CHAPTER 1

Introduction

OBJECTIVE

The purpose of this introductory chapter is to cover basic notions that I thought you will need in subsequent chapters, including a broad and limited overview of the field of statistics. I review some software products that I recommend for processing your data, and end the chapter with short survey of the concept of probability

CONTENTS

1.1. Statistics and Abstraction

Students in my statistics classes often ask me the question "Why do have to learn this?" or "What is this going to be used for?" or "What are the concrete applications of this technique?" Rather than asking questions, other students will make bold statements such as "We need concrete solutions, and not statistical theory," or that "I will probably never use this stuff in practice." In my opinion all these questions and statements actually conceal a more fundamental issue, which is resistance to abstraction. Many of my students refuse to study statistics by developing basic technical skills first before applying them to solve real-world problems. They seem to prefer statistics to emerge in the middle of a discussion about a specific concrete matter. This sounds to me as if instead of teaching a child to count from 1 to 10, you decide to teach that child how to count cars, then dogs, then trees, and so on. That child will certainly learn to count all those things, but without establishing any relationship among them, without capturing the essential notion of number, abstracted from any concrete identifiable object.

I do believe that a minimum level of abstraction is necessary if you want to learn one technique, and use it to solve many different practical problems. The technique I have used for dealing with resistance to abstraction, and which I apply in this book is to use motivational concrete framework. I always start by describing a practical business or social problem to motivate the statistical investigation. My next step is to describe what the statistical solution looks like, and justify why it can be seen as solution, and finally start by developing the statistical apparatus needed to go from the problem to the solution. I must admit that, it could be quite difficult to remain motivated in studying a statistical technique over a certain period of time without knowing where it is going to lead us.

A discussion over the nature of a statistical solution to a specific problem, can indeed be quite passionate if well conducted. And such a discussion can be done before the technical concepts are introduced. Its purpose is not to replace the statistics course, and it cannot. Instead, its purpose is to have a conversation about our destination before embarking on the journey. You may expect as much as I do, that this conversation will make the frustrations and all sort of inconveniences associated with the journey more acceptable.

Practitioners and students tend to insist on the concreteness of what they learn, not simply as a rejection of abstraction, but also because they want to be busy producing numbers. But insisting too much of the production of numbers will probably overemphasize the importance of these numbers why neglecting the more fundamental question of their intrinsic value in the context of research. Practitioners do not only need to know how to do a test of hypothesis, or how to construct a confidence interval. Having a good grasp of the very nature of the progress the statistical solution will help us achieve, makes more sense to me. You need to know whether the statistical solution is even worth the effort, and what importance should you put on it in your investigative effort.

1.2. The Field of Statistics

The field of statistics has become so vast that there is no one single point of entry that can lead to all of its branches. This situation has made the study of statistics confusing. Consequently, before deciding what statistics class to take or what statistics book to buy or what statistical consultant to hire, a broad and high-level overview of the field of statistics is necessary.

The basic statistical activity consists of organizing and summarizing a data series. Sales data in a department store could be

presented in the form of monthly, quarterly, or annual total sales. The statistical measure in this context will be total or the sum. The same data is often displayed graphically with pie and bar charts. Some creative approaches for displaying statistical data are powerful exploration techniques, which have allowed analysts and statisticians to extract useful information from databases. This activity that consists of organizing and summarizing data is referred to as *Exploratory Statistics*[1] . Its role is limited to providing practitioners with metrics and graphical methods to show progress, to evaluate risk, to compare magnitudes. However, descriptive statistics does not provide the tools to answer important and broad research questions, when the researcher's interest goes beyond what observed data can tell. For example "How is sales revenue affected by the level of advertising expenses in general ?" or "How likely a male aged 15 to 25 is to be a smoker ?" These are two general research questions that do not refer to a specific data set. Therefore the data set can only be used as a stepping stone for exploring a whole universe of possibilities, or for inferring from observed data. Here you are entering the domain of "*Inferential Statistics*."

Inferential Statistics makes data speak through various modeling techniques. These techniques are not always very precise, but remain useful as the only means for studying hidden relationships and invisible parameters. But inferential statistics itself has many sub-branches with different aims. I will confined myself to mentioning only some of these sub-branches that ordinary scientists are likely to encounter in their professional life. The 2 branches of inferential statistics I like to mention are *Classical (or Mainstream) Inferential Statistics* and *Finite Population Sampling and Inference*. You will later see that each of these branches of infe-

[1]Some texts use the term *Descriptive Statistics*. However, some of the techniques used go beyond a mere description of data, and perform further exploration to extract information

rential statistics is further subdivided into the parametric[2] and nonparametric[3] branches with different aims. But for now, I like to show the difference between *Classical (or Mainstream) Inferential Statistics* and *Finite Population Sampling and Inference*.

Finite Population Sampling and Inference

An automobile insurance company may want to know about the likelihood for a driver to have one accident or more in a given year. If the interest is limited to a specific group of drivers who can be located (e.g. all drivers who live in the state of Maryland in the US), then the likelihood for a driver to have an accident represents the actual *proportion* (or relative number) of Maryland drivers who got into a car accident in any given reference year. This proportion is a concrete measure with a well-understood operational definition[4]. The number of Maryland drivers is known, and all of them constitute a finite population. Estimating quantities in this context appeals to the techniques of Finite Population Sampling and Inference. The quality of our estimations will also be evaluated with respect to the specific concrete population that is being targeted in our study.

Classical Inferential Statistics

If you do not want to confine yourself to a particular geographic location, you could look at the driver as an abstract person who could live anywhere (an abstract person does not have a real residence after all). In this context, the likelihood for a driver to have an accident is not associated with any well-defined operational

[2]Parametric methods are based on hypothesized models

[3]Nonparametric methods are based on the processes that generated the data being analyzed

[4]The operational definition associated with a concept is seen here as the step-by-step procedure for quantifying it

definition. It does not represent a concrete measure. Instead, it is a theoretical construct that is often referred to as the *Probability* that a driver will have an accident. Calculating this probability appeals to the techniques of Classical Inferential Statistics.

Of these two branches of inferential statistics, classical inferential statistics is the oldest one. Its development started several centuries ago to support scientific research. Finite population sampling and inference was invented primarily to meet governments' needs. Since classical inferential statistics does not refer to any specific population of interest, it is with no surprise that government officials have shown little interest in it. Given the limited resources they have at their disposal and the fact that there is generally a specific group of people they have to look out for, such a broad framework is inappropriate. Until today, Finite population sampling and inference has been used primarily in government-sponsored survey projects, and the importance of government databases to researchers is the primary reason for including this inferential framework in this book.

1.3. Statistics and the Notion of Variable

Statistics in general can be seen as the study and management of variability. Houses and cars for example have different prices. Even the price associated with a particular car model may change monthly, or quarterly depending on various factors such as government tax regulations or gas price. This variability in car or house prices explains the need to compile statistics to see what happened. Without variability there is no need for statistics. *Exploratory Statistics* is more concerned with the description of that variability, while *Inferential Statistics* will be concerned with both the description and the management of variability.

