INTRODUCTION TO STATISTICS AND COMPUTATION WITH DATA

Rituparna Sen, notes by Ramdas Singh

Second Semester

List of Symbols

$\hat{\theta}$, the estimate of a variable θ.

Contents

Chapter 1

AN INTRODUCTION TO STATISTICS

January 2nd.

Commonly referred to as the science of data, *statistics* involves collecting, summarizing, presenting, and interpreting data. Randomness and variability in data neccessitate the use of statistics. If a process is deterministic, there is no real need of statistics.

Often in probability, we are given the probability of getting heads in a single coin toss and we may be tasked to find the probability of 4 heads appearing in 10 tosses. In contrast, statistics starts with observing 4 heads appearing in 10 tosses, and utilises this data to determine the probability of a head. If we are to determine the price of a house in city, factors to look for may include the city itself, the specific location and area, the kind of house, the square footage, the age of the house, and the change with time. Despite all this, there is still an element of randomness; it is not a true deterministic quantity.

1.1 Fundamental Elements of Statistics

We first discuss some fundamental elements.

- An *experimental unit*. It may be a singular item such a coin toss, or a single house as in the previous example(s).
- The *population*. The set of all experimental units. We may note that studying all experimental units is a population may not be possible.
- A *census* studies all the units in a population.
- A *sample* of the population. A subset of the population. It is ideally chosen in a way that represents the entire population. A true representative sample may not always be possible without the the subset being the entire population.
- A *variable*. For each experimental unit in the sample, we record the data on several variables. In the case of the prices of houses, the factors discussed are the variables.
- *Univariate* and *multivariate* samples. As expected, a univariate sample has only one variable per unit, and a multivariate one has multiple variables per unit. Two variables per unit in the sample is also often referred to as a *bivariate* sample.

1.2 Types of Statistics and Data

1.2.1 Descriptive Statistics

Often, in statistics, we use pictures, tables, and summary numbers to describe the data. R Studio (or just R) may be used to handle descriptive statistics.

1.2.2 Inferential Statistics

Here, we make statements about the population based on our sample observations. In the case of coin tossing, let us look at the population of infinite coin tosses, where the probability of a heads is an unknown p. A census of this population is not possible, so we may take a sample of 10 tosses. Suppose we get X heads in this sample. We know that X follows a binomial distribution as $X \sim \text{Bin}(10, p)$. Here, we describe a new varaible called the *estimate*, $\hat{p}$. In this case, we find that $\hat{p} = X/10$. This is how we work in statistics; we deal with an unknown variable of the population, say θ, by looking at a sample and describing an estimate $\hat{\theta}$.

We may also be tasked to find the *measure of reliability*; how reliable our estimate is. One such measure may be $\left|\hat{\theta} - \theta\right| < \delta$ where we target to make δ as small as possible.

Example 1.1. Suppose 1000 cola consumers participate in a blind taste test among 2 brands, A and B, and are asked their preference. To know which kind is preferred universally, let us begin by asking the following:

- Describing the population. In this case, it is all the cola consumers.
- Describing the sample. In this case, it is our chosen 1000 cola consumers.
- The variable of interest. Whether people prefer brand A or brand B.
- Our inference. The preference in the sample is extended to all the cola consumers.

1.2.3 Type of Data

Qualitative data, or categorical or nominal or ordinal data, is data with no numerical value representation of it. In reference to the previous example, preference between A and B is a qualitative piece of data. Other such examples include choice of elective courses of a student, the gender of a person, a preference of a cricket team, etc..

Quantitative data on the other hand has a numerical value which interests us in statistics. Examples include the age of a person, the semestral marks of a student, the salary of a worker, the cost of books, etc.. Quantitative data is also divided into two parts, discrete and continuous.

1.3 Collecting Data

January 7th.

To collect data to perform statistics on, one may choose of the following ways to do so; a most basic source for sampling is a *published source*. In a *designed experiment*, we select experimental units and administer some treatment on each one. In the medical field, these medical experiments are called clinical trials. An *observational study* may also be conducted.

1.3.1 Sampling

Two kinds of simple random sampling exist; one is *simple random sampling without replacement* (SRSWOR), and the other is *simple random sampling with replacement* (SRSWR). Here, the experimental units are chosen sequentially at random with or without replacement. In *cluster sampling*, the population is divided into smaller groups known as clusters. Experimental units are then randomly selected among these clusters to form a sample. In *convenience sampling*, which is a non-probability sampling, the sample is drawn from that part of the population which is close to hand.

1.3.2 Sources of Error due to Flawed Sampling

A major source of error is *selection bias*; some parts of the population are deliberately left out when choosing the sample. It is an error of bias and not of randomness. For example, online surveys leave out people without access to internet. While this may be at fault of the person sampling the population, a *non-response bias* occurs when the population does not respond. A *reponse bias* also exists where the population does not reflect the true value. An error is noting down values or measuring samples may also occure, known as a *measurement error*.

Chapter 2

DESCRIPTIVE STATISTICS

The most one can possible do with qualitative data is to pictorially represent it via tables and charts.

Appendices

Chapter A

Appendix

Extra content goes here.

Appendix

Index