PROBABILITY IS SYMMETRY

On Foundations of the Science of Probability

Krzysztof Burdzy

CONTENTS

PREFACE

This book is about one of the greatest intellectual failures of the twentieth century—several unsuccessful attempts to build a scientific theory of probability. Probability and statistics are based on very well developed mathematical theories. Amazingly, these solid mathematical foundations are not linked to applications via a scientific theory but via two competing and mutually contradictory philosophies, pretending to be science. One of these philosophical theories ("frequency") is an awkward attempt to provide scientific foundations for probability, the other theory ("subjective") is truly bizarre. A little scrutiny shows that in practice, the two ideologies are almost entirely ignored, even by their own supporters.

I will present my own vision of probability in this book, hoping that it is close to the truth in the absolute (philosophical and scientific) sense. This goal is very ambitious and elusive so I will be happy if I achieve a more modest but more practical goal—to build a theory that represents faithfully the foundations of the sciences of probability and statistics in their current shape. A well known definition of physics asserts that "Physics is what physicists do." I ask the reader to evaluate my theory by checking how it matches the claim that "Probability is what probabilists and statisticians do." As I have already mentioned, what statisticians really do is practically unrelated to the most popular philosophies of probability, except in a handful of trivial cases.

The title of the book was inspired by the following remark made by Wilfrid Kendall: "I have come to the conclusion over the years that symmetry is a matter of belief." I have come to the conclusion over the weeks that this sentiment encapsulates the difference between the subjective theory of probability and my theory presented in this book.

Is another book on the philosophy of probability needed? I think that it is, for several reasons. First, some theories of probability are false and this should be said repeatedly, even if it was said many times before. The second reason is practical. I am sure that the best statisticians and scientists ignore the most absurd parts of the official philosophies but I am not certain whether this can be said about all practitioners of probability and statistics, so it is important to expose the follies of the popular probabilistic ideologies. Finally, the discrepancy between some tenets of the official philosophies and the contents of textbooks is so large that it can be described only as hypocrisy. This should be remedied, needless to say.

So far, my research has been almost totally focused on pure mathematics. Since this is my first adventure in the area of philosophy, a word about my motivation is appropriate.

I always felt, and I still do, that subjectivity and science do not mix. At the same time, I know that many rational people, most notably "Bayesian" statisticians, have an opposite view. The critique of the subjective theory that I found in the literature is not sufficiently convincing to me and, clearly, it is ignored by the Bayesians. I could have tried to find all the answers I was looking for in the existing literature, but in the end it proved easier and more satisfying to build a theory of probability by myself. In the course of the work on this project, I realized that it is impossible to treat just a few aspects of the philosophy of probability; one really needs to discuss all relevant questions and this is why my first article-size drafts grew to the size of the present book. The book answers all questions that I was intrigued by but I had to stop and leave some philosophical puzzles unsolved, partly because I was not interested in solving all the relevant problems, but mainly because I was not capable of doing so. I am not happy about some loose ends but I do not think that achieving perfection is a realistic goal. I want to share my thoughts on probability with other people, not because my theory is perfect, but because it satisfies my craving for the common sense, and I hope that it will have a similar appeal to the reader.

It is hard to be original in philosophy, so much of what I am going to say is already known and has been expressed in some way. It is easier to explain what is new in the book rather than to list all the recycled ideas. I believe that my "scientific laws of probability" (L1)-(L5), presented in Chapter 5, are new, although their novelty lies mainly in their form and in nuances of their interpretation. I also think that my critique of the subjective theory contains novel ideas, including a proof that the subjective theory is self-contradictory.

Despite the fact that this is a philosophical book, it is not my intention to compete with professional philosophers in the area of probability. The book is written from the point of view of a scientist and it is meant to appeal to scientists rather than philosophers. Readers interested in the professional philosophical analysis of probability (especially in a more dispassionate form than mine) may want to start with one of two very accessible monographs by Gillies [2] and Weatherford [8]. An article by Primas [4] is a very interesting and useful review of many philosophical problems of probability although it is technically challenging at some places. Reference lists in [2], [4] and [8] are a great starting point for anyone wishing to explore the subject in greater depth.

Seattle, 2003

1. INTRODUCTION

1.1. Reality and philosophy.

Two and two makes four. Imagine a mathematical theory which says that it makes no sense to talk about the result of addition of two and two. Imagine another mathematical theory that says that the result of addition of two and two is whatever you think it is. Would you consider any of these theories a reasonable foundation of science? Would you think that they are relevant to ordinary life?

If you toss a coin, the probability of heads is 1/2. According to the frequency theory of probability, it makes no sense to talk about the probability of heads on a single toss of a coin. According to the subjective theory of probability, the probability of heads is whatever you think it is. Would you consider any of these theories a reasonable foundation of science? Would you think that they are relevant to ordinary life?

The frequency theory of probability is usually considered to be the basis of the "classical" statistics and the subjective probability theory was adopted as the basis of the "Bayesian" statistics (the terms will be explained in Chapters 2 and 3). According to the frequency theory of probability, the concept of probability is limited to long runs of identical experiments or observations, and the probability of an event is the long run relative frequency of the event. The subjective theory claims that there is no objective probability and so probabilities are subjective views; they have to be "consistent," that is, they have to satisfy the usual mathematical probability formulas, to be useful.

I will now present two examples showing that statisticians behave as if they did not believe in their own philosophical theories. Then I will outline the main claims of this book and discuss the organization of the material in the remaining chapters.

The "classical" statistics, the one that is based on the frequency theory of probability, developed a notion of confidence intervals. If a scientist wants to find the value of a physical constant, she can perform a large number of measurements and then find a 90%-confidence interval for the quantity. Here 90% can be replaced with some other probability close to 100%. The meaning of the "90%-confidence interval" is that the interval covers the true but unknown value of the quantity with probability 90%. To interpret this statement using the frequency theory, one would have to have a long *sequence of sequences* of measurements of the same quantity (all sequences having the same length), and then one would have to generate a series of confidence intervals, each one on the basis of a different data set. Such situations might occur occasionally in practice but they are far from common. Most of the

time, a single sequence of measurements is all that is considered—classical statisticians do not shy away from using confidence intervals in such situations.

If the reader has some spare time, I suggest performing the following experiment that will determine whether Bayesian statisticians assign probabilities in a subjective way. One should deform a coin slightly but visibly with pliers so that it is not flat and, therefore, it cannot be presumed to be symmetric (this is not essential but will make the experiment more interesting). The coin should be taken to a large conference on Bayesian statistics and one thousand Bayesian statisticians should be approached, one at a time, with the following questions.

(i) What would you like to have for dinner today?

(ii) If you toss this deformed coin 100 times and observe the results, how will you find the probability of heads on the 101-st toss?

Answers to question (i) will undoubtedly uncover a variety of preferences—some people will opt for a steak, some will request seafood and some will turn out to be vegetarian. The respondents will vary considerably in their choices of the style of cooking: Italian, Chinese, Thai, etc. The answers will prove that culinary preferences are subjective.

As for question (ii), one hundred percent of Bayesian statisticians will say that the tosses are "exchangeable" and ninety nine percent of them will choose the "uniform prior." If you do not understand these terms, do not worry, I will explain them later. The only important point here is that the Bayesian statisticians will show a striking unanimity in their opinions about probabilities in this situation, and in a great variety of other situations. If the term "subjective" has any meaning at all, Bayesian statisticians do not assign probabilities in a subjective way.

1.2. Summary of the book's main claims.

In Chapter 5, I will present five simple scientific laws of probability (I will call them (L1)-(L5)) and I will argue that they are a *de facto* standard of the applications of probability in all sciences. One of my main claims is that

> *Statisticians, classical and Bayesian, and all other scientists behave as if they believed that (L1)-(L5) were objectively true.*

The main criteria for my choice of the "scientific laws of probability" are simplicity and agreement with the current scientific practices. Any such laws have to be verifiable in some reasonable way that can be implemented in real life. They should be also directly applicable in simple situations. Personally (subjectively?) I believe that (L1)-(L5) are

objectively true but I will propose a much weaker, "minimalist," interpretation of the laws in Section 5.2.