The field of exploratory statistics can be divided into the following three components :

- The descriptive data analysis
- The summary measures of quantitative variables,
- The exploratory data analysis

In statistics, characteristics of interest such as *age*, *educational attainment*, *height*, *gender* are generally coded or recoded with numeric values. The recoded characteristics are called *variables*. Let me consider educational attainment for example. This characteristic may be defined as {Some High School with no degree, High school graduate, Some college with no degree, College degree or more}. It could be recoded as {1, 2, 3, 4}, where 1 represents the lowest level (high school with no degree), and 4 the highest (college degree or more). This characteristic, which is recoded with numeric values with only an indirect connection to education, has now become an abstract variable that I could refer to as X (capital X). Such abstraction makes it easier to identify an appropriate existing statistical technique that could be applied to the study of educational attainment. Without such abstraction, it would be necessary to develop a technique for educational attainment, and another technique for a similar characteristic such as high school grade levels (1=freshman, 2=sophomore, 3=junior, 4=senior or more). Moreover, one may not even realize that both techniques are identical, a knowledge that eliminates the need to continuously reinvent the wheel. Throughout this book, most techniques will be described in terms of abstract variables such as X, Y, or Z. When applying them to solve a problem, it will be up to the researcher to recode the raw characteristic so as to match the variable so defined to a specific technique.

The use of variables in statistics is sometimes confusing to some students. It should not be. In fact using abstract variables instead of the more concrete characteristics is an exercise that we all do on a daily basis without even realizing it. When you purchase two items from a groceries store at the cost of $5.00,

and \$6.00, you know that the total to pay is \$11.00. You implicitly took the abstract numbers 5 and 6 (with no dollar sign) and remembered learning in elementary school how to add them to obtain 11, before putting back the dollar sign. This way you do not need to learn how to add dollars, pens, houses or people independently. You first learned to add abstract numbers before applying these skills in a concrete situation. But skills are developed more effectively in an abstract context. Statistical methods will often be described in terms of abstract variables taking numeric values. It will be up to the analyst to consider the 4 levels of the educational attainment characteristic independently of the concept of education to create an abstract variable X (actually how the 4 levels are defined has no statistical value). The variable X will be processed with the statistical technique, and the result will then be associated with the concept of educational attainment to formulate the research finding. You will later see in this book that the transition from the concrete to the abstract to back to the concrete is not always as smooth as we may want it to be.

Table 1.1 shows a small dataset of 9 professionals along with their gender and educational attainment, both expressed as characteristics then as variables. The 2 variables X, and Y contain all the information needed for analysis, while only the characteristics contain all the information needed to interpret the results. In addition to facilitating the development of statistical techniques, the variables present an auxiliary benefit. The numeric values they take make it easier to refer to specific groups of subjects. In the group of professionals of Table 1.1 for example, rather than saying "all professionals with a high school degree at the minimum," we could say ($X \geq 2$) *(read "X greater than or equal 2")*. In subsequent chapters, you will see that such groups are subject to numerous manipulations, which makes frequent references to long sentences particularly unwieldy.

Table 1.1 :
Educational Attainment and Gender of 9 Professionals

Individual	*Characteristics*		*Variables*	
	Education	Gender	X	Y
1	Some High School	Male	1	1
2	Some College	Female	3	0
3	Some College	Female	3	0
4	College Degree or More	Male	4	1
5	Some High School	Female	1	0
6	College Degree or More	Male	4	1
7	High School Graduate	Female	2	0
8	High School Graduate	Male	2	1
9	College Degree or More	Male	4	1

The coding of characteristics should be done so as to maximize the usefulness of the created variables. For example, educational attainment, which is an *Ordinal* characteristic[5] must be coded in such a way that its lowest level (i.e. some high school) receives the smallest of the 4 numbers, while its highest level (i.e. college degree or more) receives the highest number. Although we have the luxury of assigning any number to any level, coding "College degree or more" as 1 and "Some high school" as 4 clearly contradicts our intuition, while leading to a statistical analysis that will be harder to read. Likewise, gender can be coded as 3 for male and 4 for female or vice versa. In this case, which gender type receives the smaller number is irrelevant because Gender is a Nominal[6] characteristic. However, a more effective coding scheme for dichotomous (or binary) characteristics such as gender, is to assign the number 0 to one of the gender

[5] An *ordinal* characteristic is one whose levels can be ranked for low to high

[6] A *nominal* characteristic is one whose values are simply labels, identifiers, or attributes to identify subjects

type and 1 to the other. Its main advantage is that all the codes will sum to the count of individuals with the gender type that receives the code of 1.

1.4. Measurement Types

Throughout this book, I will describe various statistical techniques. Each of these procedures will only be valid for variables of a certain type. Therefore, knowing the variable type will be essential for identifying the proper procedure to use. The two main variable types I am concerned about are the *Categorical Variables* and the *Measurement Variables*.

1.4.1 Categorical Variables

A variable is called categorical when it takes a limited number of values. The simplest of all categorical variables are dichotomous variables (also called binary variables) such as gender that may be arbitrarily coded as 0 for male and 1 for female. If you conduct a telephone survey of 100 individuals, some of them will agree to participate while others will decline. You may defined a categorical variable X where $X_i = 1$ if individual i is a respondent, and $X_i = 0$ if individual i is a nonrespondent.

Dichotomous variables are commonly coded using 0 and 1. Although any pair of numbers would be suitable for coding the two groups of a dichotomous variable, the 0-1 coding scheme has the major advantage mentioned in section 1.3 that you cannot overlook. That is, all coded values sum to the number of cases coded as 1. Moreover, the arithmetic mean of the coded values equals the proportion of cases coded as one, which by itself is a quantity of interest. Because some statistical procedures are complex, any simplification in the coding scheme or in the notations will pay off.

Other examples of categorical variables include motorcycle manufacturers (i.e. X = *Motorcycle Manufacturer*) that can be coded as 1 for Honda, 2 for Yamaha, 3 for Kawasaki, 4 for Suzuki, 5 for Hartley-Davidson, and 5 for other. That is $X = 1, 2, 3, 4, 5$, and these codes represent labels or identifiers, and cannot be part of an arithmetic calculation in any meaningful way. You may distribute the number of motorcycles sold in a given year across manufacturers to determine market share. Note that no ranking of motorcycle manufacturers is possible based on the codes assigned to them, nor is any ranking possible among the values (0 and 1) of the gender variable. Random variables not offering any ranking possibility form a special class of categorical variables called *nominal scale* variables. These variables are used is some inferential procedures involving proportions such as the chi-square test to be discussed in subsequent chapters.

Other categorical variables allow for ranking and are called *ordinal scale* variables. This class of variables includes for example *Educational Attainment* defined in Table 1.1 and taking the values 1,2,3, and 4. There is a natural order in these numbers because "high school graduate" is normally superior to "some high school." Does it really matter whether you code this variable as 1,2,3, and 4 as oppose to 4, 7, 23,and 31 ? The answer is no, it does not matter, because ordinal scale variables are typically ranked first and the obtained ranks are further processed with a special statistical technique (e.g. non-parametric tests of hypothesis to be discussed later in the book). Consequently, the only thing that matters when coding educational attainment for example is to code "high school graduate" with a value higher than that used to code "Some high school" in order to preserve the natural order of things. Otherwise, the analysis will be impossible to interpret.

