A Short Additional Note on Probability Theory

Jacco Thijssen

January 2008

1 Introduction

Probability theory is the branch of mathematics that deals with *random experiments*. These could be simple ones like flipping a coin or rolling a die. The main underlying idea is that there is some form of *randomness* in the outcomes that we (the experimenters) can not resolve. In the natural and life sciences experiments are readily thought of as scientists constantly perform them in laboratories. In the social sciences, however, there are no clear-cut experiments.[1] Some examples of common phenomena that social scientists are thinking of as experiments are: conducting a survey, the formation of prices on the stock exchange, the realisation of a country's GDP, the outcome of an election, etc. It is because of these analogies that we use probability theory in the same way in the social sciences as in the natural sciences.

Probability theory will form the corner stone of our treatment of inferential statistics, which is the science/art of using sample data to make inferences about the entire population. It is important to realise that probability theory is a subject in and of itself, which happens to be used in an important branch of statistics. It is neither exclusively used, nor ultimately developed for statistics. In this course we will only discuss the absolute basics needed for the study of inferential statistics.[2]

[1] This is changing, by the way. Especially in economics many studies are carried out these days where economists get a group of people together in a "laboratory" to perform an economic experiment. This usually involves letting the subjects make economic decision behind a computer.

[2] Modern probability is seen as a branch of "measure theory"; an abstract mathematical theory.

In my opinion, the textbook misses some essential elements in its treatment of probability theory.[3] In this note I try to give a relatively rigorous, but still intuitive construction of the basic ingredient of any probability analysis: the probability space. These notes should bee seen to augment Sections 4.1 and 4.2 of the book.

2 Probability Spaces

The mathematical description of a random experiment contains three ingredients:

1. A list of all the possible outcomes of the experiment. This is called the *sample space* and is usually denoted by S.

2. A set consisting of all "events" that we wish to assign a probability to. This set is called the *event space* and is denoted by $\mathcal{A}$.

3. A function which assigns a number between 0 and 1 to every event in $\mathcal{A}$. This function is called a *probability* and is denoted by P.

The collection $(S, \mathcal{A}, P)$ is called a *probability space* and is the starting point of *any* probability analysis.

Example 1 (Roll a die)**.** *Consider the experiment "roll a fair die once". The possible outcomes of this experiment are 1,2,3,4,5, or 6. So,* $S = \{1, 2, 3, 4, 5, 6\}$.[4] *What are possible "events" in this experiment? Well, there are many possibilities. For example, "number 1 comes up", or "the number that comes up is either 3 or 5", or "an even number comes up", etc. All these events can be written as sets:* $A_1 = \{1\}$, $A_2 = \{3, 5\}$, $A_3 = \{2, 4, 6\}$, *respectively. Note that all these events are* subsets of the sample space S. *That is, they contain some elements (but not necessarily all) of* S. *Finally, what probability should we assign to all these events? Before we write this down formally, think about this for a second. What probabilities would* you *assign to* A_1, A_2, *and* A_3*?*

[3] Most books water down the probability theory to such an extent that it becomes predigested pap, which confuses rather than enlightens. You are too smart for that.

[4] Mathematicians are, *im großen ganzen*, lazy people (that's why, for example, we use the summation notation). Therefore, we often write this as $S = \{1, \ldots, 6\}$.

Since the die is fair we could assign each outcome equal probability. To compute the probability of an event $A \in \mathcal{A}$ (remember $\mathcal{A}$ is the set containing all events we want to assign a probability to) we simply put

$$P(A) = \frac{|A|}{|\Omega|}, \tag{1}$$

where $|A|$ denotes the number of elements in the set A. So, in this example we would get

$$\begin{aligned} P(A_1) &= \frac{|\{1\}|}{|\{1,\ldots,6\}|} = \frac{1}{6}, \\ P(A_2) &= \frac{|\{3,5\}|}{|S|} = \frac{2}{6} = \frac{1}{3}, \quad \textit{and} \\ P(A_3) &= \frac{|\{2,4,6\}|}{6} = \frac{1}{2}. \end{aligned}$$

You probably came to the same conclusion already.

Let's summarise two aspects that are made clear by this example.

1. Events can be written as subsets of the sample space; they simply contain a certain number of possible outcomes.

2. If all outcomes are deemed equally likely we can assign the probability (1). This probability is usually referred to as the *classical probability*.

