INTRODUCTION TO STATISTICS

Chapter One

Basics of Statistics

Unit Introduction

There are several uses for statistics across a wide range of different sectors. Statistics, generally, can be defined as an approach for receiving, interpreting, evaluating, and deriving conclusions from data. In other terms, statistics is the process that mathematicians and scientists have devised for analyzing and generating findings from data that has been acquired. Statistics include every aspect of data receiving, analyzing, evaluating, and presenting, as well as the meticulous planning that goes into each process. A collection of techniques for receiving and interpreting data makes up statistics (Davies & Fearn, 2006).

Remember:

Statistics is the study of getting knowledge using categorical and numerical data.

It should be evident from the preceding that statistics include much more than simply number tabulation and graphical representation.

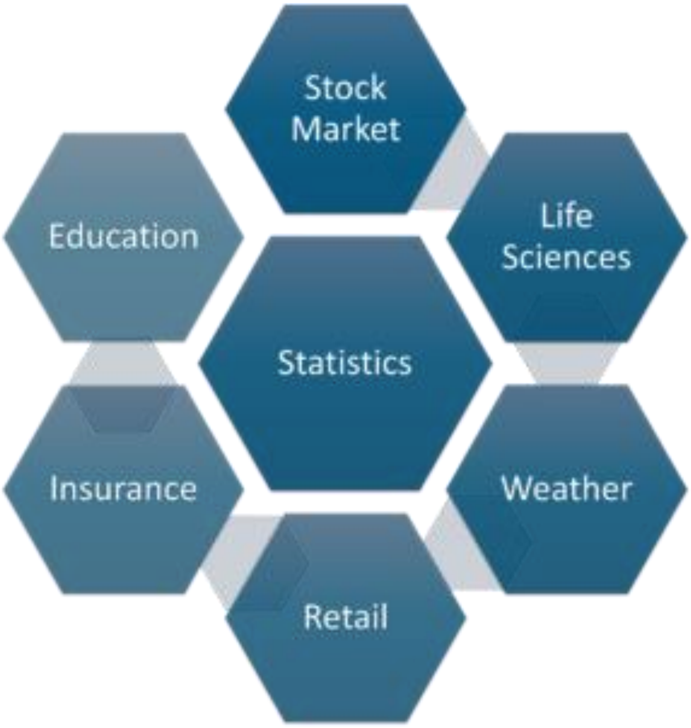

Figure 1.1. Diagram representing various applications of Statistics and its Fields (Source: Zulaikha Lateef, Creative Commons License)

Statistics is also the discipline of coping with ambiguous phenomena and situations. Statistics is successfully used in practice to investigate various topics, including the efficacy of medical interventions, audience responses to television advertising, views of young people regarding marriage and sex, and much more. Today, statistics are undoubtedly employed in all branches of science (Darling, 2019).

Learning Objectives

At the end of this chapter, readers will be able to learn:

1. Basic understanding of statistics and its interdisciplinary nature
2. Various definitions of statistics
3. Basic concepts in Statistics, including data, population, sample, parameter, and variables
4. The concept of greater statistics
5. Some daily-life examples of statistics, including spam filtering, the Sally Clark case, and inflation, among others

Key Terms

1. Numerical
2. Sequence
3. Population
4. Sample
5. Parameters
6. Variables
7. Scales
8. Mode
9. Mean
10. Median
11. Range

1.1. Some Definitions of Statistics

One helpful definition of statistics is the technology used to derive significance from the information. No definition, though, is absolute. The basis of many uses of statistics, chance, and possibility are specifically left out of this definition. So, technology for managing uncertainty could be a different working definition (Proschan & Sethuraman, 1977). However, other definitions or more exact definitions might place a greater emphasis on the roles that statistics perform. Therefore, we may say that statistics is the primary discipline for forming conclusions about the unknown, forecasting the future, or creating meaningful summaries of information. Although diverse applications will result in varied representations, these definitions broadly cover the basis of the discipline.

Examples of statistics applications include prediction, real-time supervision, fraud prevention, population count, and study of DNA sequences, but they may call for very different approaches and equipment (Pickands 1975). It is important to note that we purposefully picked the word "technology" for these definitions instead of the word "science." Statistics is an application of our knowledge of how to retrieve data-based information and our comprehension of variability. Technology is the use of science and its findings. However, the term "science" is frequently used to describe statistics. Statistical Science is the name of one of the most interesting statistical journals.