1.4.2 Measurement Variables

Length, *Width*, or *Height* are examples of *Measurement variables*. A measurement variable can be defined as a variable that takes values produced by a measuring instrument. The length takes values produced by a yardstick, while the weight takes values produced by a weighting scale. Note that a measuring instrument is not necessarily a hard physical equipment such as the weighting scale, it could also be one of these survey instruments used in psychological assessment, or a scoring model such as those used by credit card companies.

If the values that the measurement variable takes have numerical increments that correspond to equal differences in the physical entity over the entire range of measurement, then this variable is said to be an *Interval scale* variable. Measurements aimed at quantifying opinions would generally not produce interval scale variables. For example a variable measuring the level of satisfaction as "very dissatisfied" (coded as 1), "dissatisfied" (coded as 2), "neutral" (coded as 3), "satisfied" (coded as 4), and "very satisfied" (coded as 5) cannot be of interval type, since a numerical increment of 1 from very dissatisfied to dissatisfied will certainly not translate into the same chance in satisfaction level as the an increment of 1 from satisfied to very satisfied. Examples of interval data include temperature (in Celsius or Fahrenheit), year (e.g. 1980, 1981, 1982, etc...), and the different psychological test or credit scores. Note that the interpretation of a credit score for example may be debatable, but it is measured in such a way that a numerical increment will generally correspond to equal differences in abilities to reimburse debts. Several statistical procedures discussed in chapters 8 and 10 will be valid for interval scale data, but not for categorical data.

Interval scale variables such as the temperature or the credit

score have a fundamental difference with other measurement variables such as the height or the weight, which is the location and the meaning of the numerical origin. The height and the weight have both a numerical origin of 0, which corresponds to the total absence of any physical matter. For height and weight, 0 is a natural origin, which provides a clear-cut point where the measurement of the physical magnitudes begins. A 0 temperature on the other hand does not represent a total absence of temperature (very few people will dare wear t-shirts in a zero-degree temperature, especially if expressed in the Fahrenheit scale). Where does the measurement of temperature begin? Variables which take values with a natural origin are called *Ratio scale* variables.

The existence of a natural origin for ratio scale measurements has an important implication in practice, which is the possibility of making ratio comparisons. For example weighting 200 pounds means that you are weighting twice heavier than a 100-pound person. Such a ratio comparison would be impossible with interval scale variables. When the temperature is 40^0C, you will certainly not feel twice warmer than when the temperature is 20^0C. You should probably not put more emphasis on the difference between interval and ratio scale variables more than it is necessary. Most statistical techniques that are valid for one of these 2 types will also be valid for the other.

In many statistics textbooks you will often see the terms *discrete variables* or *continuous variables*. These 2 terms belong primarily to the language of mathematics. They represent notions that are relevant for a rigorous formulation of statistical theories based on the language of mathematics. The continuous variable is the mathematical idealization of a measurement variable, and is assumed to take all possible values in a continuum of possible values. The discrete variable is any variable that is not continuous. The set of possible values of a discrete variable if

often assumed to be countable. Practitioners should talk about measurement and categorical variables, while mathematical statisticians can talk about continuous and discrete variables.

A measurement variable such as height as used in practice is not really continuous, since it will generally be rounded either to the nearest integer or to a single digit after the decimal point. Still the rounded height is seen as a rough approximation of a hypothetical and more exact value that belongs to a continuum. Statistical results that are valid for continuous variables should nevertheless be applied to these pseudo-continuous numbers. There is nothing wrong with that, and it is unnecessary to waste time wondering whether your variable is continuous enough. The ride from the drawing board of theory to the messy world of hard can be bumpy. People with good judgement under those circumstances will have an edge.

1.5. The Software

If you are a practitioner, the study of statistical methods and techniques for you, is a first step towards achieving other scientific goals. Because soon or later you will need to take a dataset and actually implement those techniques you learned. This will be achieved only if you have a software product that you master reasonably well, and which you can use to process your data before interpreting the output. Unless you are dealing with a very small dataset, and perhaps wanting to implement a basic technique for producing simple descriptive statistics such averages, you are generally not going to succeed with a manual manipulation of your data. The era of statistical analysis by hand and handheld electronic calculators is over, and has in fact been over for decades. A computer, and a software product (ideally simple to learn) are mandatory if you are going to be serious about statistics.

What software? and for what purpose? It is difficult to recommend a particular software that is suitable to all needs. Such a software does not exist. I personally use multiple software packages for my consulting projects depending on the nature of the task. Among others, I used the R package, Excel, SAS, OpenOffice Calc, and only occasionally Stata, and SPSS. All examples in this book are given in the MS Excel for Windows (2007 & 2010), R, or OpenOffice Calc. My choice of these 3 products is justified by the fact that they are either free or widely available for personable use. R and OpenOffice Calc are completely free, while Microsoft Excel, although not free is widely available and often come loaded in most new computers running the Windows operating system. I will now briefly discuss the merits and limitations of each of these 3 products.

1.5.1 MS Excel (Shipped with the MS Office Suite)

Some statisticians have advised against the use of Excel for statistical analysis. I beg to disagree on this, and strongly advise using Excel when appropriate. There are indeed some odd datasets full of unusual numbers that Excel may not be able to process adequately. But it is highly unlikely that you will ever have to deal with such datasets in business or social research. However, if you are conducting high-level scientific research that requires robust algorithms to adequately handle unusual data, then Excel is certainly not for you. Excel in my opinion, is an excellent compromise between statistical capability and ease of use. I myself use Excel each time I need to perform a quick analysis on a small to medium size dataset.

Excel may not be a very efficient production tool for practitioners wanting to perform a large number of analyzes of the same type using perhaps many different datasets. If you are able to develop Excel macros then you can automate several tasks

and transform Excel into a good production tool. But investing in learning macro development is probably not justified unless you are already a heavy Excel user.

Which Version of Excel?

If you want to use Excel for statistical analysis, and your version is older than Excel 2010, I would advise that you upgrade to upgrade to Excel 2010, which is the latest version at the time of the writing of this book. In Excel 2010, the implementation of statistical functions has improved dramatically from what it was in the earlier version 2007. Excel developers have done a incredibly good job making the statistical functions work the way most practicing statisticians expect them to work. This is particularly true for the different functions used to evaluate the probability distribution functions discussed in chapter 2. Section E.8 of Appendix E describes the statistical functions used in this book.

Having said that, Excel 2007 still implements reasonably well most of the statistical techniques you may care about. If you already own Excel 2007 you may continue using it until you see the need to upgrade. You may well not need. However, a version older than 2007 may simply not be particularly useful.

Excel Add-Ins

An Excel Add-In is an external file that Excel can load when it starts up. The file contains a program that adds additional functionality to Excel, usually in the form of new functions or modules. Excel is shipped with a variety of Add-Ins ready for you to load and start using, and many third-party Add-Ins are available. The two Excel Add-Ins that I use in this book are the "Analysis ToolPak" and "Solver". Both come with MS Excel, but must be activated before they can be used for the first time. Appendix E provides all the instructions for setting up the Analysis

ToolPak. These instructions are to be used for setting up Solver as well (you would select Solver Add-In instead of Analysis ToolPak).