I will argue that classical statisticians do not limit their applications of probability to long runs of experiments or observations and Bayesian statisticians do not assign probabilities in a subjective way. However, I will not limit my arguments to pointing out a discrepancy between the officially supported philosophies and the scientific practice. I see profound philosophical problems with the frequency and subjective theories. When we attempt to interpret the frequency theory, we face the following alternative. Either (i) the frequency theory is not meant to make any predictions, or (ii) it is supposed to be used for making predictions. In the first case, the frequency theory descends into pure philosophy, with no practical significance. In the second case, the frequency theory turns out to be highly incomplete—it does not provide an account for a large number of common applications of probabilistic techniques.

The above alternative is also relevant in the context of the subjective theory but the most challenging alternative that the subjective theory has to face is somewhat different. An interpretation of the subjective theory can be based either on the assumption that (i) symmetries and physical independence need not be taken into account when "subjective" probability assignments are made, or (ii) the subjectivists should base their opinions on physical independence and symmetry (among other things). In the first case, the subjective theory can be proved to be false by a reasonable scientific test. In the second case, the philosophical structure of the subjective theory collapses. Needless to say, all these assertions will be expanded upon in the following chapters.

I would like to stress that I do not see anything absurd about using the frequency and subjective interpretations of probability as mental devices that help people to do abstract research and to apply probability in real life. Classical statisticians use probability outside the context of long runs of experiments or observations, but they may imagine long runs of experiments or observations, as it may help their mental processes. In this sense, the frequency theory is a purely philosophical theory—some people regard long run frequency as the true essence of probability and this conviction may help them apply probability even in situations where no real long runs of experiments exist.

Similarly, Bayesian statisticians assign probabilities to events in a way that appears objective to other observers. Some Bayesians may hold on to the view that, in fact, everything they do is subjective. This belief may help them apply probability even though there is a striking difference between their beliefs and actions. The subjective theory is a

purely philosophical theory in the sense that some people find comfort in "knowing" that in essence, probability is purely subjective, even if all statistical and scientific practice suggests otherwise.

Some readers will be disappointed by my "scientific laws of probability" (L1)-(L5) (to be presented in Chapter 5), because they are not much more than a formalization of a few intuitive "folk laws." Hence, my contribution may be no more than a codification of some universally accepted ideas. I find both formal and informal presentations of the currently popular philosophies, frequency and subjective, so misleading that even if (L1)-(L5) are no more than a clarification of some ideas, I consider them worth publishing. It is my strong opinion that both frequency and subjective theories are fundamentally and irreparably flawed, and I will present arguments to this effect. I expect that not all readers will share my point of view, to say the least, but I will be happy if they find (L1)-(L5) helpful in clarifying their own philosophical positions.

1.3. Organization of the material.

In the next chapter, I will review some elementary mathematical methods of probability and statistics—they will be needed later in the book, for example, in the discussion of the classical and Bayesian statistics. Next, I will give a brief description of the main trends in the philosophy of probability. That will be followed by two chapters devoted to science, first to my philosophical views on science in general, and second to my specific proposal for the scientific laws of probability. A separate chapter on decision theory will come next. My criticism of the two currently popular philosophies of probability will take the next four chapters; it will be split into the discussion of the frequency theory of probability, the corresponding branch of statistics ("classical statistics"), the subjective theory of probability, and the Bayesian statistics, (allegedly) based on the subjective philosophy. The remaining chapters will be devoted to teaching practices in the area probability, abuse of langauge in probability, and "concluding remarks" that did not fit anywhere else.

I will end the introduction with an explanation of the usage of a few terms, because readers who are not familiar with probability and statistics might be confused when I refer to "philosophy of probability" as a foundation for statistics rather than probability. I am a "probabilist;" among my colleagues, this word refers to a mathematician whose focus is a field of mathematics called "probability." An ideal probabilist is concerned exclusively with pure mathematics, but in fact most probabilists have strong interests in applications. The probability theory is applied in all natural sciences, social sciences, business, politics, etc., but there is only one field of natural science (as opposed to the deductive science of

mathematics) where probability is the central object of study and not just a tool—this field is called "statistics." For historical reasons, the phrase "philosophy of probability" refers to the philosophical and scientific foundations of statistics.

2. MATHEMATICAL METHODS OF PROBABILITY AND STATISTICS

This chapter is challenging for two reasons. First, I want to present a review of elementary methods of probability and statistics that would be a sufficient basis for the philosophical arguments presented in later chapters, but that would not take much space. I feel that it is hardly possible to achieve both goals simultaneously but I hope that my imperfect review is better than no review.

Another reason why this chapter is challenging is that it is hard to avoid controversy in the area of probability even at the introductory level (perhaps *especially* at the introductory level) as some probabilistic intuition corresponds to mass phenomena and some to decision problems in situations involving uncertainty. Nevertheless, I will try to paint an unbiased and accurate picture in this and the next chapter.

2.1. Probability.

The mathematics of probability is based on Kolmogorov's axioms. The fully rigorous presentation of the axioms requires some definitions from the "measure theory," a field of mathematics. This material is not needed in this book, so I will present the axioms in an elementary way. Any probabilistic model, no matter how complicated, is represented by a space of all possible outcomes Ω. The individual outcomes ω in Ω can be very simple (for example, "heads," if you toss a coin) or very complicated—a single outcome ω may represent the temperatures at all places around the globe over the next year, for example. Individual outcomes may be combined to form events. If you roll a die, individual outcomes ω are numbers $1, 2, \ldots, 6$, that is $\Omega = \{1, 2, 3, 4, 5, 6\}$. The event "even number of dots" is represented by a subset of Ω, specifically, by $\{2, 4, 6\}$. Every event has a probability, that is, a number between 0 and 1. Probabilities of some events are 0 or 1. If you roll a "fair" die then all outcomes are equally probable, that is, $P(1) = P(2) = \ldots = P(6) = 1/6$. Kolmogorov's axioms put only one restriction on probabilities—if events $A_1, A_2, \ldots, A_n$ are disjoint, that is, at most one of them can occur, then the probability that at least one of them will occur is the sum of probabilities of $A_1, A_2, \ldots, A_n$. In symbols,

$$P(A_1 \text{ or } A_2 \text{ or } \ldots \text{ or } A_n) = P(A_1) + P(A_2) + \ldots + P(A_n).$$

A curious feature of Kolmogorov's axiomatic system is that it does not include at all the notion of independence. We call two events (mathematically) independent if the probability of their joint occurrence is the product of their probabilities, in symbols,

$P(A \text{ and } B) = P(A)P(B)$. Without independence, the mathematical theory of probability and the applications of probability would not be meaningful branches of science. The intuitive meaning of independence is that the occurrence of one of the events does not give any information about the possibility of occurrence of the other event.

If a quantity X depends on the outcome ω of the experiment or observation then we call it a random variable. For example, if the experiment is a roll of two dice, the sum of dots is a random variable. If a random variable X may take values $x_1, x_2, \ldots, x_n$ with probabilities $p_1, p_2, \ldots, p_n$ then the number $EX = p_1x_1 + p_2x_2 + \ldots + p_nx_n$ is called the expected value or expectation of X. Intuitively speaking, the expectation of X is the (weighted) average, mean or central value of all possible values, although each one of these descriptions is questionable. The expected value of dots on a fair die is $1/6 \cdot 1 + 1/6 \cdot 2 + \ldots + 1/6 \cdot 6 = 3.5$. Note that the "expected value" of dots is not expected at all because the number of dots must be an integer.

