Essentials of Mathematical Probability and Statistics

Essentials of Mathematical Probability and Statistics

John Travis
Mississippi College

[3] November 2, 2020

John Travis grew up in Mississippi and did his graduate work at the University of Tennessee and Mississippi State University. As a numerical analyst, since 1988 he has been a professor of mathematics at his undergraduate alma mater Mississippi College where he currently serves as Professor of Mathematics.

John is married to Ruth Page Travis and together they have three unique daughters. All are good at mathematics.

You can find him playing racquetball or guitar but not generally at the same time. He is also an active supporter and organizer for the opensouce online homework system WeBWorK.

Preface

This text is intended for a one-semester course in probability and statistics that presumes calculus knowledge up to integration techniques. It is perhaps helpful if a student has already been exposed to sequences and series but much of what is needed is reviewed in the text.

An interactive version of this text is available at

http://math.mc.edu/travis/mathbook/Probability/Essentials_Probability_And_Statistics.html

http://math.mc.edu/travis/mathbook/Probability/Essentials_Probability_And_Statistics.html

and an pdf version with active links is available at

http://math.mc.edu/travis/mathbook/Probability/Essentials_Probability_And_Statistics.pdf

http://math.mc.edu/travis/mathbook/Probability/Essentials_Probability_And_Statistics.pdf

.

A collection of WeBWorK online homework problems are available to correlate with the material in this text. Copies of these sets of problems are available by contacting the author. These exercises are considered an integral part when using this text although a static version of these is included in the appendix.

To successfully utilize this text, a student should review the requisite material and perhaps review the proofs and derivations if appropriate. While moving through the text, the student should review each of the examples and then attempt each of the interactive WeBWorK exercises. Whenever an interactive cell comes up, the student should play around with the cell and perhaps change the input data as appropriate to experiment. When a section is completed, a student should work the WeBWorK exercises (not part of this text) or some other exercises provided by the instructor and attempt the exercises provided in this text...many of which are famous examples or exercises that might have special significance. Some are (of course) just easy and most of the textbook exercises have solutions provided.

WeBWorK (webwork.maa.org) is an open-source online homework system for math and science courses. WeBWorK is supported by the MAA and the NSF and comes with a Open Problem Library (OPL) of over 35,000 homework problems. Problems in the OPL target most lower division undergraduate math courses and some advanced courses. Supported courses include college algebra, discrete mathematics, probability and statistics, single and multivariable calculus, differential equations, linear algebra and complex analysis.

Sage (sagemath.org) is a free, open-source, software system for advanced mathematics, which is ideal for assisting with a study of abstract algebra. Sage can be used either on your own computer, a local server, or on SageMathCloud (https://cloud.sagemath.com).

R (https://www.r-project.org/) is a programming language and free software environment for statistical computing and graphics supported by the R Foundation for Statistical Computing. The R language is widely used among statisticians and data miners for developing statistical software and data analysis. In this text, the sage cell is used also for interactive computations related to R.

John Travis
Clinton, Mississippi 2015-2019

Contents

Chapter 1

Statistical Measures

1.1 Making Inferences

To compute your final grade in a class your teacher will likely consider the scores you have earned on various assignments and examinations completed during the duration of the course. However, she ultimately will likely be required to assign some numerical score indicating your level of success in the course. One grade to rule them all. This final grade can only be one value and it would make sense that the grade be a reflection of your work on these tasks. So, what is a fair way for your teacher to complete this task?

Through this process, you will also often need to take into account whether that data set is the entire list of possibilities--known as the population--or just a subset of that population perhaps obtained by taking repeated measurements --that is, a sample.

In general, it is often useful to make decisions using quantitative data but making those decisions can be somewhat arbitrary without a mathematical basis supporting those decisions. In this chapter, you will consider a number of ways to use point values to represent a given set of data. Each of these quantitative metrics will be called a ”statistical measure” and will, in some fashion, describe using one number some property of the entire data set. Such measures are part of what is known as ”descriptive statistics”. Later, you will learn about how other metrics can be used to predict properties of the underlying situation. Doing this is part of what is known as ”inferential statistics”.

So, let’s go and hopefully you will in some measure enjoy the ride!

1.2 Measurement Scales

In creating statistical measures, you might want to consider one of the following general types.

- Nominal measures - In this case, data falls into mutually exclusive and exhaustive categories for which the numerical value is only used for identification purposes. For example, assigning Male = 1, Female = -1.

- Ordinal measures - In this case, data consists of discrete numerical values which can be ranked from lowest to highest or vice versa. For example, your grades in a number of classes are used to compute your GPA--which is a single number.

- Interval measures - In this case, data possesses an order and where the distance between data values is of significance. For example, heights and weights.
- Ratio measures - In this case, data can be expressed as a position in some interval and where ratios between observations have meaning. For example, percentile rankings

In the subsequent sections of this chapter, you will see that a number of different measures are available for most data sets. Determining which ”correct” measure to use for describing any given data set will depend the actual situation surrounding the collection of the data.

Below are a couple of active WeBWorK exercises that you can continue to attempt until you get the correct answer. Try to avoid clicking on the provided solution unless you are hopelessly stuck on exercises like this throughout the text.

Checkpoint 1.2.1 WebWork - Types of Data. Before leaving a particular restaurant, patrons are asked to respond to a questionaire containing the questions given below. For each question, indicate (using the pull-down menu) whether the possible responses are Interval, Nominal, or Ordinal.

1. Would your overall rating of this restaurant be excellent, good, fair, or poor?
2. Would you recommend this restaurant to a friend?
3. Have you eaten at this restaurant previously?
4. Do you consider our prices to be high, average, or low?

