this book, you'll find discussions of the issues involved, along with examples of how those issues work out in practice. And, by the way, how you work them out using Microsoft Excel.

This page intentionally left blank

About Variables and Values

1

IN THIS CHAPTER

It must seem odd to start a book about statistical analysis using Excel with a discussion of ordinary, everyday notions such as variables and values. But variables and values, along with scales of measurement (discussed in the next section), are at the heart of how you represent data in Excel. And how you choose to represent data in Excel has implications for how you run the numbers.

With your data laid out properly, you can easily and efficiently combine records into groups, pull groups of records apart to examine them more closely, and create charts that give you insight into what the raw numbers are really doing. When you put the statistics into tables and charts, you begin to understand what the numbers have to say.

Variables and Values

When you lay out your data without considering how you will use the data later, it becomes much more difficult to do any sort of analysis. Excel is generally very flexible about how and where you put the data you're interested in, but when it comes to preparing a formal analysis, you want to follow some guidelines. In fact, some of Excel's features don't work at all if your data doesn't conform to what Excel expects. To illustrate one useful arrangement, you won't go wrong if you put different variables in different columns and different records in different rows.

A *variable* is an attribute or property that describes a person or a thing. Age is a variable that describes you. It describes all humans, all living organisms, all objects—anything that exists for some period of time. Surname is a variable, and so are Weight in Pounds and Brand of Car. Database jargon often