We have discussed the field of statistics thus far in this book, most notably in the paragraph just before, but the word "statistics" also has another interpretation: it represents the plural of "statistic." A statistic is a summarization of numerical data—for instance, a summary of statistics outlining a population size, fertility rate, or violence rate. Therefore, in a certain sense, this book is about specific numerical values (Dehling, & Taqqu, 1989). But it is about a lot more than that. It concerns how to receive, manage, examine, and draw conclusions from those numerical values. It has to do with technology in general. The reader will therefore be let down if they were expecting to see tables of figures in this book, such as "sports statistics."

However, a reader interested in learning how corporations make choices, astronomers explore different types of stars, medical researchers determine the genes linked to a specific disease, insurance providers choose how much to charge for a premium, spam screens are built to keep offensive advertisements out of your e-mail inbox, and so on, will be commended. This clarifies why the word "statistics" can be used both single and plural: statistics is a subject, but considerable numbers constitute statistics (Baik et al., 2005).

There goes the word "statistics" yet again. The word "data" was also included in my initial working definition. Data is the plural form of the Latin term datum, which means "something provided," and comes from the verb dare, which means "to give." It follows that it must be used as a plural noun instead of a singular one: "the data are poor," and "these data demonstrate that," as opposed to "the data is bad" and "this data shows that." The English language does, though, evolve through time. The term "data" is used more frequently these days to describe a spectrum, as in "the water is wet" instead of "the water is wet." Our propensity is to use anything that sounds most euphonious in the given situation. This usually implies adhering to the plural form (Cuevas, 2014).

Data represent numerical outcomes of measurements, counts, or other operations in most cases. Such information might be viewed as offering a condensed portrayal of whatever we are studying. If we are interested in schoolchildren's academic performance and suitability for different occupations, we may decide to look at the statistics revealing their performance on other tests and examinations. These figures would indicate their aptitudes and preferences. The portrayal would undoubtedly be imperfect. A low grade can result from being unwell at the time of the test. A missing value only indicates that the person did not take the test, not much about their aptitude. Later, we will talk more about the quality of data. It is essential due to the general rule (applied to all aspects of life, not just statistics) that if we deal with inadequate data, poor results will arise. Statisticians can accomplish remarkable things when deciphering data, but they cannot help tremendously (Sanathanan, 1972).

Of course, many instances do not seem to generate numerical data directly. There is a great deal of raw data that looks to be in the case of pictures, phrases, or even electronic or auditory signals. Thus, numbers do not reflect in satellite photographs of fields or rainforest coverage, verbal explanations of adverse effects from medicine use, or speech sounds. Moreover, a closer examination reveals that these things are converted into numerical figures or representations that can potentially be turned into numbers when they are recorded and measured. For instance, satellites and other images are made up of millions of tiny components called pixels, each characterized in terms of the (numerical) values of the many colors that make it up (Stephens, 1974).

Text can be analyzed to produce word counts or metrics of a word and phrase frequency; web search engines like Google employ this representation. The numerical values of the waveforms that make up the various components of speech represent spoken words. While not all data are numerical, most are converted into numerical form. Additionally, most statistics interact with numerical data.

People who use statistics in a biased manner frequently raise suspicion. People can choose to stress distinct summaries if there are several ways to summarize a piece of information, each looking at a somewhat different perspective. Crime statistics serve as a specific illustration. The British Crime Survey is perhaps the country's most significant source of crime data. This directly questions a sample of people about crimes they were the victim of over the past year to evaluate the degree of crimes. In contrast, the Recorded Crime Statistics series contains all crimes that must be reported to the Home Office that the police have documented. By definition, this does not include some infractions (Cavalier, 2011).

