Elements of Statistics

Elements of Statistics:

A Hands-on Primer

By

Raghubar D. Sharma

Cambridge
Scholars
Publishing

Elements of Statistics: A Hands-on Primer

By Raghubar D. Sharma

This book first published 2017

Cambridge Scholars Publishing

Lady Stephenson Library, Newcastle upon Tyne, NE6 2PA, UK

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

ISBN (10): 1-5275-0034-9
ISBN (13): 978-1-5275-0034-1

For Anjana, Vivek, Shilpa and Pratik

Dedicated to: My High School Mathematics Teacher
Mr. C.L. Sharma whose gentle teaching style instilled
love for mathematics in me and my brothers.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

Figures

Tables

INTRODUCTION

Statistics has become a form of logic or rhetoric that everyone needs to learn to navigate the modern world. Though this book is primarily aimed at undergraduate students who are required to take at least one compulsory statistics course before graduation, it is also a valuable guide for anyone who has little or no background in statistics and wants to become statistics literate. Without the pretensions of the famous book that the learning of statistics can be without tears or that you don't need to understand symbols, formulae, and equations, this book will prepare you to understand basic statistics and to complete your statistics course without anxiety. This book has been written with the conviction that you don't need to be a mathematician to learn statistics. It is a crucial resource for students taking a required statistics course who are intimidated by statistical symbols, formulae, and daunting equations.

The application of statistics in social research has become imperative. A gap usually exists between the time when students take their first statistics course and when they engage in their first serious research project (typically during an internship, or a research methods course, or their final year project/thesis). Because of this gap, students often don't remember basic statistics well enough to apply it effectively in their research. Hence, there is a need for a "desk reference," "refresher," or "core concept" text—an Elements of Statistics for burgeoning researchers, à la Strunk and White's Elements of Style. This book will serve as an excellent desk reference, refresher, or core concept text for the budding researcher. It will also be helpful while interning or working as a research assistant or research associate.

Those who feel left out when their colleagues, supervisors, or bosses use statistics will benefit from this book. This particular group of people has been on my mind for a long time. When I was employed as a workforce data analyst for an employment equity office in the early 1990s, I routinely prepared reports and presentations for senior management. The basic education of my manager was grade 11, and her boss, an assistant deputy minister (ADM), had grade 12 with an accounting certification. They both had missed the opportunity to learn statistics in school. One day, my manager returned from her boss's office after discussing a report I had prepared and told me that the ADM preferred "circles" not "hills" in the report. After some pondering, I realized that her boss preferred pie charts over bar graphs. This report was on employment equity designated groups, which included Aboriginal peoples, women, and persons with disabilities—where a respondent could be counted more than once. For example, the same person could be counted as a woman, an Aboriginal person, and a person with a disability. In Chapter 3, I explain that a pie chart is not appropriate where there is double-counting of respondents or observations.

Because my manager had little to no statistical literacy, it would have been futile to explain to her why pie charts are not a good idea for this kind of data. But since then, I have felt that I should write a book on statistics that could help people like my manager and her boss. Today, even if you work in non-statistical areas such as policy, communications, and journalism, you need to have some knowledge of statistics. Nowadays, statistical literacy is as important as literacy itself. This book is written in a self-help, hands-on learning style so the reader can easily attain the skills needed to achieve a basic understanding of statistics and be comfortable with presentations loaded with statistics.

This book follows an easy-to-comprehend format. It gives a strong foundation in the basics, while calculations elaborate on the basics in sequences designed for students and general readers who have never taken a statistics course. Simply put, it is a hands-on primer. The idea is that when you're reading it, you won't need a calculator or a computer.

The book contains 17 chapters along with statistical tables in the appendix. Chapter 1 provides a refresher on basic math and statistical symbols, while Chapter 2 discusses the levels of data measurement and types of variables. Chapter 3 deals with the visual representation of data. It also cautions researchers on the use and misuse of graphics. Chapters 4 and 5 discuss the measures of central tendency and variability, respectively. Chapter 6 familiarizes the reader with the basic concepts of probability. As researchers are always required to work with samples, Chapter 7 is devoted to methods for selecting appropriate samples. The next three chapters deal with important concepts that a researcher must know before embarking on applying a statistical technique to data. Chapter 8 discusses the sampling distribution and the normal curve, particularly with respect to generalizing from a sample to the population. Chapter 9 elucidates the relationship between the normal distribution and standard scores. It also includes conversion of raw scores into standard scores, an essential requirement for comparative research. Chapter 10 examines relationships between variables as well as the procedure and essential concepts for testing a hypothesis.

