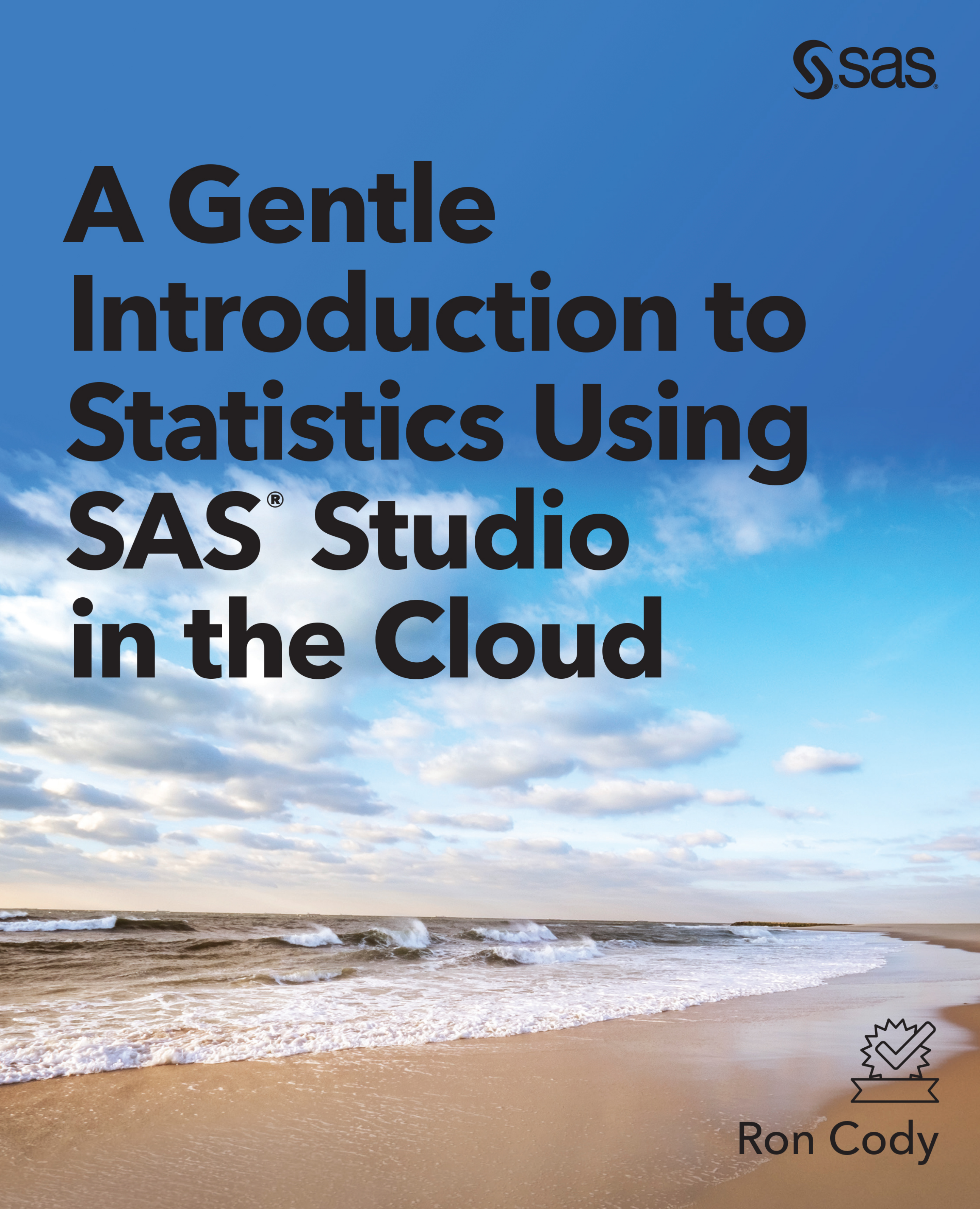


A Gentle Introduction to Statistics Using SAS® Studio in the Cloud

Ron Cody

The correct bibliographic citation for this manual is as follows: Cody, Ron. 2021. *A Gentle Introduction to Statistics Using SAS® Studio in the Cloud*. Cary, NC: SAS Institute Inc.