The Analysis ToolPak contains a long list of statistical modules each of which implements a specific statistical procedure. Solver on the other hand, is a power tool for solving optimization problems. The reason I introduced solver in this book, is to allow you to use many important statistical techniques that were previously accessible only to individuals with a good grasp of Calculus. With Solver, you no longer need Calculus to understand and implement advanced statistical techniques such as nonlinear regression, maximum likelihood estimation and more.

This book does not teach you how to use Excel. However, I will show you what can be done with Excel, with the Analysis ToolPak, and Excel Solver. But feel free to further explore these tools if you are going to use them to perform your own analysis.

1.5.2 OpenOffice Calc (it's free)

OpenOffice is a free fully-fledged Office suite that comprises among other products a text processor called Writer, and a spreadsheet called Cale that I will be concerned about. In this book, I use OpenOffice 3.2 that can be downloaded at,

`http://www.openoffice.org`

OpenOffice Calc 3.2 mimics the 2007 version of Excel very closely at least as far as statistical functions are concerned. The statistical functions in both Calc 3.2 and Excel 2007 generally have the same names. The only exception is that the function arguments in Excel are separated with commas, while they are separated with semi colons (;) in Calc.

Note that with OpenOffice Calc, you will not have access to powerful add-ins such as the Analysis ToolPak and Solver of Excel. Again, you may want to proceed with the reading of the book until you see what you need and what you do not need. I have used Calc on many occasions, and I was satisfied with what I was able to do with it.

1.5.3 The R Package (it's free)

The R package has become an immensely popular statistical package across the world. If you are going to do statistical analysis on a regular basis for many years, and you do not know which statistical software to learn, this is the one to get. No doubt. You will enjoy the support of an extended online support group where you will be able to ask questions. Moreover, the product is entirely free, and numerous books have been published to help practitioners and scientists learn this product.

The R package can be downloaded at,

`http://www.r-project.com`

Furthermore, the PDF file "Using R for Introductory Statistics" by John Verzani, which provides a short and friendly introduction to the R package, and a good overview of its capabilities can be downloaded at,

`http://cran.r-project.org/doc/contrib/Verzani-SimpleR.pdf`

R is an interactive computing environment that makes a large collection of statistical functions available to you. Using R is about finding the right function and learning how to use it. I provide several examples in this book that uses R. If you have never previously used R, you may want to proceed with the reading of this book to see what one can do with this package before

deciding whether you want to use it or not. R gives you the opportunity to develop your own functions for performing routine tasks as well as develop completely new packages for advanced users.

Appendix F describes some of the R functions most frequently used in this book. In order to avoid filling the body of the book with R material that non R users are not interested in, I will regularly refer to Appendix F where interested readers could find more details on the use of R.

1.6. Basic Exploratory Statistics

In the beginning of this chapter, I indicated that exploratory statistics was concerned about the description of the variability associated with random variables. Although variability at the population level will often be your primary interest, the sample is what you will often get. Therefore the description of variability will be done based on sample data. I now like to formally introduce the important notions of population and sample.

Populations and Samples

Statistics is often made difficult either because the study population has been ill-defined or because defining it has been downright omitted. Occasionally, the study population will be well defined then ignored afterwards, when in reality it should be driving the formulation and the implementation of the statistical procedure. The way you define your population must reflect your interests at the time of the investigation. This shows how critical it is for you to take the time to clearly envision what you want to focus on. This phase of the investigation is not statistical. It is about setting specific goals.

If you want to study income for example, you will have two

types of populations you can define. The first population will be the specific group of individuals whose income levels are of interest to you. If you investigate income as it relates to residents of the city of Los Angeles, it will be wise to consider defining your population as all residents of Los Angeles who are included in the labor force, and to see income as the characteristic of interest. This population is a *Population of Units*. If your goal is to study income as it is affected by educational attainment with no reference to a specific group of individuals, then you will want to define a *Numerical population*, where one member is an income figure such as $46,900.00. Your numerical population will then be a large collection possible numbers.

Populations of units are finite by nature. That is they represent a finite group of units that are the focus of the investigation. A correct and rigorous investigation of such populations may require the use of survey sampling methods that are discussed in chapter 12. Numerical populations on the other hand, have an unspecified number of numerical values that you will probably not be interested in, and which can be arbitrarily large. The mathematical idealization of these numerical populations consists of saying that these are *infinite populations*, which can take any values in a continuum. This idealization facilitates the formulation of statistical theories that only mathematical statisticians should care about, not practitioners.

To conduct your investigation you will generally focus on a small portion of your population called the Sample. Because the numbers in the numerical population are not tied to specific physical units, these populations are abstract in nature. The selection of samples from these populations is generally simple, since you do not have to worry about reflecting a particular structure that may be inherent to them. The only requirement being to select a random sample in order to remove any possible selection

bias from the process. Most techniques presented in standard statistics textbooks assume that you are dealing with numerical populations. If you are dealing with a concrete finite population of units, then sampling should be carried out carefully.

The remaining portion of this chapter is devoted to the study of basic exploratory techniques that you would apply to the sample data and initiate your statistical analysis. These techniques revolve around a graphical description of sample frequency distributions as a rough approximation of the underlying probability distributions, and the use of sample summary statistics as numerical approximations of the probability distribution parameters.

1.6.1 Frequency Distributions

Studying the frequency distribution of categorical and measurement variables requires different approaches. I will start with categorical variables and will discuss measurement variables afterwards.

Frequency Distribution of Categorical Variables

For categorical variables, the description of variability is essentially done with the *Frequency table* and the *Bar graph*. Consider for example Table 1.2 that shows the distribution of 500 subscribers of a newspaper by the type of community where the subscriber resides. The variable of interest is X = *Type of community*, which takes 3 possible values, {city, suburb, rural} conveniently coded as 1, 2, and 3 respectively. Table 1.2 shows the raw count of subscribers (also called frequency)as well as the relative frequency (expressed in percentages) representing the ratio of the frequency to the total number of subscribers.

The column of relative frequencies is the most important since

it describes the distribution of subscribers without being much affected by the number of subscribers. Relative frequencies are likely to remain stable in other studies based on a different number of respondents. This is stable information you can rely upon during a decision-making process.

Table 1.2 : Distribution of subscribers by community type

Type of Community	*Coded Value*	*Number of Subscribers*	*Relative Frequency (%)*
City	1	190	38
Suburb	2	170	34
Rural	3	140	28
Total		500	100

Figure 1.1 shows a bar graph that depicts the distribution of subscribers by type of the community of residence. Its main advantage is its ability to provide a quick and visual comparison of the different levels of the categorical variable X.

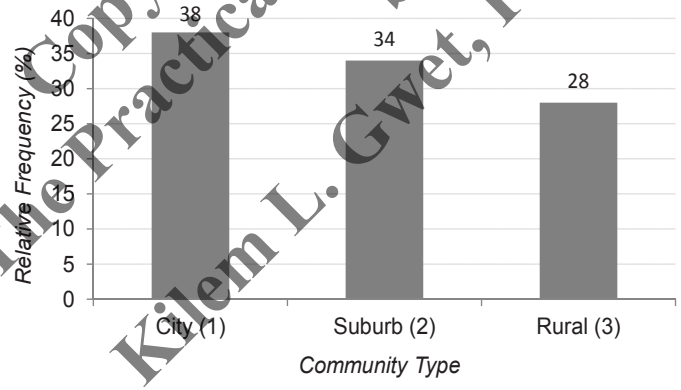

Figure 1.1. Bar graph for categorical variable "Type of Community"

The ordering of categories in the graph is arbitrary as are the codes 1,2, and 3 assigned to them. However, ordinal categorical variables will generally suggest a more natural ordering.