Chapters 11, 12, and 13 are devoted to tests of significance for the nominal-, the ordinal-, and the interval/ratio-level variables, respectively. Chapter 14 discusses the correlation coefficient, and Chapter 15 is devoted to the statistical power of a statistical test. Chapter 16 provides the basics on analysis of variance (ANOVA), and Chapter 17 focuses on regression. Thus, the book ends with a comprehensive survey of applied statistics and fills the lacunae left by the majority of statistics books.

All exercises and examples in the book have been developed by the author. Due care has been taken to credit the sources used in the book. Any omission in referencing is, of course, unintentional and once pointed out will be rectified in the next edition.

CHAPTER ONE

BASIC MATH AND SYMBOLS USED IN STATISTICS

Learning Objectives

In order to learn statistics, you need some knowledge of basic mathematics, which most of you have already acquired during your grade and high school years. Because memories tend to fade over time, this chapter serves as a refresher of basic mathematical operations. It also introduces some commonly used statistical symbols. Specifically, you will learn about:

- the order of operations;
- fractions, decimals, exponents, and logarithms; and
- the most frequently used statistical symbols.

Introduction

H.G. Wells once said, "Statistical thinking one day will be as necessary for the efficient citizenship as the ability to read and write."[1] Arguably, that day has arrived, as we are bombarded daily with statistics from television commentators, newspapers, popular multimedia, and advertising billboards. Terms and phrases such as *batting average*, *the chance of winning an election*, *outliers*, and *median income* are all statistics that are routinely used in popular media. Yet many people, including university students, think that statistics is not relevant to their learning. Statistics is not taken seriously because we continuously hear that statistics is not an objective science. Yet, experts in various fields often use statistics to

[1] Wells, H.G. 1903. *Mankind in the Making*, London: Chapman & Hall, Page 204.

refute each other's claims (Haan[2], 2008). Politicians frequently use statistics during election campaign debates to score points over one another. The idea that statistics can be used to lie was made popular by Darrell Huff[3] in 1954. In fact, Huff's book, *How to Lie with Statistics*, does not propagate lying with statistics; it illustrates how statistics can be misused. The following example shows that the *average* is meaningless without reference to the spread (variability) of data. Let's say that five students in a class brought $5, $6, $2, $2, and $60 each to buy lunch. If you used only the average, you would say that on average a student brought $15 for lunch. The use of the average suggests that students brought more-than-sufficient money for lunch, whereas at least two, or 40% of the students had enough money only to buy a soft drink. In statistical terms, the $60 is an *outlier,* which increases the spread of data. Using an *average* without considering the spread of data around the *mean* value can create a misleading impression.

Another misconception is that to learn basic statistics you need to know advanced mathematics, when, in fact, knowledge of basic math is sufficient to understand most statistics. In the next section, we review the necessary math needed to learn statistics. Though some equations in statistics textbooks may look daunting, you can easily understand and apply basic statistics without attempting to solve these intimidating equations.

Basic Math Needed to Learn Statistics

You can learn to apply advanced-level statistics without learning advanced-level math. These days, most statistical calculations are done by computer software. You can learn to interpret statistics produced by statistical software such as SPSS and SAS without learning advanced-level mathematics. Even if your basic math is rusty, the following review will be sufficient to learn essential statistics.

The Order of Operations

The order of operations is the basic principle that governs the sequence of operations, which is bracket (parenthesis), exponent, division, multiplication,

[2] Haan, Michael. *Introduction to Statistics for Canadian Social Scientists*, Don Mills, Ontario: Oxford University Press, 2008.

[3] Huff, Darrell, and Irving Geis (illustrator). *How to Lie with Statistics,* New York: W.W. Norton and Company Inc., 1954.

addition, and subtraction. You may be familiar with **BEDMAS**, an acronym you likely learned in grade or high school to memorize the sequence of operations, where:

B stands for **Bracket**;
E stands for **Exponent**;
D stands for **Division**;
M stands for **Multiplication**;
A stands for **Addition;** and
S stands for **Subtraction.**

Let's say we have to solve the following equation to find the value of *x*:

$$x = 2 \times (3 + 5)^2 \div 4 + 6 - 7$$

Applying BEDMAS, we will first find a solution for the equation in the **bracket**: (3 + 5) = 8.

Next, we will find the value of the **exponent** of 8 = 8^2 = 64.

The next step will be **division**: 64 ÷ 4 = 16.

After division, the next operation is **multiplication**: 2 × 16 = 32.

After multiplication, the next step is **addition**: 32 + 6 = 38.

The final step is **subtraction**: 38 – 7 = **31**.