Example 2 (Coin flip)**.** *Consider the experiment "flip a fair coin twice". There are four possible outcomes of this experiment: Heads both times (HH), Tails both times (TT), Heads the first time and Tails the second (HT), and vice versa (TH). In other words $S = \{HH, HT, TH, TT\}$. Let's consider the following two events: "the first coin flip gives H" and "both flips give a different outcome". In set notation: $A_1 = \{HH, HT\}$, and $A_2 = \{HT, TH\}$, respectively. What probabilities to assign? Well, the coin is fair, so for each flip each outcome is equally likely. In other words,*

$$P(A_1) = \frac{|\{HH, HT\}|}{|S|} = \frac{2}{4}, \quad \textit{and} \quad P(A_2) = \frac{1}{2}.$$

After studying these two examples you might be left with some questions, like:

1. do I have to assume that each outcome is equally likely? and

2. how do I know that the die or coin is fair?

The answers to these two questions are as follows:

1. no, not at all! See Example 3 below, and

2. you don't. That's exactly why we study statistics! Take the coin example. In real life we never know whether a coin is actually fair. So, we flip the coin repeatedly and compute the frequency of Heads to *estimate* the probability of heads coming up. The study of statistics is all about procedures to make exactly these kind of *inferences*.

Example 3 (Example 1 cont'd)**.** *Suppose that you have a feeling that the die you are about to roll is loaded. It may have been given to you by some River Boat Gambler (or worse, a stock broker), and you suspect that it is much more likely that the number 1 comes up. How to assign a probability in this case? You could start by assigning a number p_k to all possible outcomes $k \in S$. A natural probability would then be to set*

$$P(A) = \sum_{k \in A} p_k. \tag{2}$$

For example, if $A = \{1, 2\}$ and $B = \{1, 3, 5\}$, then

$$P(A) = p_1 + p_2, \quad \text{and} \quad P(B) = p_1 + p_3 + p_5,$$

respectively.

Maybe you have rolled the die a few times and conclude that $p_1 = \frac{1}{2}$ and that all other outcomes are equally likely. Or maybe you simply believe *that this is the case. In the former case we speak about the* relative frequency *definition of probability, whereas in the latter case we speak of the* subjective *definition. Whatever the case, could you now, for example, set $p_2 = \cdots = p_6 = \frac{1}{6}$? Well, no. If you were to do that then you would run into a lot of problems later on.*

In fact, I haven't been entirely honest earlier on when I implied that the probability could be *any* function assigning numbers between 0 and 1 to events in $\mathcal{A}$. I should have added the following restrictions, which in probability theory have the status of axioms.[5]

[5]In mathematics an "axiom" is a statement that represents a convention, a statement that we do not question. The trick is to build useful mathematics with as few axioms as possible. For the set of real numbers, $\mathbb{R}$, for example, it is an axiom that $x + y = y + x$. The fact that $0 \cdot x = 0$, all $x \in \mathbb{R}$, however, is not an axiom, even though you probably accept it without questioning. In fact, it can be proved from surprisingly few axioms.

Axiom 1. *A probability is such that* $P(S) = 1$.

Axiom 2. *A probability is such that if* $A_1, A_2, \ldots$ *are mutually exclusive events in* $\mathcal{A}$ *(that is,* $A_i \cap A_j = \emptyset$*, all* i, j*), then*

$$P(A_1 \cup A_2 \cup \cdots) = \sum_i P(A_i).$$

The second axiom looks difficult, but makes intuitive sense. If I have two events that can not occur simultaneously, then the probability of the one *or*[6] the other taking place is just the sum of their respective probabilities.

Example 4 (Example 3 cont'd)**.** *If we take* $p_1 = \frac{1}{2}$ *and* $p_2 = \cdots = p_6 = \frac{1}{6}$*, then* P *– as defined in (2) – does not satisfy the axioms. Note that* $\{1\}$*,* $\{2\}$*,* $\{3\}$*,* $\{4\}$*,* $\{5\}$*, and* $\{6\}$ *are mutually exclusive events (you can not have two or more numbers coming up at the same time). So, from the second axiom it follows that*

$$P(\{1\} \cup \cdots \cup \{6\}) = \sum_{i=1}^{6} p_i = \frac{1}{2} + \frac{1}{6} + \cdots + \frac{1}{6} = \frac{8}{6} = \frac{4}{3}.$$