A Gentle Introduction to Statistics Using SAS® Studio in the Cloud

978-1-954844-49-0 (Hardcover)
978-1-954844-45-2 (Paperback)
978-1-954844-46-9 (Web PDF)
978-1-954844-47-6 (EPUB)
978-1-954844-48-3 (Kindle)

SAS Institute Inc., SAS Campus Drive, Cary, NC 27513-2414

April 2021

Contents

About this Book

What Does This Book Cover?

This book is designed to fulfill two purposes: one is to teach statistical concepts and the other is to show you how to perform statistical analysis using SAS Studio.

The book starts out with two short, introductory chapters describing statistical analysis (both descriptive and inferential) and experimental design. Following the introductory chapters are several chapters that show you how to register for SAS OnDemand for Academics, use some of the built-in SAS Studio tasks, how to upload data from your local computer to the SAS cloud, and how to convert data from multiple sources (such as Excel or CSV files) and create SAS data sets. There is one chapter on descriptive statistics, summarizing data both in table and graphical form. The remainder of the book describes most of the statistical tests that you will need for an introductory course in statistics.

Is This Book for You?

As the title suggests, this book is intended for someone with little or no knowledge of statistics and SAS, but it is also useful for someone with more statistical expertise who might not be familiar with SAS. One of the important points for beginners or people with more extensive knowledge of statistics, is a discussion of the assumptions that need to be satisfied for a particular statistical test to be valid. That is especially important because with SAS Studio tasks, anyone can click a mouse and perform very advanced statistical tests.

What Should You Know about the Examples?

Because you can download all of the programs and data sets used in this book from the SAS website, you can run any or all of the programs yourself to better understand how perform them.

Example Code and Data

You can access the example code and data for this book by linking to its author page at https://support.sas.com/cody.

SAS OnDemand for Academics

This book is compatible with SAS Studio and the SAS product called OnDemand for Academics. This is a cloud-based application that is free for anyone wanting to learn how to use SAS, not just college students. Although all the examples in the book were run using SAS OnDemand for Academics, you can run these tasks and programs on other versions of SAS Studio.

Where Are the Exercise Solutions?

Solutions to all the odd-numbered exercises are included at the end of the book. For those individuals who are not students, are working on their own, or are faculty members, please contact SAS Press for solutions to all of the exercises.

We Want to Hear from You

SAS Press books are written *by* SAS Users *for* SAS Users. We welcome your participation in their development and your feedback on SAS Press books that you are using. Please visit sas.com/books to do the following:

- Sign up to review a book
- Recommend a topic
- Request information on how to become a SAS Press author
- Provide feedback on a book

About the Author

Ron Cody, EdD, is a retired professor from the Rutgers Robert Wood Johnson Medical School who now works as a national instructor for SAS and as an author of books on SAS and statistics. A SAS user since 1977, Ron's extensive knowledge and innovative style have made him a popular presenter at local, regional, and national SAS conferences. He has authored or co-authored numerous books, as well as countless articles in medical and scientific journals.

Learn more about this author by visiting his author page at http://support.sas.com/cody. There you can download free book excerpts, access example code and data, read the latest reviews, get updates, and more.

1.

Descriptive and Inferential Statistics

Overview

Many people have a misunderstanding of what statistics entails. The trouble stems from the fact that the word “statistics” has several different meanings. One meaning relates to numbers such as batting averages and political polls. When I tell people that I am a statistician, they assume that I am good with numbers. Actually, without a computer I would be lost.

The other meaning, the topic of this book, is to **describe collections of numbers** such as test scores and to describe properties of these numbers. This subset of statistics is known as **descriptive statistics**. Another subset of statistics, **inferential statistics**, takes up a major portion of this book. One of the goals of inferential statistics is to determine whether your experimental results are “statistically significant.” In other words, what is the probability that the result that you obtained could have occurred by chance, rather than an actual effect?