The expectation of $(X - EX)^2$, that is, $E(X - EX)^2$ is called the variance of X and denoted $\mathrm{Var}X$. Its square root is called the standard deviation of X and denoted σ_X, that is $\sigma_X = \sqrt{\mathrm{Var}X}$. It is much easier to explain the intuitive meaning of the standard deviation than that of variance. Most random variables take values different from their expectations and the standard deviation signifies a typical difference between the value taken by the random variable and its expectation. The strange definition of the standard deviation, via variance and square root, has an excellent theoretical support, a mathematical result known as the Central Limit Theorem (CLT), to be reviewed next.

2.1.1. Strong Law of Large Numbers, Central Limit Theorem and Large Deviation Principle.

A sequence of random variables $X_1, X_2, X_3, \ldots$ is called i.i.d. if these random variables are independent and have identical distributions.

The Strong Law of Large Numbers says that if $X_1, X_2, X_3, \ldots$ are i.i.d. then the averages $(X_1 + X_2 + \ldots X_n)/n$ converge to EX_1 with probability 1.

A random variable Y is said to have the standard normal distribution if $P(Y < y) = (1/\sqrt{2\pi}) \int_0^y \exp(-x^2/2)dx$. Intuitively, the distribution of possible values of a standard normal random variable is represented by a bell-shaped curve centered at 0.

Suppose that $X_1, X_2, X_3, \ldots$ are i.i.d., the expectation of any of these random variables is μ and its standard deviation is σ. The Central Limit Theorem says that for large n, the normalized sum $(1/\sigma\sqrt{n}) \sum_{k=1}^n (X_k - \mu)$ has a distribution very close to the standard normal distribution.

Roughly speaking, the Large Deviation Principle (LDP) says that under appropriate assumptions, observing a value of a random variable far away from its mean has a probability much smaller than a naive intuition might suggest. For example, if X has the standard normal distribution, the probability that X will take a value greater than x is of order $(1/x)\exp(-x^2/2)$ for large x. The probability that the standard normal random variable will take a value 10 times greater than its standard deviation is about 10^{-38}. The Central Limit Theorem suggests that the Large Deviations Principle applies to sums or averages of sequences of i.i.d. random variables. In fact, it does, but the precise formulation of LDP will not be given here. The LDP-type estimates are not always as extremely small as the above example might suggest.

2.1.2. Exchangeability and de Finetti's theorem.

A permutation π of a set $\{1, 2, \ldots, n\}$ is any one-to-one function mapping this set into itself. A sequence of random variables $(X_1, X_2, \ldots, X_n)$ is called exchangeable if it has the same distribution as $(X_{\pi(1)}, X_{\pi(2)}, \ldots, X_{\pi(n)})$ for every permutation π of $\{1, 2, \ldots, n\}$. Informally, $(X_1, X_2, \ldots, X_n)$ are exchangeable if for any sequence of possible values of these random variables, any other ordering of the same values is equally likely. Recall that a sequence of random variables $(X_1, X_2, \ldots, X_n)$ is called i.i.d. if these random variables are independent and have identical distributions.

A celebrated theorem of de Finetti says that an infinite exchangeable sequence of random variables is a mixture of i.i.d. sequences. In other words, for any given infinite exchangeable sequence of random variables, one can generate a sequence with the same probabilistic properties by first choosing randomly a probability distribution $\mathcal{P}$ and then generating an i.i.d. sequence whose elements X_k have distribution $\mathcal{P}$. For example, you can generate an exchangeable sequence $(X_1, X_2, \ldots, X_n)$ by randomly deforming a coin and then tossing it n times; random variables X_k are indicators of heads here, that is, $X_k = 1$ if the k-th toss is heads and $X_k = 0$ otherwise.

2.2. Classical statistics.

Statistics is concerned with the analysis of data, although there is no unanimous agreement on whether this means "inference," that is, the search for the truth, or making decisions, or both.

One of the "classical" methods of statistics is "estimation"—I will explain it using an example. Suppose that you have a deformed coin and you would like to know the probability p of heads (this formulation of the problem contains an implicit assumption

that the probability p is objective). We can toss the coin n times and encode the results as a sequence of numbers (random variables) $X_1, X_2, \ldots, X_n$, with the convention that $X_k = 1$ if the result of the k-th toss is heads and $X_k = 0$ otherwise. Then we can calculate $\overline{p} = (X_1 + X_2 + \ldots + X_n)/n$, an "estimator" of p. The estimator $\overline{p}$ is our guess about the true value of p. One of its good properties is that it is "unbiased," that is, its expectation is equal to p. The standard deviation of $\overline{p}$ is $\sqrt{npq}$.

Another procedure used by classical statisticians is hypotheses testing. Consider the following drug-testing example. Suppose that a new drug is expected to give better results than an old drug. Doctors adopt (temporarily) a hypothesis H that the new drug is *not* better than the old drug and choose a *level of significance*, often 5% or 1%. Then they give one drug to one group of patients and the other drug to another group of patients. When the results are collected, the probability of the observed results is calculated, assuming the hypothesis H is true. If the probability is smaller than the significance level, the hypothesis H is rejected and the new drug is declared to be better than the old drug.

On the mathematical side, hypotheses testing proceeds along slightly different lines. Usually, at least two hypotheses are considered. Suppose that you can observe a random variable X whose distribution is either P_0 or P_1. Let H_0 be the hypothesis that the distribution is in fact P_0, and let H_1 be the hypothesis that the distribution of X is P_1. An appropriate number c is found, corresponding to the significance level. When X is observed and its value is x, the ratio of probabilities $P_0(X = x)/P_1(X = x)$ is calculated. If the ratio is less than c then the hypothesis H_0 is rejected and otherwise it is accepted. The constant c can be adjusted to make one of the two possible errors small: rejecting H_0 when it is true or accepting it when it is false.

Finally, I will outline the idea of a "confidence interval," as usual using an example. Suppose a scientist wants to find the value of a physical quantity θ. Assume further that he has at his disposal a measuring device that does not generate systematic errors, that is, the errors do not have a tendency to be mostly positive or mostly negative. Suppose that the measurements are $X_1, X_2, \ldots, X_n$. The average of these numbers, $\overline{X}_n = (X_1 + X_2 + \ldots + X_n)/n$, can be taken as an estimate of θ. The empirical standard deviation $\sigma_n = \sqrt{(1/n)\sum_{k=1}^{n}(X_k - \overline{X}_n)^2}$ is a measure of accuracy of the estimate. If the number of measurements is large, and some other assumptions are satisfied, the interval $(\overline{X}_n - \sigma_n, \overline{X}_n + \sigma_n)$ covers the true value of θ with probability equal to about 68%. If the length of the interval is increased to 4 standard deviations, that is, if we use $(\overline{X}_n - 2\sigma_n, \overline{X}_n + 2\sigma_n)$, the probability of coverage of the true value of θ becomes 95%.

2.3. Bayesian statistics.

The Bayesian statistics derives its name from the Bayes theorem. Here is a very simple version of the theorem. Let $P(A \mid B)$ denote the probability of an event A given the information that an event B occurred. The conditional probability $P(A \mid B)$ can be calculated by dividing the probability of a simultaneous occurrence of A and B by the probability of B. Suppose that events A_1 and A_2 cannot occur at the same time but one of them must occur. The Bayes theorem is the following formula,

$$P(A_1 \mid B) = \frac{P(B \mid A_1)P(A_1)}{P(B \mid A_1)P(A_1) + P(B \mid A_2)P(A_2)}.$$

Intuitively, the Bayes theorem is a form of a retrodiction, that is, it gives the probability of one of several causes (A_1 or A_2), given that an effect (B) has been observed.