In addition, we have that

$$\{1\} \cup \cdots \cup \{6\} = \{1, 2, 3, 4, 5, 6\} = S.$$

In other words, the second axiom would imply that

$$P(S) = P(\{1\} \cup \cdots \cup \{6\}) = \frac{4}{3} > 1,$$

which contradicts the first axiom. So, this choice for P *is* not *a probability. It is easy to see that the* only *probability that satisfies the axioms, that assigns probability 1/2 to the outcome 1 and equal probabilities to the other outcomes, and that is consistent with the frequency definition (2) has*

$$p_1 = \tfrac{1}{2}, \quad \textit{and} \quad p_2 = \cdots = p_6 = \frac{1}{10}.$$

(Check that this indeed satisfies the axioms!)

This example shows a feature that I find so important that I would want you to put in a frame and hang above your bed:

[6] We always use the mathematical "or": either one or the other or both.

The (mathematical) theory of probability does not tell you *what* probability to choose. It merely gives you a *consistent* framework to make probability calculations, *once YOU have decided on the appropriate probability.*

This is a very important statement. It implies that – from a probability theory point of view – *the* probability does not exist. It is something we choose using procedures that lie outside the realm of mathematics: experience, intuition, etc. In statistics we often (implicitly) choose a particular family of probabilities, among we then choose one based on sample information.[7]

3 Continuous Sample Spaces

So far, we have only looked at examples with a discrete state space. Let's take a look at a simple example with a continuous state space and the problems it creates. The example is meant as an illustration of the many subtleties that appear in probability theory. It will not show up in the exam, but some feeling for the problems with continuous state spaces will make life easier in a few weeks time. Continuous sample spaces, namely, are used very often in statistics (any time we deal with a continuous variable, for example).

Example 5 (Draw a random number)**.** *Consider the experiment "draw a random number from the interval* $[0, 1]$*". Modelling this simple example leads to complicated problems. The sample space is easy enough:* $S = [0, 1]$*. However, the event space* $\mathcal{A}$ *is very complicated. Essentially, you cannot assign a probability to* every *subset of the interval* $[0, 1]$*, as that would lead to inconsistencies. The reasons why are way beyond the scope of this course and would take us firmly into measure theory territory. The event space is usually taken to be the "Borel* σ*-field on* $[0, 1]$*". The only thing to know about this animal is that it contains all subsets of* $[0, 1]$ *that we will ever be interested in. Defining the probability on* $\mathcal{A}$ *is also not an easy thing. Let's first look at subsets of the form* $[a, b] \subseteq [0, 1]$*. It would make sense to define the probability* P *such that*

$$P([a, b]) = b - a, \tag{3}$$

so, for example, $P([0.25, 0.75]) = 0.5$ *and* $P([0, 1/3]) = 1/3$*, which seems a logical choice. This definition is then extended to all sets in* $\mathcal{A}$*. A consequence*

[7] This sounds rather cryptic, but should become clearer later on.

of this choice of probability is that

$$P(\{a\}) = 0, \quad \textit{all } a \in [0,1].$$

The reason for this is – again – too complicated, but it should be understood that it is NOT linked to (1) whatsoever. Many people reason as follows:

$$P(\{a\}) = \frac{|\{a\}|}{|S|} = \frac{1}{|[0,1]|} \text{“} = \text{”} \frac{1}{\infty} \text{“} = \text{”} 0.$$

This reasoning is false and has no mathematical meaning. To see that it does not work, consider the set $\mathbb{Q} \cap [0,1]$.[8] *Applying (1) then leads to*

$$P(\mathbb{Q} \cap [0,1]) = \frac{|\mathbb{Q} \cap [0,1]|}{|[0,1]|} \text{“} = \text{”} \frac{\infty}{\infty} = ?$$

Instead it can be shown that

$$P(\mathbb{Q} \cap [0,1]) = 0.$$

This leads to the following interesting paradox. The probability of any *number being drawn is 0. In fact the probability of drawing one of infinitely many rational numbers is 0 as well. Still, however, the probability of* some *number being drawn is 1, since* $P([0,1]) = 1$ *by the first axiom. In other words,* with probability 1 an event with probability 0 occurs*! This implies that "probability 0" does not mean "impossible".*

Is your head spinning yet?

[8] That is, the set of all rational numbers (fractions) in the interval $[0,1]$. This set contains infinitely many elements, but not, for example, $\frac{1}{2}\sqrt{2}$.