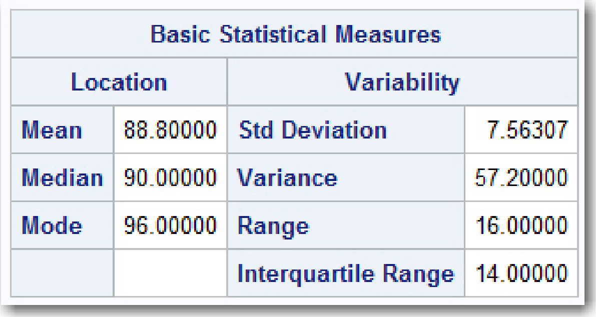
Descriptive Statistics

I am sure that every reader of this book is already familiar with some aspects of descriptive statistics. From early in your education, you were assigned a grade in a course, based on your average. Averages (there are several types) describe what statisticians refer to as **measures of location** or **measures of central tendency**. Most basic statistics books describe three indicators of location: the mean, median, and mode. To compute a mean, you add up all the numbers and divide by how many numbers you have. For example, if you took five tests and your scores were 80, 82, 90, 96, and 96, the mean would be (80 + 82 + 90 + 96 + 96)/5 or 88.8. To compute a median, you arrange the numbers in order from lowest to highest and then find the middle number. This number is the median. Half the numbers will be below the median and half of the numbers will be above the median. In the example of the five test scores (notice that they are already in order from lowest to highest), the median is 90. If you have an even number of numbers, one method of computing the median is to average the two numbers in the middle. The last measure of central tendency is called the mode. It is defined as the most frequent number. In this example, the mode is 96 because it occurs more than any other number. If all the numbers are different, the mode is not defined.

Besides knowing the mean or median (the mode is rarely used), you can also compute several measures of dispersion. **Dispersion** describes how spread out the numbers are. One very simple measure of dispersion is the range, defined as the difference between the highest and lowest value. In the test score example, the range is 96 – 80 = 16. This is not a very good indicator of dispersion because it is computed using only two numbers—the highest and lowest value.

The most common measure of dispersion is called the **standard deviation**. The computation is a bit complicated, but a good way to think about the standard deviation is that it is similar to the average amount each of the numbers differs from the mean, treating each of the differences as a positive number. The actual computation of a standard deviation is to take the difference of each number from the mean, square all the differences (that makes all the values positive), add up all the squared differences, divide by the number of values, minus one, and then take the square root of this value. Because this calculation is a lot of work, we will let the computer do the calculation rather than doing it by hand.

Figure 1.1 below shows part of the output from SAS when you ask it to compute descriptive statistics on the five test scores.



Basic Statistical Measures

Location		Variability	
Mean	88.80000	Std Deviation	7.56307
Median	90.00000	Variance	57.20000
Mode	96.00000	Range	16.00000
		Interquartile Range	14.00000

Figure 1.1: Example of Output from SAS Studio

This shows three measures of location and several measures of dispersion (labeled Variability in the output). The value labeled “Std Deviation” is the standard deviation described previously, and the range is the same value that you calculated. The variance is the standard deviation squared, and it is used in many of the statistical tests that we discuss in this book.

Descriptive statistics includes many graphical techniques such as histograms and scatter plots that you will learn about in the chapter on SAS Studio descriptive statistics.

Inferential Statistics

Let’s imagine an experiment where you want to test if drinking regular coffee has an effect on heart rate. You want to do this experiment because you believe caffeine might increase heart rate, but you are not sure. To start, you get some volunteers who are willing to drink regular coffee or decaf coffee and have their heart rates measured. The reason for including decaf coffee in the experiment is so that you can separate the placebo effect from a possible real effect. Because some of the volunteers might have a preconceived notion that coffee will increase their heart rate, their heart rate might increase because of a psychological reason, rather than the chemical effect of caffeine in the coffee.

You divide your 20 volunteers into two groups—to drink regular or decaf coffee. This is done in a random fashion, and neither the volunteers nor the person measuring the heart rates knows whether the person is drinking regular or decaf coffee. This type of experiment is referred to as a **double-blind, placebo-controlled, clinical trial**. We will discuss this design and several others in the next chapter.

Suppose the mean heart rate in the regular coffee group is 76 and the mean heart rate in the decaf (placebo) group is 72. Can you conclude that caffeine increases heart rate? The answer is "maybe." Why is that? Suppose that caffeine had no effect on heart rate (this is called the **null hypothesis**). If that were true, and you measured the mean heart rate in two groups of 10 subjects, you would still expect the two means to differ somewhat due to chance or natural variation. What a statistical test does is to compute the probability that you would obtain a difference as large or larger than you measured (4 points in this example) by chance alone if the null hypothesis were true. Statisticians typically call a difference statistically significant if the probability of obtaining the difference by chance is less than 5%. Be really careful here. The term significant is also used by non-statisticians to mean important. In a statistical setting, significant only means that the probability of getting the result by chance is less than 5% or some other number that you specified before the experiment began. Because statisticians like to use Greek letters for lots of things, the predefined **significance level** is called alpha (α).