One of the simplest examples of the Bayesian methods is the analysis of a deformed coin tossing. A popular Bayesian model assumes that the coin tosses are exchangeable. According to de Finetti's theorem, this is mathematically equivalent to the assumption that there exists an unknown number Θ (a random variable), between 0 and 1, representing the probability of heads on a single toss. If we assume that the value of Θ is θ then the sequence of tosses is i.i.d. with the probability of heads on a given toss equal to θ. The Bayesian analysis starts with a *prior* distribution of Θ. A typical choice is the uniform distribution on $[0, 1]$, that is, the probability that Θ is in a given subinterval of $[0, 1]$ of length r is equal to r. Suppose that the coin was tossed n times and k heads were observed. The Bayes theorem can be used to show that, given these observations and assuming the uniform prior for Θ, the *posterior* probability of heads on the $(n + 1)$-st toss is $(k + 1)/(n + 2)$. Some readers may be puzzled by the presence of constants 1 and 2 in the formula—one could expect the answer to be k/n. If we tossed the coin only once and the result was heads, then the Bayesian posterior probability of heads on the next toss is $(k + 1)/(n + 2) = 2/3$; this seems to be much more reasonable than $k/n = 1$.

3. MAIN PHILOSOPHIES OF PROBABILITY

My general classification of the main philosophies of probability is borrowed from [2] and [8]. However, I will pay much more attention to the frequency and subjective theories than to other theories, in contrast to [2] and [8], because these two theories are widely recognized as the foundation of modern statistics and other applications of probability. My presentation will not follow the chronological order—I will discuss less popular philosophies of probability first.

3.1. The classical theory.

The "classical" definition of probability gives a mathematical recipe for calculating probabilities in highly symmetric situations, such as tossing a coin, rolling a die or playing cards. It does not seem to be concerned with the question of the "true" nature of probability. In 1814, Laplace stated the definition in these words (English version after [2]):

> *The theory of chance consists in reducing all the events of the same kind to a certain number of cases equally possible, that is to say, to such as we may be equally undecided about in regard to their existence, and in determining the number of cases favorable to the event whose probability is sought. The ratio of this number to that of all the cases possible is the measure of this probability, which is thus simply a fraction whose numerator is the number of favorable cases and whose denominator is the number of all the cases possible.*

Since the definition applies only to those situations in which all outcomes are (known to be) equally "possible," it does not apply to a single toss or multiple tosses of a deformed coin. The definition does not make it clear what one should think about an experiment with a deformed coin—does the concept of probability apply to that situation at all? The classical definition seems to be circular because it refers to "equally possible" cases—this presumably means "equally probable" cases—and so probability is defined using the notion of probability.

The "classical philosophy of probability" is a modern label. That "philosophy" was a practical recipe and not a conscious attempt to create a philosophy of probability, unlike all other philosophies reviewed below. They were developed in the XX century, partly in parallel.

3.2. The logical theory.

The "logical" theory of probability maintains that probabilities are relations between sentences. They are weak forms of logical implication, intuitively speaking. According to this theory, the study of probability is a study of a (formal) language. Keynes and Carnap were the most prominent representatives of this philosophical view. The version advocated by Keynes allows for non-numerical probabilities. The logical theory is based on the Principle of Indifference which asserts that, informally speaking, equal probabilities should be assigned to alternatives for which no reason is known to be different. The Principle of Indifference does not have a unique interpretation. If you toss a deformed coin twice, what is the probability that the results will be different? The Principle of Indifference suggests that the answer should be 1/2 but a generalization of this claim to a large number of tosses implies that the tosses are independent and the probability of heads is 1/2 for each toss. This implies the impossibility of learning from experience. Since this conclusion is not palatable, Keynes and Carnap argued that the probability that the first two results will be different should be taken as 1/3.

The logical theory seems to be almost unknown among mathematicians, probabilists and statisticians. One reason is that some of the philosophical writings in this area, such as those of Carnap, are hard to follow for non-experts. Moreover, the emphasis on the logical aspect of probability seems to miss the point of the real difficulties with this concept. This is in sharp contrast with mathematics which benefited substantially from the careful examination of its logical foundations.

3.3. The propensity theory.

The term "propensity theory" is applied to recent philosophical theories of probability which consider probability an objective property of things or experiments just like mass or electrical charge. Popper developed the first version of the propensity theory.

The following example illustrates a problem with this interpretation of probability. Suppose a company manufactures identical computers in plants in Japan and Mexico. The propensity theory does not provide a convincing interpretation of the statement "This computer was made in Japan with probability 70%," because it is hard to imagine what physical property this sentence might refer to.

3.4. The frequency theory.

The development of the foundations of the mathematical theory of probability at the beginning of the nineteenth century is related to observations of the stability of relative

frequencies of some events in gambling. Much later, at the beginning of the twentieth century, von Mises formalized this idea using the concept of a collective. The collective is a long (ideally, infinite) sequence of isomorphic events or objects. Examples include plants in the field or molecules of gas in a macroscopic-size vessel. Von Mises defined the collective using its mathematical properties. For a sequence of observations to be a collective, the relative frequency of an event must converge to a limit as the number of observations grows. The limit is identified with the probability of the event. Von Mises wanted to eliminate from this definition some naturally occurring sequences that contained patterns. For example, many observations related to weather show seasonal patterns, and the same is true for some business activities. Von Mises did not consider such examples as collectives and so he imposed an extra condition that relative frequencies of the event should be equal along "all" subsequences of the collective. The meaning of "all" was the subject of much controversy and some non-trivial mathematical research. Of course, one of the subsequences is the sequence of those times when the event occurs but it is clear that this is not the meaning that the definition is trying to capture. Hence, one should limit oneself to subsequences chosen without clairvoyant powers, but as I said, this is harder to clarify and implement than it may seem at the first sight. The issue is further complicated by the fact that in real life, only finite sequences are available, and then the restriction to *all* sequences chosen without clairvoyant powers is not helpful at all.

Another, perhaps more intuitive, way to present the idea of a collective is to say that a collective is a sequence that admits no successful gambling system. This is well understood by owners of casinos and roulette players—the casino owners make sure that every roulette wheel is perfectly balanced (and so, the results of spins are a collective), while the players dream of finding a gambling system or, equivalently, a pattern in the results.

I will later argue that the problem with subsequences is quite tractable from the scientific point of view. The real problem with the frequency theory is that it has little to say about many situations when "collectives" are not present but probability theory is very successful.

3.5. The subjective theory.

Two people arrived independently at the idea of the subjective theory of probability: Ramsey and de Finetti. Ramsey did not live long enough to develop fully his thoughts so de Finetti was the founder and the best known representative of this school of thought.

The "subjective" theory of probability identifies probabilities with subjective opinions about unknown events. This idea is deceptively simple. First, the word "subjective" is

ambiguous and so I will spend a lot of time trying to clarify its meaning in the subjective philosophy. Second, one has to address the question of why the mathematical probability theory should be used at all, if there is no objective probability.

3.5.1. Interpreting subjectivity.

De Finetti emphatically denied the existence of any objective probabilistic statements or objective quantities representing probability. He summarized this in his famous saying "Probability does not exist." This slogan and the claim that "probability is subjective" are terribly ambiguous and lead to profound misunderstandings. Here are four interpretations of the slogans that come naturally to my mind.

(i) Although most people think that coin tosses and similar long run experiments displayed some patterns in the past, scientists determined that those patterns were figments of imagination, just like optical illusions.

(ii) Coin tosses and similar long run experiments displayed some patterns in the past but those patterns are irrelevant for the prediction of any future event.

(iii) The results of coin tosses will follow the pattern I choose, that is, if I think that the probability of heads is 0.7 then I will observe roughly 70% of heads in a long run of coin tosses.

(iv) Opinions about coin tosses vary widely among people.