Note: If there is more than one operation within the bracket, follow the same sequence. For example, if the equation in the bracket is (5 + 6 ÷ 2 × 4 – 1), then perform the operations in the following sequence: first, divide 6 by 2 = 3; next, multiply 3 by 4 = 12; then, add 5 to 12 = 17, and then subtract 1 from 17 = 16. The value in the bracket is 16.

Decimals and Fractions

It is necessary to be comfortable with decimals because many statistics and statistical relationships are expressed in decimals. You can use a calculator to calculate a decimal from a fraction. One way to convert a fraction to a decimal is to divide the *numerator* by the *denominator*. For example, $\frac{8}{16}$ is equal to 8 divided by 16, which is equal to 0.5. A fraction can also be expressed as a percentage by converting the fraction to a decimal and then multiplying it by 100: 0.5 × 100 = 50%.

You will also learn that the concept of probability is central to statistical prediction. Probability is expressed in decimal points, and a chance of an event happening is expressed in percentage. For example, if there is a 30% chance of catching a fish from a river, you could say that the probability of catching a fish is 0.3. The following is a refresher from grade or high school mathematics on fractions, decimals, and percentages, which you can calculate by hand or with a basic calculator.

Finding the Percent of a Number

Example: *to find out what is 92% of 28.*

- Multiply the number by the percent: 28 × 92 = 2576.
- Divide the total by 100: 2576 ÷ 100.
- To find out the answer, move the decimal point two places to the left: **25.76**.

Finding Percentage

Example: *to find out what percent is 28 of 92?*

- Divide the first number by the second: 28 ÷ 92 = 0.3043
- Multiply the answer by 100: 0.3043 × 100
- Move the decimal point two places to the right: **30.43%**.

Converting a Fraction to a Decimal

Example: *Convert* $\frac{1}{3}$ *to a decimal.*

- Divide the numerator of the fraction by the denominator: 1 ÷ 3 = **0.3333**.

Converting a Fraction to a Percent

- After converting a fraction to a decimal, simply multiply by 100 and move the decimal point two places to the right: 0.3333 × 100 = **33.33%**.

Converting a Percent to a Fraction

Example: *Convert 75% to a fraction.*

- Remove the percent sign from 75.
- Make a fraction with the percent as the numerator and 100 as the denominator: $\frac{75}{100}$ = 75 ÷ 100 **= 0.75**.

Converting a Decimal to Percent

Example: *Convert 0.75 to a percent.*

- Multiply the decimal by 100: 0.75 × 100 = 75.
- Add a percent sign to 75: **75%**.

Converting a Percent to a Decimal

Example: Convert 75% to a decimal.

- Divide the percent by 100: 75 ÷ 100 = **0.75**.

Rounding Decimals

The rounding of decimal points makes calculation easier and presentation of the numbers clearer. One reason we usually need to round a number is because a decimal may extend endlessly. For example, 1/3 results in 0.3333333333. In the rounding of decimals, we need to consider two questions.

1. How many places should we carry the decimal point?
2. How do we decide that the last number reflects the remainder?

The answer to the first question is that it depends on the type of data. Some demographic data, such as survival rates, may extend to sixth decimal point; whereas in many other situations, you might decide to keep only one decimal place. Conventionally, we round to two decimal places.

The answer to the second question is that we retain the value of the last decimal place if the value next to the retained decimal place number is less than 5. For example, 2.344 is rounded to 2.34. We increase the value of the retained decimal place by 1 if the next decimal place is 5 or greater. For example, 3.765 is rounded to 3.77.

Some statisticians suggest that in a dataset, the number ending with 5 after the decimal point should be rounded up for one-half of the time and down for the other half of the time. For example, say the first number ending with 5 after the decimal point is 2.235; it is rounded down to 2.23. Say the second number ending with 5 after the decimal point is 6.475; it is rounded up to 6.48. **However, I would suggest that for numbers ending with 5 after the decimal point, you round up every time**.

Truncation

When we retain the decimal place just as it is, without changing the value of the decimal place, it is called *truncation*. Some computer programs truncate the numbers without changing the value of the retained decimal place; for example, 2.334 and 2.337 are both retained as 2.33. In other words, both numbers are truncated to 2.33.

Exponents

In statistics, exponents are used quite routinely; therefore, it is important to familiarize yourself with them. Basically, an exponent indicates the number of times a numeral should be multiplied with another. For example, 2^3 means that 2 is multiplied 3 times: $2 \times 2 \times 2$. In the expression 2^3, the 2 is called the *base* and the 3 is called the *exponent*. It is also commonly called *2 raised to the power of 3*. Here, the 2 is called a *base* and the 3 a *power*.