Now for some terminology: You already know about the null hypothesis and a significance level. If caffeine had no effect on heart rate, what is the probability that you would see a difference of 4 points or more by chance alone? This probability is called the ***p*-value**. If the p-value is less than alpha, you reject the null hypothesis and accept the alternate hypothesis. The **alternate hypothesis** in this example is that caffeine does affect a person's heart rate. Most studies will reject the null hypothesis whether the difference is positive or negative. As strange as this sounds, this means that if the decaf group had a significantly higher heart rate than the regular coffee group, you would also reject the null hypothesis. The reason you ran this experiment was that you expected that caffeine would increase heart rate. If this was a well-established fact, supported by several clinical trials, there would be no need to conduct the study—if the effect of caffeine on heart rate was never tested, you need to account for the possibility that it could either increase or decrease heart rate. Looking for a difference in either direction is called a **2-tailed test** or a **non-directional test**.

As you saw in this example, it is possible for the null hypothesis to be true (caffeine has no effect on heart rate) and, by chance, you reject the null hypothesis and say that caffeine does affect heart rate. This is called a **type I error** (a **false positive** result). If there is something called a type I error, there is probably something called a **type II error**—and there is. This other error occurs when the alternate hypothesis is actually true (caffeine increases heart rate) but you fail to reject the null hypothesis. How could this happen? The most common reason for a type II error is that the experimenter did not have a large enough **sample** (usually a group of subjects). This makes sense: If you had only a few people in each group, it is easy to see that the means of the two groups would be different. If you had several thousand people in each group, and caffeine really had an effect on heart rate, you would be pretty sure of confirming the fact. Just as you set a significance level before you started the experiment (the 5% value that statisticians call alpha), you can also compute the probability that you make the wrong decision and claim that caffeine has no effect on heart rate when it really does (a **false negative**). The probability of a type II error is called beta (β), (more Greek).

Rather than think about the probability of a false negative, it is easier to think about the probability of a true positive. This probability is called **power**, and it is computed as 1 – beta. The last chapter of this book shows how to compute power for different statistical tests. Typically, the only way to increase power is to have a larger sample size.

Before we leave this chapter, there are two more terms that you should be familiar with. Remember in our coffee experiment, we assigned people to drink regular or decaf coffee. The 10 people in each group is called a **sample**. When you do your experiment and state your conclusion, you are not merely making a statement about your sample. You are saying that anyone who drinks regular coffee will have a higher heart rate that you estimate to be 4 points. You are making a statement about anyone who drinks regular or decaf coffee, and the name given to this theoretical group of people that you are making conclusions about is called a **population**. In practice, you define your population (everyone in the US or the world, for example), you take samples from this population, and make inferences about the effect your intervention on an outcome. That is why this branch of statistics is called **inferential statistics**.

Summary of Statistical Terms

- Measures of central tendency – statistics such as a mean or median that describe the center of a group of data values.
- Dispersion – a measure that describes how spread out the data values are. The most common measure of dispersion is the standard deviation.
- Sample – the group of subjects on whom you are conducting your experiment.
- Population – a theoretical group of subjects on whom you make inferences, based on the results from your sample.
- Type I error – a false positive result from a study. For example, concluding a drug or treatment works when it does not.
- Alpha (α) – the significance level. It is the probability that you are willing to accept for a type I error, usually set at .05.
- *p*-value – the probability of making a false positive (type I) error. If the *p*-value is less than alpha, you declare the results as significant. Some researchers prefer to just list the p-value and not set a specific alpha level.
- Type II error – a false negative result. For example, you have a drug or treatment that works but the results from the study is not significant (the probability of obtaining your result is greater than alpha.)

- Beta (β) – the probability of getting a false negative (type II) result.
- Power – the probability of a positive result when there truly is an effect. For example, you claim that your drug or treatment is better than a placebo or standard treatment, and you are correct in this decision. The power is the probability of the study rejecting the null hypothesis (finding the effect). Typically, powers of 80% or higher are considered acceptable. Large expensive studies might require powers of 90% or even 95%.