Each one of the above interpretations is false in the sense that it is not what de Finetti said or what he was trying to say. The first interpretation involves "patterns" that can be understood in both objective and subjective sense. De Finetti never questioned the fact that some people noticed some (subjective) patterns in the past random experiments. De Finetti argued that people should be "consistent" in their probability assignments (I will explain the meaning of consistency momentarily), and that recommendation never included a suggestion that the (subjective) patterns observed in the past should be ignored in making one's own subjective predictions of the future, so (ii) is not a correct interpretation of de Finetti's ideas either. Clearly, de Finetti never claimed that one can affect future events just by thinking about them, as suggested by (iii). We know that de Finetti was aware of the clustering of people's opinions about some events, especially those in science, because he addressed this issue in his writings, so again (iv) is a false interpretation of the basic tenets of the subjective theory. I have to add that I will later argue that the subjective theory contains implicitly assertions (i) and (ii).

The above list and its discussion were supposed to convince the reader that interpreting

subjectivity is much harder than one may think. A more complete review of various meanings of subjectivity will be given in Section 9.5.

The correct interpretation of "subjectivity" of probability in de Finetti's theory requires some background. The necessity of presenting this background is a good pretext to review some problems facing the philosophy of probability. Hence, the next section will be a digression in this direction.

3.5.2. Verification of probabilistic statements.

The mathematics of probability was never very controversial. The search for a good set of mathematical axioms for the theory took over 100 years, until Kolmogorov came up with an idea of using measure theory in 1933. But even before then, the mathematical probability theory produced many excellent results. The challenge always lied in connecting the mathematical results and real life events. In a nutshell, if you have a real life event, how do you determine its probability? If you make a probabilistic statement, how do you verify whether it is true?

It is a good idea to have in mind a concrete elementary example. Take a coin and deform it with pliers so that it is not flat. What is the probability that it will fall heads up? Problems associated with this question and possible answers span a wide spectrum from practical to purely philosophical. Let us start with some practical problems. A natural way to determine the probability of heads for the deformed coin would be to toss the coin a large number of times and take the relative frequency of heads as the probability. This procedure is suggested by the frequency theory of probability. The first problem is that, in principle, we would have to toss the coin an infinite number of times. This, of course, is impossible, so we have to settle for a "large" number of tosses. How large is large?

Another practical problem is that a single event is often a member of two (or more) "natural" sequences. The experiment of tossing a deformed coin is an element of the sequence of tosses of the same deformed coin, but it is also an element of the sequence of experiments consisting of deforming a coin (a different coin every time) and then tossing it. It is possible that the frequency of heads will be 30% in the first sequence (because of the lack of symmetry) but it will be 50% in the second sequence (by symmetry).

On the philosophical side, circularity is one of the problems lurking when we try to define probability using long run frequencies. Even if we toss the coin a "large" number of times, it is clear that the relative frequency of heads is not necessarily equal to the probability of heads on single toss, but it is "close" to it. How close is close? One can use a mathematical technique to answer this question, such as a "confidence interval." A 95%

confidence interval covers the true value of the probability of heads with probability 95%. This probabilistic statement is meaningful only if we can verify it, but if the verification is based on the long run frequency idea, it will require constructing a long sequence of confidence intervals. This leads either to an infinite regress (sequence of sequences of sequences ...) or to a vicious circle of ideas (defining probability using probability).

Another philosophical problem concerns the relationship between a sequence of events and a single element of the sequence. If we could perform an infinite number of experiments and find the relative frequency of an event, that would presumably give us some information about other infinite sequences of similar experiments. But would that provide any information about any specific experiment, say, the seventh experiment in another run? In other words, can the observations of an infinite sequence provide a basis for the verification of a probability statement about any single event?

3.5.3. Subjectivity as an escape from the shackles of verification.

The previous section should have given the reader a taste of the nasty philosophical and practical problems related to the verification of probability statements. The radical idea of de Finetti was to get rid of all these problems in one swoop. He declared that probability statements cannot be verified at all—this is the fundamental meaning of subjectivity in his philosophical theory. This idea can be presented as a great triumph of thought or as a great failure. If you are an admirer of de Finetti, you may emphasize the simplicity and elegance of his solution of the verification problem. If you are his detractor, you may say that de Finetti could not find a solution to a philosophical problem, so he tried to conceal his failure by declaring that the problem was ill-posed. De Finetti's idea was fascinating but, alas, many fascinating ideas cannot be made to work. This is what I will show in Chapter 9.

I will now offer some further clarification of de Finetti's ideas. Probability statements are "subjective" in de Finetti's theory in the sense that "No probability statement is verifiable or falsifiable in any objective sense." Actually, according to de Finetti, probability statements are not verifiable in any sense, "subjective" or "objective." In his theory, when new information is available, it is not used to verify any probability statements made in the past. The subjective probabilities do not change at all—the only thing that happens is that one starts to use different probabilities, based on the old *and* new information. This does not affect the original probability assignments, except that they become irrelevant for making decisions—they are not falsified, according to de Finetti. The observation of an event or its complement cannot falsify or verify any statement about its probability.

One of the most important aspects of de Finetti's interpretation of subjectivity, perhaps *the* most important aspect, is that his philosophical theory is devoid of any means whatsoever of verifying any probability statement. This extreme position, not universally adopted by subjectivists, is an indispensable element of the theory; I will discuss this further in Chapter 9. A good illustration of this point is the following commentary of de Finetti on the fact that beliefs in some probability statements are common to all scientists, and so they seem to be objective and verifiable (quote after [2], page 70):

> *Our point of view remains in all cases the same: to show that there are rather profound psychological reasons which make the exact or approximate agreement that is observed between the opinions of different individuals very natural, but there are no reasons, rational, positive, or metaphysical, that can give this fact any meaning beyond that of a simple agreement of subjective opinions.*

The subjective theory is rich in ideas—no sarcasm is intended here. In the rest of this chapter, I will discuss some of these ideas: the "Dutch book" argument, the axiomatic system for the subjective theory, the identification of probabilities and decisions, and the Bayes theorem.

3.5.4. The Dutch book argument.

Probability does not exist in an objective sense, according to the subjective theory, so why should we use the probability calculus at all? The subjective theory justifies using probabilities using a "Dutch book" argument. Someone can make a Dutch book against me if I place various bets in such a way that no matter which events occur and which do not occur, I will lose some money. One can prove in a rigorous way that it is possible to make a Dutch book against me if and only if my "probabilities" are not "consistent", that is, they do not satisfy the usual formulas of the mathematical probability theory.

I will illustrate the idea with a simple example. Consider an experiment with only two possible (complementary) outcomes A and B, such as a toss of a deformed coin. In this elementary example, all that the mathematical theory of probability requires is that the probabilities of A and B add up to 1, that is, $P(A) + P(B) = 1$. Suppose that I harbor "inconsistent" views, that is, my personal assignment of probabilities does not satisfy $P(A) + P(B) = 1$, for example, $P(A) = 0.9$ and $P(B) = 0.8$. Since I am 90% sure that A will happen, I am willing to pay someone \$0.85, assuming that I will receive \$1.00 if A occurs (and nothing otherwise). The expected gain is positive because $\$0.15 \cdot 0.9 - \$0.85 \cdot 0.1 = \$0.05$, and so accepting this "bet" is to my advantage. A similar

calculation shows that since I am 80% sure that B will occur, it is to my advantage to pay \$0.75 in advance to get \$1.00 in case B occurs. If I place both bets, I will have to pay \$0.85 + \$0.75 = \$1.60 but no matter which event occurs, A or B, I will receive the payoff equal to \$1.00 only. In either case, I am going to lose \$0.60. A Dutch book was formed against me because I did not follow the usual rules of probability, that is, I used "probabilities" that did not satisfy the condition $P(A) + P(B) = 1$.

Consistency protects me against creating a situation resulting in certain loss and so I have to use the mathematics of probability in my judgments, the subjective theory advises. Note that the claim here is not that inconsistency will necessarily result in a Dutch book situation (in a given practical situation, there may be no bets to be made), but that consistency protects me against the Dutch book situation under all circumstances.