Multiplication and Division of Two Exponents with Identical Bases

If two exponents with identical bases are multiplied, the rule is to add the exponents. For example, to multiply 2^4 by 2^3, you add the exponents 4 and 3. Thus, $2^4 \times 2^3$ becomes $2^{(4+3)}$, or 2^7. Its value is: $2 \times 2 \times 2 \times 2 \times 2 \times 2 \times 2$ = **128**.

If two exponents with identical bases are divided, the rule is to subtract the exponents. For example, to divide 2^5 by 2^3, you subtract the exponent 3 from 5. Thus, $2^5 \div 2^3$ becomes $2^{(5-3)}$, or 2^2. Its value is: 2×2 = **4**.

Square Root

The square root of a number is the reverse operation of a square of that number. For example:

Square of $n = n^2$
If $n = 3$, then $3^2 = 3 \times 3 =$ **9.**
The square root of $n = \sqrt{n}$.
If $n = 9$, then $\sqrt{9} = \sqrt{3 \times 3} =$ **3.**

Logarithms

You may have learned in high school that a logarithm is a special type of exponent. The base of a logarithm is either 10 or 2.718. When the base is 10, it is called the *common logarithm;* when the base is 2.718, it is called the *natural logarithm.* What this means is that if we raise 10 to the power of 3 (10^3), the answer is $10 \times 10 \times 10 =$ **1000**, and, hence, the common log of 1000 is 3. Similarly, if we raise 2.718 to the power of 3 (2.718^3), the answer is about 20 ($2.718 \times 2.718 \times 2.718 \approx 20.01$), which means the natural log of 20 will be about 3, or, to be precise, 2.9957. The $\approx$ stands for "approximately equal to."

Some Common Symbols Used in Statistics

The symbols used in statistical equations may intimidate a non-mathematical person. The fear of symbols sometimes disheartens a person to learn statistics. This fear is unnecessary. Once you understand the meaning of symbols, the fear disappears and the learning of statistics becomes easy.

Most statistics textbooks use X and Y as symbols for *variables*. Basically, a variable is a characteristic (such as sex or social class) or a quantity (such as age or income). A variable varies between its categories. For example, sex can take a value of a male or a female, and class might vary between the lower, middle, or upper class. Similarly, age might take any value from one day old to 100 years old, and income could vary from 0 dollars to billions of dollars. The symbol N is used for the number of persons or the number of cases in a population, and the symbol n is used for the number of persons or the number of cases in a sample. Generally, uppercase letters (X, Y, Z) are used to represent population characteristics, and lowercase letters (x, y, z) are used to denote sample characteristics.

The most dreaded symbol for a person unfamiliar with statistics is a Greek-alphabet uppercase sigma, which is written as ∑ and denotes the adding up or summing up of numbers. For example, $\sum(X_1, X_2, X3)$ indicates that we are adding quantities represented by the symbols X_1, X_2, and X_3. Simply put, if $X_1 = 2$, $X_2 = 3$, and $X_3 = 4$, then $\sum(X_1, X_2, X_3) = 2 + 3 + 4 = \mathbf{9}$. It is that simple. You will see sigma written as follows:

$$\sum_{i=1}^{N} X_i$$

In the above example, $N = 3$. X_i indicates that X takes i values. Because X takes three values (2, 3, and 4), in this example i is equal to 3.

The summation sign, ∑, is the most frequently used symbol in statistics. The following rules will be helpful to understand its use.

Summation (∑) and Constant (*c*)

Written in symbols: $\sum c = N \times c$.
$\sum c$ means that the sum of constants is equal to the number of times a constant appears in the series multiplied by the value of the constant:

if $c = 10$ and $N = 6$, then $N \times c = 10 \times 6 = 60$. It is the same as:

$$10 + 10 + 10 + 10 + 10 + 10 = 60.$$

Summation (∑), Constant (*c*), and a Variable

Written in symbols: $\sum cX_i$
The symbols above suggest that you first multiply each value of variable X with the constant and then add them up. If $c = 10$, $X_1 = 3$, $X_2 = 4$, and $X_3 = 5$, then,

$$\sum \mathrm{c}X_i = 10 \times 3 + 10 \times 4 + 10 \times 5 = 30 + 40 + 50 = \mathbf{120}.$$

Summation (∑) and Two Variables (X and Y)

a. $\sum(X - Y)^2$ (Sum of Squared Deviations of X and Y)
b. $\sum XY$ (Sum of Products of X and Y)
c. $\sum X \sum Y$ (Product of Summations of X and Y)

Table 1-1 provides the calculations for a., b., and c. for two variables X and Y with three values.