I created Figure 1.1 using Excel, although the same bar graph may be produced with R by typing the appropriate commands in the R console as shown in Figure 1.2. The formatting of this type of charts is generally more convenient with Excel. However, if you are already an R user, you may continue using it for the purpose of creating charts as well.

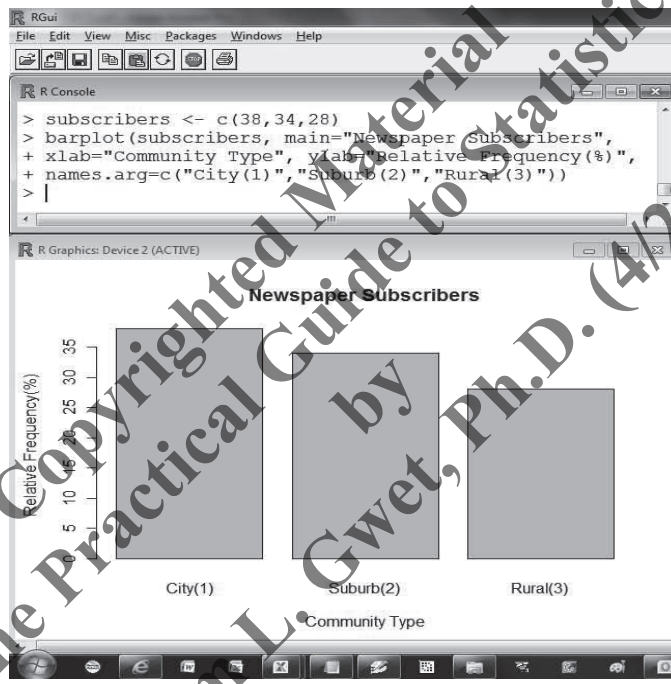

Figure 1.2. Bar graph for categorical variable "Type of Community" Using R

Frequency Distribution of Measurement Variables

Analyzing the frequency distribution of a measurement variable requires that you first categorize it before creating a frequency table and a *histogram*. The frequency table shows counts of observations within each of the categories called *bins*. The his-

togram is a graphical representation of the distribution of measurement data. It is similar to the bar graph, with the exception that the bars are adjacent and joined at the class boundaries to reflect the continuous nature of measurement data.

Table 1.3 shows the dollar amount that 40 families spent on food on a given day in an amusement park. Such a flat list of numbers does not allow us to tell any useful story regarding the spending habits of visiting families. The measurement variable of interest is $X =$*Daily dollar amount spent on food by a family*. In order to explore the distribution of this variable, I used the Histogram tool of Excel's Analysis ToolPak to create the frequency table 1.4 and the histogram in Figure 1.3. The details for creating Figure 1.3 can be found in section E.8 of Appendix E.

Table 1.3 : Daily food expenses of 40 families

$77	$18	$63	$84	$38	$54	$50	$59	$54	$56
$36	$26	$50	$34	$44	$41	$58	$58	$53	$51
$62	$43	$52	$53	$63	$62	$62	$65	$61	$52
$60	$60	$45	$66	$83	$71	$63	$58	$61	$71

The purpose of creating a histogram is to take a first look at the form of the distribution of a measurement variable. Are the amounts spent by families in the amusement park symmetrically distributed ? Are some of these amounts unduly high ? Unduly low ? Is the distribution skewed to right ? Skewed to left ? Is there a particular value around which most amounts are concentrated ?

There is no unique number of classes (or bins) that will allow you to best explore all aspects of the frequency distribution. To determine the number of bins, most statistics book recommend the use of the so-called "2 to the k rule" that recommends to use as number of bins, the smallest integer value k for which 2^{k-1} exceeds the number of observations. This rule is also known as

the Sturges' rule because it was proposed by Sturges (1926).

Excel appears to use the simpler rule that recommends the number of bins to equal the square root of the number of observations, rounded up to the nearest integer. It does not really matter much which rule you use, you may still need to modify the bins to obtain a better look at the distribution. For example, if your data series contains outliers (these are values that are located far away from the bulk of the data), you will need to have additional bins (some of which will have 0 observation) in order to have a more detailed look at the middle of the distribution. This is due to the fact that the bin width is calculated as $(Max\ Value - Min\ Value)/(k-1)$ where k is the number of bins. Therefore, a large extreme value will increase the bin width dramatically creating a high concentration of observations in the middle of the distribution, which will conceal the form of the distribution at that location.

Table 1.4 : Frequency Table of daily food expense data

Bin	*Bin Code*[a]	*Frequency*
\$0 to \$18	18	1
\$18 to \$29	29	1
\$29 to \$40	40	3
\$40 to \$51	51	7
\$51 to \$62	62	18
\$62 to \$73	73	7
\$73 to \$84	More	3
Total		40

[a]The bin code is all Excel produces. The associated bin definition as interval is in the first column.

If the construction of histograms is an issue that inspires you, you may want to look at other rules that have been suggested in the literature such as those of Freedman and Diaconis (1981),

and Scott (1979) .

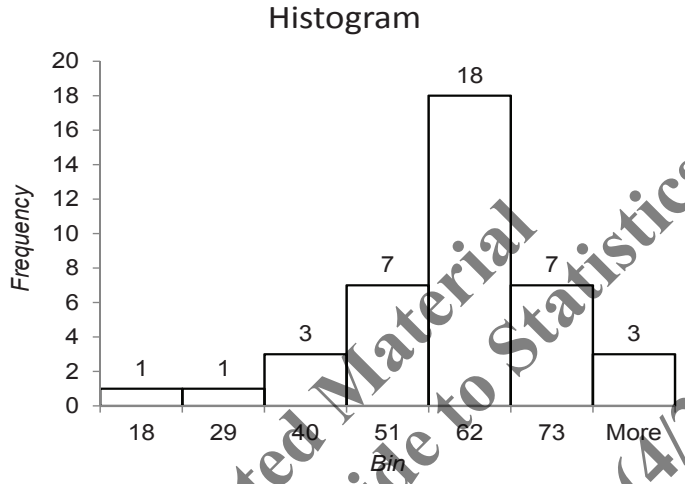

Figure 1.3. Histogram of Table 1.3 data

Figure 1.3 shows a rather regular smoothed distribution, which does not appear to have outliers. There is probably no need to to have both bins 18 and 29. Both may well be collapsed into a single bin 29 to obtain a smoother histogram. If a family had spent about \$6 or \$150 for example these would have been outliers resulting in a distortion of the histogram.

The R package too, offers an easy way to plot histograms with the function `hist()`. In its simplest form, this function will take a single argument representing the vector containing the data series. Numerous other parameters can be specified to have other features added to the graph (see the R documentation for more details).

1.6.2 Summary Statistics

In the last section, I discussed two of the most commonly-used graphical methods for exploring the probability distribution

of a random variable X, which are the bar graphs for categorical variables, and the histograms for measurement variables. There are other graphical techniques such as the boxplot, the pie chart, or the frequency polygon that I did not discuss due to their limited implications in statistical inference. In this section, I want to add to these visualization techniques, some numerical summary measures that will describe specific aspects of the variable distribution in a more precise way.