3.5.5. The axiomatic system.

The subjective theory of probability is sometimes introduced using an axiomatic system, as in [1]. This approach gives the subjective theory of probability the flavor of a mathematical (logical, formal) theory. An axiomatic system such as in [1] may start with statements about decisions, not probabilities. These statements are intuitively appealing, even obvious, just as one would expect from axioms. A typical example is an axiom that states that if a decision $\mathcal{A}$ is preferable to decision $\mathcal{B}$ and $\mathcal{B}$ is preferable to $\mathcal{C}$ then $\mathcal{A}$ is preferable to $\mathcal{C}$. One can prove that if someone's choices between various decisions satisfy the axioms then there exist a "utility function" and a probability distribution P on all possible events such that a decision $\mathcal{A}$ is preferable to a decision $\mathcal{B}$ if and only if the expected utility is greater if decision $\mathcal{A}$ is made. The utility function represents value of assets and gains that may have no natural or universal numerical values, such as real estate or even peace or love, but it also represents the "real" or "operational" value of money. One dollar gain has typically different utility for a pauper than for a millionaire. It is commonly assumed that the utility function is increasing and convex, that is, people prefer to have more money than less money (you can always give away the unwanted surplus), and the subjective satisfaction from the gain of an extra dollar is smaller and smaller as your fortune grows larger and larger.

One could argue that logical consistency is a desirable intellectual habit with good practical consequences but there exist some mathematical theories, such as non-Euclidean geometries, which do not represent anything real (at the human scale in ordinary life). Hence, adopting a set of axioms does not guarantee a success in practical life—one needs an extra argument, such as empirical verification, to justify the use of any given set of ax-

ioms. The subjective theory claims that probability statements cannot be verified (because probability does not exist in an objective sense) so this leaves the Dutch book argument as the only subjectivist justification for the use of the mathematical rules of probability and the implementation of the axiomatic system.

3.5.6. Identification of probabilities and decisions.

The identification of probabilities and decisions is an aspect of the axiomatic system. The axioms formalize intuitive ideas about rational decisions. Then a mathematical argument shows that if a set of decisions satisfies the axioms, then the decisions can be represented using probabilities and a utility function. The important point here is that probabilities obtained in this way do not correspond directly to anything objectively existing in the real world—they are only a way of encoding a set of consistent decisions. If we do not assume anything about the relationship between the decisions and the real world, there is nothing we can conclude about the relationship between the probabilities and the real world.

3.5.7. The Bayes theorem.

The subjective theory is implemented in the Bayesian statistics in a very specific way. The essence of statistics is the analysis of data so the subjective theory has to supply a method for incorporating the data into a consistent set of opinions. On the mathematical side, the procedure is called "conditioning," that is, if some new information is available, the holder of a consistent set of probabilistic opinions is supposed to use the probability distribution *conditional* on the extra information. The mathematical theorem that shows how to calculate the conditional probabilities is called the Bayes theorem (see Section 2.3). The consistent set of opinions held before the data are collected is called the "prior" and the probability distribution obtained from the prior and the data is called the "posterior."

4. WHAT IS SCIENCE?

I will later present a severe criticism of some existing philosophies of probability. I feel that, for my critique to be effective, I have to place it in a proper context. In other words, I have to outline my view of science in general.

It is a daunting task to explain the essence of science because the problem involves the fundamental questions of ontology and epistemology. It is not my intention to compete in this field with the greatest minds in philosophy. I will present a simple vision of science that is an appropriate basis for my discussion of theories of probability. I feel that many thinkers investigating probability were steeped in sophisticated and abstract theories of science that blinded them to some obvious facts about probability.

A unique feature of humans among all species is our ability to communicate using language. Many other species, from mammals to insects, can exchange some information between each other, but none of these examples comes even close to the effectiveness of human oral and written communication. The language gives us multiple sets of eyes and ears. Facts observed by other people are accessible to us via speech, books, radio, etc.

The wealth of available facts is a blessing and a problem. We often complain that we are overwhelmed with information. A simple solution to this problem emerged in human culture long time ago—data compression. Families of similar facts are arranged into patterns and only patterns are reported to other people. Pattern recognition is not only needed for data compression, it is also the basis of successful predictions. People generally assume that patterns observed in the past will continue in the future and so knowing patterns gives us an advantage in life (an important example of patterns are "laws of science"). Since not all people are equally good at pattern recognition, language communication gives them not only access to multiple sets of eyes and ears but also access to multiple brains.

The process described above is not perfect. Our own senses are imperfect, our own memory is imperfect and out own ability to recognize patterns is imperfect. Clearly, other people are afflicted by the same imperfections. On the top of that, communication adds its own errors. Some of them are random but some of them are typically human. What we say may be colored by our political or religious beliefs, for example. Some people pursue their goals by spreading misinformation, that is, they lie. Experience taught people to be somewhat skeptical abut information acquired from other people. Information is categorized and different batches of information are considered to be reliable to different degree. Science may be defined as the most respected and most reliable knowledge that

people offer to other people. Science achieved this status in various ways. Scientific claims are often repeatedly verified. The ethical standards imposed in science are much higher than, for example, in politics. I will illustrate these points by comparing science and religion. The utterly counterintuitive claims of the quantum theory and general relativity are widely accepted by populations as diverse as democratic societies and communist societies, Catholics and Muslims. Religion seems to lie at the other extreme of major ideologies—the humanity seems to have reconciled itself to the coexistence of various religions without any hope for the final coordination of their beliefs. In other words, religious information conveyed from one person to another may be met with total skepticism, especially if the two people are followers of different religions.

In view of the above, one of the primary roles of science is to present facts and patterns in the most reliable way. The most general patterns are called "laws" in natural sciences. The history of science showed that we cannot fully trust any laws, for example, the highly successful gravitation theory discovered (or invented) by Newton was later fundamentally revised by Einstein. The laws of science are the most reliable information on facts and patterns available at this time, they are not necessarily absolute truths.

The success of science (and human communication in general) depends very much on universality of certain basic perceptions. In other words, almost all people agree on certain basic facts, such as numbers and colors. When I look at five red apples, I am quite sure that any other person would also see five red apples, not seven green apples. Of course, we do make counting mistakes from time to time. The further we are from numbers, the harder it is to agree on directly perceived facts. If two people cannot agree on an answer to a question such as "Is the water in the lake cold or warm?", they can use a scientific approach to the question by translating the problem into the language of numbers. In this particular case, one can measure the water temperature using a thermometer. Numbers displayed by thermometers and other measuring devices are a highly reliable way to relay information between people. One has to note, however, that no scientific equipment or method can be a substitute for the prescientific agreement between different people on some basic facts. For example, suppose that a distance is measured and it is determined that it is 8 meters. A person may want to communicate this information to another person in writing. This depends on the ability of the other person to recognize the written symbol "8" as the number "eight." The problem cannot be resolved by measuring and describing the shape of the symbol because a report on the findings of such a procedure would contain words and symbols that might not be recognized by another person. The example seems

to be academic but it is less so if we think about pattern recognition by computers.

Different areas of intellectual activity are concerned with different aspects of reality and their success depends on the ability of different people to recognize the basic facts in the same way as it is done by other people. One of the main reasons for the success of the natural sciences is that most of their findings are based on directly and reliably recognizable facts, such as numbers, or they can be translated into such language. Measuring the spin of an electron is far beyond the ability of an ordinary person (and even most scientists) but the procedure can be reduced to highly reliable instructions and the results can be represented as numbers. The further we go away from natural sciences, the harder it is for people to agree, in part because they cannot agree even on the basic facts and perceptions. A statement that "Harsh laws lead to the alienation of the people" contains the words "harsh" and "alienation" whose meaning is not universally agreed upon. A very precise definition, legal-style, may be proposed for these words but such a definition need not be universally accepted.