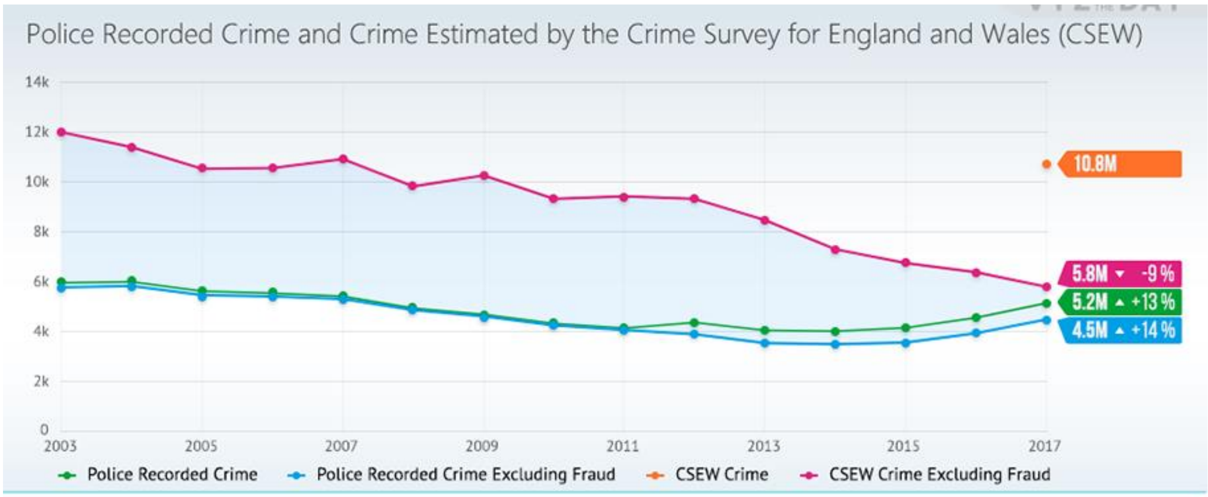

Figure 1.2. Graphical representation of Statistical Data for measuring Crime rate in the United Kingdom (Source: Knoema, Creative Commons License)

Of course, the fact that it does not include crimes never addressed by the police is more significant. With such variations, it is obvious why the data could fluctuate between the two sets, even to the point where specific crime categories might appear to rise over time according to one set of numbers while dropping to the other. The data on crime also show another possible reason to be skeptical of statistics. People may decide to focus on a metric when it is employed as a system performance indicator, increasing its value but at the expense of other system components. The chosen metric becomes meaningless as a gauge of system performance due to excessive improvement (Aki, 1986). For instance, by concentrating all of their efforts on it, the police may lower the rate of stealing at the expense of enabling other types of crime to increase. Consequently, the stealing rate is no longer applicable as a gauge for the crime. In honor of Charles Goodhart, a former Chief Adviser to the Bank of England, this occurrence has been given the title "Goodhart's law."

The primary point is that the issue does not lie in the statistics themselves but in how they are used and how true meaning is misunderstood. Perhaps being cautious of something we do not comprehend is quite normal. The sheer nature of scientific advancement raises another reason for skepticism in an effective manner (Tabachnick et al., 2007). As a result, we might read in the papers one day about scientific research that seems to demonstrate that a particular food type is dangerous for us and beneficial the next day. Consequently, this confuses and gives the impression that the scientists are uninformed and possibly unreliable.

This distrust partly extends to statistics because such scientific inquiries invariably rely on statistics. But novel discoveries that alter our thinking are fundamental to scientific progress. Further research may have made us realize that there are numerous types of fats, some advantageous and others harmful, contrary to what we may have previously believed about dietary fat. It is not unexpected that the initial investigations are contradictory and incongruent outcomes because the situation is more intricate than we initially anticipated. Fourth-level errors in understanding fundamental statistics are a source of suspicion (Johnstone, 2001). The reader could use the following assertions as an experiment to attempt to determine what is odd about each one:

i. According to a report, screening programs are helpful because early diagnosis of a medical problem results in longer retention spans.

ii. We are informed that a stated price has previously been discounted by 25% for eligible clients; since we are not suitable, we must pay the additional 25%.

iii. According to a simple extrapolation of growth over the last 100 years, it is predicted that the average lifespan will reach 150 years in the following century.

iv. According to the information, "every year since 1950, the number of American youngsters shot to death has doubled."

Sometimes the misconceptions are not that simple, or at least they come from quite complex statistical ideas. After more than a century of growth, it would be remarkable if statistics did not contain any profoundly paradoxical concepts (Johnstone, 2001). The Prosecutor's Fallacy is an example of one of these. It depicts a conflict between the likelihood that something—for example, the suspect is guilty—will be true if you have some proof, such as the suspect's gloves at the crime site, and the likelihood of discovering that proof if you think the suspect is guilty. We will look into this issue in more detail later because it occurs frequently and is not just in the courts (Collet, 2001).

If there is skepticism or suspicion of statistics, it is evident that the usage of those numbers, rather than their creation or calculation, is to blame. Blaming the subject or the statistician who derives the significance from the information is unfair. Instead, responsibility rests with those who intentionally misuse the data or do not comprehend what the stats are trying to tell them. We accuse a man who fired the gun, not the weapon itself, of killing the victim (Grégoire, 2016).