Checkpoint 1.2.2 WebWork - Types of Data Again. Determine whether the following possible responses should be classified as ratio, interval, nominal or ordinal data.

1. The letter grades received by students in a computer science class
2. The number of students in a statistics course
3. Your hometown
4. The college (Arts and Science, Business, etc.) you are enrolled in

1.3 Statistical Measures of Position

Given a collection of data, sorting the data may provide several useful descriptors. When sorting data, you can easily use something like a spreadsheet for larger data sets but in this section you will also see there are ways to perform a sort by hand. In either case, statistical measures of position generally involve very little computational work once the data is sorted and take into account only the order of the data from lowest to highest. To assist with notation, we will generally use x-values to represent the original raw data and y-values to represent that same data when ordered with the subscript indicating the positional placement.

Definition 1.3.1 Order Statistic. From the data set

$$x_1, x_2, ..., x_n,$$

assume that when sorted it is denoted

$$y_1, y_2, ..., y_n$$

where

$$y_1 \leq y_2 \leq ... \leq y_n.$$

Then, y_k is known as the kth order statistic. ◊

Example 1.3.2 Age of Presidents - order statistics. The age at inauguration for presidents from 1981-2019 gives the data

$$x_1 = 69, x_2 = 64, x_3 = 46, x_4 = 54, x_5 = 47, x_6 = 70$$

(Reagan, Bush, Clinton, Bush, Obama, Trump). For this data, the order statistics are denoted

$$y_1 = 46, y_2 = 47, y_3 = 54, y_4 = 64, y_5 = 69, y_6 = 70.$$

□

Once the data is sorted, it should be very easy for you to locate the smallest and largest values.

Definition 1.3.3 Minimum/Maximum:. For a given data set, the smallest and largest values are known as the minimum and maximum, respectively. In our notation and presuming a data set of size n, the minimum $= y_1$ and the maximum $= y_n$ ◊

Example 1.3.4 Age of Presidents - Minimum/Maximum. Using the President inauguration data 1.3.2, minimum $= y_1 = 46$ and maximum $= y_6 = 70$. □

A value that separates ordered data into two groups with a desired percentage on each side is called a percentile. There are multiple ways that have been created that achieve this goal. In this text we present two and will consistently use the first one presented below. For each, in general, a given percentile is a numerical value at which approximately a given percentage of the data is smaller.

The definition presented below provides for a unique measure for each unique value of s that corresponds to the PERCENTILE.EXC macro in Excel. This version starts by computing $(n + 1)s$ where $0 < s < 1$ and using this to linearly interpolate between two adjacent entries in the sorted list. Another option that corresponds to PERCENTILE.INC (and PERCENTILE) in Excel is to start with $(n - 1)p + 1$ for determining how to pick the two adjacent entries and then proceeding with linear interpolation. Again, the definition below utilizes the first approach.

Definition 1.3.5 Percentiles. For $0 < s < 1$ and for order statistics $y_1, y_2, ..., y_n$ define the 100s-th percentile to be

$$P^s = (1 - r)y_m + ry_{m+1}$$

where m is the integer part of $(n + 1)s$, namely

$$m = \lfloor (n + 1)s \rfloor$$

and

$$r = (n + 1)s - m,$$

the fractional part of $(n + 1)s$.

In Excel, this is PERCENTILE.EXC. ◊

Definition 1.3.6 Alternate Percentile Definition. For $0 < s < 1$ and for order statistics $y_1, y_2, ..., y_n$ define the 100s-th percentile to be

$$P^s = (1 - r)y_m + ry_{m+1}$$

where m is the integer part of $(n - 1)s + 1$, namely

$$m = \lfloor (n - 1)s + 1 \rfloor$$

and

$$r = (n - 1)s + 1 - m,$$

the fractional part of $(n - 1)s + 1$.

In Excel, this is PERCENTILE.INC or just PERCENTILE. ◊

Compute the following percentile values using the alternate formula 1.3.6.

Checkpoint 1.3.7 WeBWorK - Computing Percentiles. Consider the following data set:

47	37	30	65	20	38	37	45	59
49	53	21	23	37	49	20	62	62

Find the 15th and 89th percentiles for this data.

15th percentile = ____________

89th percentile = ____________

Answer 1. 20.85

Answer 2. 62

Example 1.3.8 Presidential Percentile. To compute, say, the 42nd percentile using the definition 1.3.5 for the President inauguration data presented earlier 1.3.2 consider s = 0.42. Since there are 6 numbers in our data set, then

$$(n + 1)s = 7 \cdot 0.42 = 2.94$$

and so m = 2 and r = 0.94. Thus, the percentile will lie between $y_2 = 47$ and $y_3 = 54$ and much closer to 54 than 47. Numerically

$$P^{0.42} = 0.06 \cdot 47 + 0.94 \cdot 54 = 53.58.$$

□

Both formula approaches for percentiles determine a weighted average between y_m and y_{m+1} which is unique for distinct values of p provided each of the data values are distinct. Note that if some of the y-values are equal then some of these averages might be averages of equal numbers and will therefore be the common value.

Some special percentiles are provided special names...

Definition 1.3.9 Quartiles. Given a sorted data set, the first, second, and third quartiles are the values of

$$Q_1 = P^{0.25}, Q_2 = P^{0.5}$$

and

$$Q_3 = P^{0.75}.$$

◊

It should be noted that many graphing calculators often compute quartiles using a straight average of two adjacent entries rather than by using the formula