My philosophy of science can be described as utilitarian and social. It is utilitarian because it stresses practical success, and it is social because it emphasizes the exchange of the information in the society. I see also behaviorist elements in my philosophy of science because it is concerned with the outward manifestations of beliefs. I expect science to give me reliable practical advice and I think most other people expect the same from science. The meaning and sources of this reliability are among the fundamental philosophical problems. However, both ordinary life and science must proceed forward with a simple and straightforward understanding of reliability, whether or not the philosophy can supply a theory on which we all could agree.

I am a strong critic of the subjective theory of probability but my own theory of science is somewhat subjective in the sense that it uses *human* interactions as its reference point. The difference between my theory and de Finetti's subjective theory is that his subjectivism implies the impossibility of exchanging objectively useful information between people in the area of probability, except for the raw facts. Recall that any agreement between people on probabilities is attributed by de Finetti totally to psychological effects.

I leave the question of objectivity of the universe and our knowledge to philosophers, because this is not a scientific question. Science adopted certain procedures and intellectual honesty requires that we follow them as closely as possible, if we want to call our activity science. A number of ideologies—political, philosophical and religious—tried to steal the prestige of science by presenting themselves as scientific. The subjective theory

of probability is one of them. What most people expect from science is not an "objective" knowledge in some abstract philosophical sense but an honest account of what other people learnt (or what they think they learnt) in their research. Scientists cannot say whether this knowledge is objective.

4.1. Decision making.

The unique characteristic of statistics among all natural sciences is that the decision theory is embedded in it in a seemingly inextricable way. I will try to separate the inseparable in Chapters 5 and 6. Here, I will only outline my philosophy of decision making in relation to my philosophy of science.

In deterministic situations, the decision making is not considered a part of science. For example, it is up to a physicist to find the melting temperature of gold (1064°C) but it is left to potential users of physics to implement this piece of scientific knowledge. If anybody needs to work with melted gold, he or she has to heat it to 1064°C. The decision to heat or not to heat a piece of gold is not considered a part of physics. The laws of deterministic sciences can be presented as instructions or logical implications: if you heat gold to 1064°C then it will melt. If you want to achieve a goal, all you have to do is to consult a book, find a law which explains how to achieve that goal, and implement the recipe. This simple procedure fails when a decision problem involves probability because the goal (the maximum possible gain, for example) often cannot be achieved with certainty. A standard decision theory model assumes that the decision maker would like to maximize his or her gain. If no decision maximizes the gain with certainty, the decision maker has to choose among available decisions using some criterion different from the sure attainability of the goal. The choice is not obvious in many practical situations and so decision making is historically a part of statistics—scientists feel that it would be unfair to leave this matter in the hands of lay people, who might not be sufficiently knowledgable about decision making.

The decision making problem is not scientific in nature. Science can predict the results of different decisions, sometimes with certainty and sometimes with some probability, but it is not the business of science to tell people what decisions they should make.

The identification of decision making and probability assignments by the subjective theory of probability is misleading. The identification is implicitly presented as a universal law; in fact, it is only a mathematical trick. The subjectivist claim that your decision preferences uniquely determine your probabilities (and vice versa) refers to nothing more than a purely abstract way of encoding your preferences using probabilities. This part of

the subjective theory shows only a mathematical possibility of passing from probabilities to decisions and the other way around, using a well defined mathematical algorithm. There is nothing automatic about the identification of decision preferences and probabilities, if we assume that probabilities (or some relations between them) are objective.

5. THE SCIENCE OF PROBABILITY

The science of probability must provide a recipe for assigning probabilities to real events. I will argue that the following list of five "laws of probability" is a good representation of our accumulated knowledge related to probabilistic phenomena and that it is a reasonably accurate representation of the actual applications of probability in science. I will also argue that my laws of probability are superior to the frequency and subjective theories.

(L1) Probabilities are numbers between 0 and 1, assigned to events whose outcome may be unknown.

(L2) If events A and B cannot happen at the same time, the probability that one of them will occur is the sum of probabilities of the individual events, that is, $P(A \text{ or } B) = P(A) + P(B)$.

(L3) If events A and B are physically unrelated then they are independent in the mathematical sense, that is, $P(A \text{ and } B) = P(A)P(B)$.

(L4) If there exists a symmetry on the space of possible outcomes which maps an event A onto an event B then the two events have equal probabilities, that is, $P(A) = P(B)$.

(L5) An event has probability 0 if and only if it cannot occur. An event has probability 1 if and only if it must occur.

A reader with some prior knowledge of probability theory must be surprised by (L1)-(L5). These laws look completely trivial and obvious. How can this be a new revolutionary approach to the probability theory? The obvious answer is that there is nothing really new about (L1)-(L5). I consider it an embarrassment to the scientific community that (L1)-(L5) are presented in textbooks only in an implicit way.

The discussion of (L1)-(L5) will be divided into many subsections, dealing with their various scientific and philosophical aspects.

5.1. Interpretation of (L1)-(L5).

The laws (L1)-(L5) should be easy to understand for anyone who has even a minimal experience with probability but nevertheless it is a good idea to spell out a few points.

(i) The relationship between (L1)-(L5) and the real probabilistic and statistical models is analogous to the relationship of, say, Maxwell's equations for electromagnetic fields and a model for the radio transmitter. The laws (L1)-(L5) are supposed to be the common denominator for a great variety of models, but there is no presumption that it should be

trivial to derive popular scientific models, such as linear regression or geometric Brownian motion (a model for the stock price) from (L1)-(L5) alone. One may find other conditions for probabilities besides (L1)-(L5) in some specific situations but none of those extra relations seems to be as fundamental or general as (L1)-(L5).

(ii) The word "symmetry" should be understood as any invariance under a transformation preserving the structure of the outcome space (model) and its relation to the outside world. Elementary symmetries include the mirror symmetry (left and right hands are symmetric in this sense), and translations in space and time.

A simple example of (L4) is the assertion that if you toss a coin (once) then the probability of heads is $1/2$. From the statistical point of view, the most important example to which (L4) applies is a sequence of exchangeable events. When observations are ordered chronologically, exchangeability can be thought of as a symmetry in time. Natural sciences provide many examples involving symmetries in space, not time. Examples include translation invariance in models of crystals and rotation invariance in models of planets.

I will now discuss a fundamentally important aspect of the interpretation of (L4). This law does not refer to the symmetry in a gap in our knowledge but it refers to the physical (scientific) symmetry in the problem. For example, we know that the ordering of the results of two tosses of a deformed coin does not affect the results. But we do not know how the asymmetry of the coin will affect the result of an individual toss. Hence, if T and H stand for "tails" and "heads" then TH and HT have equal probabilities according to (L4), but TT and HH do not necessarily have the same probabilities.

The above remark about the proper application of (L4) is closely related to the perennial discussion of whether the use of the "uniform" distribution can be justified in situations when we do not have any information. In other words, does the uniform distribution properly formalize the idea of the total lack of information? The short answer is "no." The laws (L1)-(L5) formalize best practices when some information is available and have nothing to say when there is no information available.

Let us have a look at the relationship between (L4) and the uniform distribution via some examples. Some random quantities can take values in an interval, for example, the percentage of vinegar in the mixture of vinegar and water can take values between 0% and 100%, that is, in the interval $[0, 1]$. If a quantity has the "uniform probability distribution" on $[0, 1]$ then its value is equally likely to be in any interval of the same length, for example, it is equally likely that the quantity is in any of the intervals $(0.1, 0.2)$, $(0.25, 0.35)$ or $(0.85, 0.95)$. The law (L4) can be applied to justify the use of the uniform

distribution only if there is an appropriate symmetry in the model. For example, if you throw darts and you want to describe the position of the dart in the board relative to the bull's eye, you may use a "random variable" Q to describe the angle between the line from the bull's eye to the dart, and a horizontal line emanating from the bull's eye to the right. If we measure the angle Q using degrees, the dart game is sufficiently symmetric (invariant under rotations around the bull's eye) for us to conclude that Q has a uniform distribution between 0 and 360. However, if you have a sample of vinegar solution in water and you do not know how it was prepared, there is no symmetry that would map the percentage of vinegar in the interval $(0.5, 0.6)$ onto the interval $(0.85, 0.95)$. In this case, (L4) does not support the use of the uniform distribution.