Two types of summary measures will be described in this section. These are the measures of location that will inform you about the location of the bulk of the data, and the measures of dispersion intended to inform you about the spread of the possible values of the characteristic of interest. A too wide range of values is an indication that any given value would be expected to be situated far away form the middle of the distribution. This situation will require a massive collection of data to obtain reliable information about the population being investigated.

Measures of Location

The measures of location that I like to mention are the mean, and the median. Each of these quantities has a population version and a sample version.

The Mean

When your population of interest is a finite and concrete population of N identifiable units, then the population mean of a variable X will be the average of all values $X_1, X_2, \cdots, X_N$ taken by the units. This population mean is denoted by $\overline{X}$ (read capital X bar), and has the following algebraic expression:

$$\overline{X} = \frac{1}{N}\sum_{i=1}^{N} X_i. \tag{1.1}$$

If X = *annual income* and your target population is made up of the 300,000 residents of a given city, then $\overline{X}$ as defined by equation 1.1 will be the population mean income, and will be seen as the finite population parameter

If the population of interest is an abstract numerical population as is often the case in most textbooks on classical statistics, the mean is actually the expected value of characteristic X and will be denoted by μ (read mu) or $E(X)$. That is $\mu = E(X)$. This expected value can be computed only if the theoretical probability distribution of X is specified. Several theoretical probability distributions will be discussed in chapter 2 that could be used for that purpose.

Whether the population parameter is a concrete (although unknown) quantity such as $\overline{X}$ or a theoretical construct such as μ, you will be able to approximate it numerically by selecting a sample of size n. The unknown population mean will then be approximated by the arithmetic mean (also called the sample mean) of the n sample values $x_1, x_2, \cdots, x_n$, which will be denoted by $\overline{x}$ (read small x bar). The equation of the sample mean is given by:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i. \tag{1.2}$$

The Median

Along with the mean, the median is another popular measure of central tendency that represents the middle of the distribution of a characteristic X. For a finite population of size N, the population median of X is the numeric value $\widetilde{X}$ that splits the N population values $X_1, X_2, \cdots, X_N$ in half. For a numerical population (viewed mathematically as *infinite*), the median denoted by $\widetilde{\mu}$ is a theoretical construct, which the random variable X will

exceed with a probability of 0.5. The population median ($\widetilde{\mu}$ or $\widetilde{X}$) will generally be approximated in practice by the sample median $\widetilde{x}$, a quantity that splits the n sample observations $x_1, x_2, \cdots, x_n$ in half.

If your sample contains 5 values {2,5,4,3,1} for example, its median is obtained by first sorting it as {1,2,3,4,5}, and by taking it middle value, which in this case is 3. That is, $\widetilde{x} = 3$. When the sample size is an odd number (e.g. 5), the median will always be one of the numbers in the sample. The situation is different when the sample size is an even number. If {2,5,3,1} represents your sample, then its sorted version will be {1,2,3,5}, which has 2 middle points 2 and 3. The median in this case will be the average of the 2 middle points. That is, $\widetilde{x} = (2+3)/2 = 2.5$.

If the mean is already available as a measure of central tendency, why would you need the median as an alternative measure? The reason is that the mean is highly sensitive to the presence of outliers in the sample, making the sample median the preferred choice when the distribution of data is skewed. The mean income calculated from a sample {\$55,000, \$50,000, \$75,000, \$63,000, \$67,000} is $\overline{x} = (\$55,000 + \$50,000 + \$75,000 + \$63,000 + \$67,000)/5 = \$62,000$. However, replacing \$67,000 with much larger number such as \$280,000 will yield a dramatically high mean income of \$104,600, which does not provide a good representation of the bulk of your data. By the way, the median in both samples remains at the same level of $\widetilde{x} = \$63,000$.

Proportions and Probabilities

The mean and the median are typically calculated for measurement variables. With categorical variables, you generally want to calculate proportions or probabilities. The *proportion* of students who receive an A grade in the last statistics represents the relative number of students to have received that grade, and will

be denoted by p.

Suppose that you want to evaluate the chance of a new graduate to get a job within 3 months after graduation. In this context, you have not specified any particular population of students in any school or any city, nor have you specified any particular timeframe for your inquiry. You are then dealing with an abstract population to which you cannot associate a concrete measure such as a proportion. Therefore, you will not talk about the proportion of new graduates to get a job within 3 months after graduation. Instead, you will talk about the *probability* for a new graduate to get a job. The probability represents a population parameter that you will denote by π (read pi).

Measures of Dispersion

The measures of dispersion that I like to present are the variance, the standard deviation, and the percentiles. If you have n sample data points $\{x_1, x_2, \cdots, x_n\}$, the *sample variance* denoted by s^2 represents the mean squared difference from the sample observations to the overall mean. This measure summarizes the spread of the sample observations around the sample mean, and is mathematically formulated as,

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2. \tag{1.3}$$

If you are seeing the above expression for the first time, you may be intrigued by its denominator, which is $n-1$ as opposed to the n often used in the calculation of averages. In practice using n or $n-1$ in the denominator will not affect the numerical values in a noticeable way. However, using $n-1$ gives the sample variance a mathematical property called unbiasedness, which mathematical statisticians seem to care about very much.

Example 1.1

If your sample contains the following numbers {\$55,000, \$50,000, \$75,000, \$63,000, \$67,000} then the sample mean will be $\overline{x} = \$62,000$, and the sample variance calculated as follows:

$$\begin{aligned} s^2 &= \big[(55000-62000)^2 + (50000-62000)^2 + (75000-62000)^2 \\ &\quad + (63000-62000)^2 + (67000-62000)^2\big]/(5-1), \\ &= 97,000,000. \end{aligned}$$

The difficulty with the use of sample variance stems from the fact that it is not expressed in the same original units as your sample data. In the above example 97,000,000 does not represents a dollar amount because the original values were squared to obtain the sample variance. To correct this problem, it is recommended to use the *sample standard deviation*, which represents the square root of the sample variance. The standard deviation associated with the sample variance of example 1.1 is $s = \sqrt{97,000,000} = \$9,848.86$, which compares very well with the initial sample values.

At the population level, the *Population Variance* measures how large you expect the squared difference of any given observation to its expected value to be. It is a population parameter that is denoted by S^2 when the population is finite and concrete, and is denoted by σ^2 (read sigma square) for abstract numerical populations. The population variance is generally unknown and should be estimated from the sample. The population standard deviation is the square root of the population variance and is denoted by S or σ for finite and abstract populations respectively.

Percentiles

Just as the median divides the sample data into two equal parts, the *Quartiles* divide the sample data into 4 equal parts,

the *Deciles* into 10 parts, and the *Percentiles* into 100 equal parts. There are 3 quartiles denoted by Q_1, Q_2, and Q_3, which respectively represent the $25th$, $50th$,and the $75th$ percentiles. Likewise, the $2nd$ decile is actually the $20th$ percentile. Consequently, knowing how to calculate percentiles is sufficient for calculating the quartiles and the deciles.

The general approach for calculating the p^{th} percentile is to first sort your n sample data points $x_1, x_2, \cdots, x_n$ in ascending order. Let L_p be the location (or the rank) of the p^{th} percentile on the sorted list. Therefore,

$$L_p = \textit{Integer part of} \left((n+1)\frac{p}{100} \right). \tag{1.4}$$

Example 1.2

Consider the following bi-weekly salaries of 15 entry-level accountants: \$2,038 ; \$1,758 ; \$1,721 ; \$1,637 ; \$2,097 ; \$2,047 ; \$2,205 ; \$1,787 ; \$2,287 ; \$1,940 ; \$2,311 ; \$2,054 ; \$2,406 ; \$1,471 ; \$1,460. To calculate the third quartile Q_3, I need to first sort the series in ascending order as follows: \$1,460 ; \$1,471 ; \$1,637 ; \$1,721 ; \$1,758 ; \$1,787 ; \$1,940 ; \$2,038 ; \$2,047 ; \$2,054 ; \$2,097 ; \$2,205 ; \$2,287 ; \$2,311 ; \$2,406. Therefore, the location of the third quartile (i.e. $75th$ percentile is,

$$L_{75} = \textit{Integer part of} \left((15+1) \times \frac{75}{100} \right),$$
$$= 12.$$

Consequently the $12th$ number on the sorted list, which \$2,205 is the third quartile ($Q_3 = \$2,205$).

Percentiles are useful statistics, since they allow you to identify the top 10% units in a given group for example. Note that p^{th} percentile of an abstract numerical population for which only

the probability distribution might be available, will generally be defined as the number x_p that will be exceeded with a probability of $1 - p$. I will further discuss these notions in subsequent chapters.

MS Excel can be used for producing basic descriptive statistics. One option would be to use the "Descriptive Statistics" module of the Analysis ToolPak. You will want to ensure that the "Summary Statistics" check box of the dialog form is checked. I have not discussed some of the summary statistics that Excel produces. One of them is the *Standard error*. It represents the standard deviation of the sample mean[7], and is generally obtained as the ratio of the simple standard deviation to the square root of the sample size. You will also want to know that Excel has functions for computing the percentiles, such as PERCENTILE.EXC, PERCENTILE.INC, PERCENTRANK.EXC, and PERCENTRANK.INC for Excel 2010, and PERCENTILE, and PERCENTRANK for Excel 2007. However, these functions do not compute the percentiles according to the algorithm presented earlier. I still advise to use them unless you are conducting a high-level scientific research where any tiny loss of precision is a big loss.

The R package too offers the summary() function that generates some limited basic summary statistics. Many R packages have been developed, and can be installed to produce a wider range of summary statistics. See the R documentation for more details.

[7]Note that as an average of n random variables, the sample mean itself is a random variable, which has a mean and a variance of its own.

1.7. The Calculus of Probability

I previously indicated that statistics in general can be seen as the study and management of variability. One notion that happened to play a pivotal role in the study of variability is the law of probability. It is the quantitative representation of variability, and the foundation of statistical inference.

Suppose you conduct telephone interviews for a survey research firm at the cost of \$50 per completed interview, and the response rate is known to be 70%. If potential respondents are selected randomly, then the dollar amount spent per interview is a two-value random variable that I will call X and that takes the 2 values \$5 (if respondent declines the interview) and \$50 otherwise. The law of probability associated with X is defined as follows:

$$P(X = \$50) = 0.70 \text{ and } P(X = \$5) = 0.30, \qquad (1.5)$$

where $P(X = \$50)$ and $P(X = \$5)$ represent the probability for the interview cost to be \$50 and \$5 respectively, and 0.30 is obtained as $1 - 0.70$. Once you know the law of probability of the variable of interest, you can use it to derive various statistical measures such as the mean, the median, the variance, and even conduct tests of hypotheses and more. Therefore, it should not come as a surprise if several chapters are often devoted to the notion of probability in many statistics textbooks. How can you compute the expected value of X (the interview expense) using the simple law of probability 1.5? Let $E(X)$ be the expected value of X. It is obtained as follows:

$$E(X) = \$50 \times 0.7 + \$5 \times 0.3 = \$36.5. \qquad (1.6)$$

That is the expected value of X is the sum of its possible values, weighted by their respective probabilities of occurrence. On the other hand, if you want to know how far you should expect either dollar amount ($5 or $50) to be from the expected value, you could compute the *standard deviation* as follows:

$$\begin{aligned} SD(X) &= \sqrt{(50-36.5)^2 \times 0.7 + (5-36.5)^2 \times 0.3}, \\ &= \sqrt{425.25} = 20.62. \end{aligned}$$

A standard deviation of 20.62 for an expected value of 36.5 indicates that any given value of the random variable X can be expected to deviate from the overall mean by as much 56 percent[8]. This substantial variation in the data was calculated using the possible values of X along with their respective probabilities of being observed.

But when I use the expression "probability of response" to a survey, what number am I exactly referring to? In this section, I am going to clarify the concept of probability. Because different experiments will lead to different estimations of the same probability, you need to have a good grasp of this notion to have an adequate interpretation of a statistical result. This is a good place to have a thorough discussion about the probability concept since in chapter 2, I will introduce a number of probability distributions that will be used throughout this book. I will also take this opportunity to mention a few fundamental concepts underlying the basic mathematical theory of probability.

1.7.1 What is a Probability ?

The ordinary law of nature describes the conditions under which an event will occur or will not. For example *every*

[8]Note that 56% is obtained as the ratio of the standard deviation to the expected value. This ratio is often called the *Coefficient of Variation*

stone placed in water will sink to the bottom if no external force is exercised on it. Once you describe the conditions under which the experiment will be conducted, you will know before the experiment even takes place, what the outcome will be. However, you will be led to statistical methods in those situations where the prediction of the course of events is impossible. When conducting a telephone interview survey, no matter how careful the planning, sometimes a solicited individual will agree to participate, sometimes that person will decline. It is the *Law of probability* that can describe the conditions under which an individual will agree to participate with a specified level of chance called the probability of response. How do I know what is exactly the magnitude of that probability of response? Intuitively one may think that if I attempt to contact similar individuals under the same conditions a large number of times, the relative number of times someone will agree to participate will tend to stabilize around a certain number, giving me a sense of the probability magnitude.

Many statistics textbooks indeed define probability as a limiting value obtained following an unlimited series of experiments executed under the exact same conditions. This definition, although widespread, can only give you a broad intuitive sense of what probability actually is, and does not have a solid mathematical foundation. A formal, rigorous, and concrete definition of the notion of probability is sometimes possible in specific situations. For example, consider an experiment that consists of selecting 5 individuals randomly out of a predetermined group of 20. The total number of samples of 5 out of 20 is 15,504. Therefore, the probability for a particular event occurring is defined as the relative number of samples of 5 individuals for which that event occurs.

For experiments involving a predetermined group of units or individuals in their design, it will generally be possible to have

a concrete definition of the notion of probability. Probabilities based on experiments for which the experimental conditions are not fully specified will generally not be defined in a way that is tied to reality. Suppose for example that you want to compute the probability the a light bulb manufactured under certain conditions will burn after 2,000 hours. The concrete meaning of such a number in this context is unclear, although you and I have some sense of what it represents. You should note however that, whether the real meaning of probability is clarified in a given context or not, the existence of the law of probability will be a reality that can and should be studied.

The notion of probability is intimately tied to that of experiments, primarily because it is with a repetition of experiments that one often quantifies the probability. Here are few definitions that will be needed later in this section:

Definition. 1.1.

The *Experiment* is seen as a process that produces a single of several possible observations.

In the conduct of an telephone interview survey, you may define your experiment in a number of ways. The simplest of the experiments consists of selecting randomly one potential participant from a specified pool of individuals, and determining whether that person agrees to participate or not. Only two observations are possible. Either the selected individual agrees to participate or she declines.

Definition. 1.2.

> An *Outcome* is a particular result of an experiment. And the set of all possible outcomes of an experiment is called the *Sample space*

If your experiment consists of selecting 5 potential survey participants for the purpose of determining the number X of those who are willing to participate, then one possible outcome is $X = 4$. All the possible outcomes are {0,1,2,3,4,5}.

Definition. 1.3.

> An (experimental) *Event* is a collection of one or several outcomes from an experiment.

It is common practice to label an event with capital letters such as A, or B for easy reference. Using the previous experiment of a survey of 5 randomly selected individuals, I can define an event as *E="More than 3 individuals agreed to participate in the survey"*. This event is the collection of the following 3 events: *A="Exactly 3 persons agreed to participate"*, *B="Exactly 4 persons agreed to participate"*, and *C="Exactly 4 persons agreed to participate."* I will then say that event E is a *Compound event*, which the union[9] of the 3 *Simple events* A, B, and C. A simple event is one that matches one of the possible outcomes, while a compound event is derived from two or more simple events.

When an event E is the union of three events A, B, or C, one will generally express this as,

$$\mathsf{E} = \mathsf{A} \cup \mathsf{B} \cup \mathsf{C}.$$

A second compound event could be defined as $\mathsf{F} = \mathsf{A} \cup \mathsf{B}$, which

[9]The union of 3 events occurs when any one of the 3 events occurs.

means that *"3 or 4 individuals agreed to participate"*. Additional operations are possible with events. For example, you could define event G that occurs only when both events E and F occur at the same time. Event G will be called "E *intersection* F" and denoted by,

$$\mathsf{G} = \mathsf{E} \cap \mathsf{F}.$$

Definition. 1.4.

Two events A and B are said to be *mutually exclusive* or *disjoint* when they have no outcome in common.

In practice, saying that two events are mutually exclusive amounts to saying that when one of the two occurs, the other cannot. For example the number of respondents to a survey cannot exceed 100 and still be below 55 at the same time.

Definition. 1.5.

A *Random Event* is an event which has the property that, following the execution of a long sequence of experiments repeated under the same fixed conditions, their frequencies tend to become stable and concentrated around a certain level called its *Probability*.

In the next section, I will present some elementary mathematical tools that will enable you to manipulate probabilities under various circumstances. A mathematical framework for studying laws of probabilities is necessary because researchers often start by establishing certain laws of probability by experiment or in a subjective manner, then proceed to derive new laws of probability by logical means under general assumptions. The mathematical framework of the next section will provide guidelines for an ef-

fective use of probabilities.

1.7.2 The Axioms of the Mathematical Theory of Probability

The whole mathematical theory of probability is based on the following 3 axioms[10]:

(1) The probability $P(A)$ (read P of A) of any given event A is assumed to lie between 0 and 1. That is,

$$0 \leq P(A) \leq 1, \text{ for any given event } A. \tag{1.7}$$

(2) The probability of the union of two mutually exclusive events is the sum of the probabilities of the individual events. That is, for any two mutually exclusive events A and B,

$$P(A \cup B) = P(A) + P(B). \tag{1.8}$$

(3) Any event that is certain will have the maximum probability of 1. That is, if event U is certain, then

$$P(U) = 1. \tag{1.9}$$

Axiom 1.7 is merely a mathematical formulation of the fact that the probability is often represented in practice in the form of a proportion. Therefore, it can only take values between 0 and 1. As for the second axiom 1.8, it represents the fact that you will observe the union of 2 disjoint events by observing one after the other, and the frequency of the union will be the sum of the frequencies of the individual components. The last axiom 1.9 reflects the fact that any event, which systematically occurs following the execution of an experiment should have the maximum probability of 1.

[10] An axiom is a statement considered to be true, and which is used in the foundation of a system of thoughts

The 3 axioms 1.7, 1.8, and 1.9 can be used to deduce several other probabilities once a few initial fundamental probabilities have been provided through experimentation or prior knowledge. I now present some of the best known and most useful properties of the law of probability that are deduced from the 3 foundational axioms. In order to help understand how probabilities must be evaluated here is what the great Russian mathematician A. N. Kolmogorov said (see Aleksandrov et al. (1999), Vol. Two, Part 3, Page 252):

> *The basic concepts of the theory of probability, namely random events and their probabilities, are completely analogous in their properties to plane figures and their areas. It is sufficient to understand by $A \cap B$ the intersection (i.e. common part) of two figures, by $A \cup B$ their union, by $\emptyset$ the conventional "empty" figure, and by $P(A)$ the area of figure A, whereupon the analogy is complete.*

I found this analogy very useful for conceptualizing the notion of probability when dealing with complex probabilistic issues. It will be further illustrated in chapter 2. Here are a few properties that are derived from the 3 foundational axioms:

- For any event A, $P(A) = 1 - P(A')$ where A' is the complement of A (i.e. any outcome that is not included in A).
- If events A and B are disjoint, then $P(A \cap B) = 0$.
- For any two events A and B, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Here is an example that illustrates how the probability can be calculated in practice.

Example 1.3

A random sample of 2,000 licensed drivers revealed the following numbers of speeding violations.

Table 1.5 : Distribution of 2,000 Drivers by Number of Speeding Violations

Number of Violations	Number of Drivers
0	1,910
1	46
2	18
3	12
4	9
5 or more	5
Total	2,000

a What is the experiment?

The experiment in this example consists of selecting a group of 2,000 drivers randomly and to obtain the number of speeding violations of each of them. You cannot be more specific than that because a particular sampling scheme has not been described. As for the sample space, depending on the timeframe used for observing the number of speeding violations, this one may vary from 0 to a large number.

b List one possible event.

One possible event is A=$\{$*A selected driver has 3 speeding violations or more*$\}$

c What is the probability that a particular driver had exactly two speeding violations ?

Let me define a generic event $A_k = \{$*A selected driver has exactly k speeding violations*$\}$ for $k = 0$ through 4, and $A_5 = \{$*A selected driver has 5 violations or more*$\}$. The problem is to compute the probability $P(A_2)$, which is the

ratio of the number of drivers with 2 violations to the total number of drivers. That is, $P(A_2) = 18/2000 = 0.009$..

A alternative, and perhaps more convenient way to approach this is to define the random variable $X = \{$*The number of speeding violations of a randomly selected driver*$\}$, and compute $P(X = 2) = 18/2000 = 0.009$.

- The method that I used to respond to question (c) is the *Empirical approach* for assigning probabilities. Empirical probability is calculated as the ratio of the number of times the event occurs to the total number of observations. In other experiments, you may have to use the *Classical approach* for assigning probabilities. The *Classical probability* is calculated as the ratio of the number of favorable outcomes to the total number of possible outcomes.