(iii) In relation to (L3), one should note that there exist pairs of events which are physically "related" but independent in the mathematical sense. If you roll a fair die, the event A that the number of dots is even is independent from the event B that the number is a multiple of 3, because $P(A \text{ and } B) = 1/6 = 1/2 \cdot 1/3 = P(A)P(B)$.

(iv) An implicit message in (L1)-(L5) is that there exist situations in which one cannot assign probabilities in a scientific way. Actually, laws (L1)-(L5) do not say how to assign values to probabilities—they only specify some conditions that the probabilities must satisfy. In some cases, such as tosses of a symmetric coin, (L1)-(L5) determine probabilities of all events in a unique way.

5.2. A philosophy of probability and scientific verification of (L1)-(L5).

It is best to keep philosophical and scientific aspects of the discussion of probability as far as possible because the history of probability teaches us that mixing the two sides of the phenomenon leads to confusion and bizarre theories. In principle, I could try to follow this advice but my own philosophical interpretation of (L1)-(L5) is so closely related to the scientific methods of verification of probabilistic statements that it would be artificial to keep them apart. I stress that my philosophical interpretation of (L1)-(L5) is independent of their scientific value—(L1)-(L5) should be judged by how well they describe known probabilistic phenomena, how well they match widely accepted scientific practices, and whether they generate reliable predictions.

My philosophy of probability says that the role of the probability theory is to identify events of probability 0 or 1 because these are the only probability values which can be verified or falsified by observations. In other words, knowing some probabilities between 0 and 1 has some value only to the extent that such probabilities can be used as input in calculations leading to the identification of events of probability 0 or 1.

Single events do have probabilities, if these probabilities have values 0 or 1. As for probabilities between 0 and 1, they can be thought of as catalysts needed to generate probabilities of interest, perhaps having no real meaning of their own. I personally think about all probabilities as "real" and "objective" but this is only to help me build a convenient image of the universe in my mind.

I am tempted to steal de Finetti's slogan "Probability does not exist" and give it a completely new meaning. Laws (L1)-(L5) may be interpreted as saying that probability does not exist as an independent physical quantity because probability can be reduced to symmetry, lack of physical influence, etc. This is similar to the reduction of the notion of "temperature" to the average energy of molecules, in physics.

Before I turn to the question of the scientific verification of (L1)-(L5), I will discuss the idea of a scientific proof. I have learnt it in the form of an anecdote told by a fellow mathematician. Mathematicians believe that they have the strictest standards of proof among all intellectuals and so they sometimes have a condescending view of the methods of natural sciences, including physics. This attitude is no doubt the result of the inferiority complex—physics discoveries find its way to the front page of The New York Times much more often than mathematical theorems. The idea of a "proof" in mathematics is this: you start with a small set of axioms and then you use a long chain of logical deductions to arrive at a statement that you consider interesting, elegant or important. Then you say that the statement has been proved. Physicists have a different idea of a "proof": you start with a large number of unrelated assumptions, you combine them into a single prediction, and you check if the prediction agrees with the observed data. If the agreement is within 20%, you call the assumptions proved.

The procedure for the verification of (L1)-(L5) I advocate resembles very much the "physics' proof." Consider a real system and assign probabilities to various events using (L1)-(L4), before observing any of these events. Then use the mathematical theory of probability to find an event A with probability very close to 1 and make a prediction that the event A will occur. The occurrence of A can be treated as a verification of the assignment of probabilities and its non-occurrence can be considered its falsification.

A very popular scientific method of verifying probability statements is based on repeated trials—this method is a special case of the general verification procedure and is obviously related to the frequency theory of probability. Suppose that we want to verify the statement $P(A) = 0.7$, where A is some event. Then we find events $A_1, A_2, \ldots, A_n$ such that n is large and the events $A, A_1, A_2, \ldots, A_n$ are i.i.d. (or exchangeable). Here,

"finding events" means designing an experiment with repeated trials. Sometimes repeated trials can be performed and sometimes they cannot—I will say more on this topic later on. The mathematics of probability says that if $P(A) = 0.7$ and $A, A_1, A_2, \ldots, A_n$ are i.i.d. then the observed relative frequency of events in the whole sequence will be very close to 70%, with very high probability. If the observed frequency is indeed close to 70%, this is considered a proof of the assertion that $P(A) = 0.7$ and the assumption that $A, A_1, A_2, \ldots, A_n$ are i.i.d. Otherwise, one concludes that the probability of A is different from 0.7, although in some circumstances one may instead question the assumption that $A, A_1, A_2, \ldots, A_n$ are i.i.d.

Recall the discussion of the two interpretations of the "proof," the mathematical one and the physical one. Traditionally, the philosophy of probability concerned itself only with the verification of probability statements—this resembles the mathematical proof. One needs to take the physics' attitude when it comes to the verification of probability assignments based on (L1)-(L5), or (L1)-(L5) themselves—it is not only the probability statements but also symmetries and lack of physical influence that are tested.

The general verification method described above works at (at least) two levels. It is normally used to verify specific probability assignments or relations. However, the combined effect of numerous instances of the application of this procedure constitutes a verification of the whole theory, that is, the laws (L1)-(L5).

The method of verification of (L1)-(L5) proposed above works only in the approximate sense, for practical and fundamental reasons: no events in the universe are absolutely "unrelated," no symmetry is perfect, mathematical calculations usually do not yield interesting events with probabilities exactly equal to 1, and the events of probability "very close" to 1 occur "almost" always, not always. A probabilist willing to test (L1)-(L5) using the method described above can choose an arbitrary degree of accuracy by seeking events with probabilities arbitrarily close to 1.

The fact that the verification procedure of (L1)-(L5) is imperfect in both practical and fundamental sense opens a door to an attack, especially from the subjectivist direction. To answer this challenge, I invoke my philosophy of science from Chapter 4 and I propose the following minimalist interpretation of (L1)-(L5). The laws (L1)-(L5) are no more than an account of the facts and patterns observed in the past—they are the best compromise (that I could find) between accuracy, objectivity, brevity, and utility in description of the past situations involving uncertainty. I will later argue that (L1)-(L5) are a better summary of what we have observed than the von Mises theory of collectives. The subjective theory

provides no such summary at all—this is one of the fatal flaws in that theory.

The actual implementation of experiments or observations designed to verify (L1)-(L5) is superfluous, except for didactic reasons. Scientists accumulated an enormous amount of data over the centuries and if someone thinks that the existing data do not provide a convincing support for (L1)-(L5) then there is little hope that any additional experiments or observations would make any difference.

5.3. Are (L1)-(L5) circular?

Laws (L3), (L4) and (L5) present some philosophical problems.

Law (L5) refers to events of probability 0 or 1. Practically no events of any interest have such probabilities, although there exist many important events whose probabilities are very close to 0 or 1. Recall that in my philosophy of science, (L1)-(L5) are an account of the past events and patterns. Hence, (L5) can be considered to be a philosophically imperfect but practical way of communicating past observations. A related problem is whether (L5) can be used for making predictions. Most people use an approximate form of (L5) in making their decisions and so they effectively believe in (L5). This, of course, does not prove that (L5) is objectively true, but this gives it support in my framework of philosophy of science, which stresses the exchange of information between people on procedures that they consider reliable.

From the philosophical point of view, (L3) and (L4) are the most important difference between the subjective theory of probability and my theory. The subjective theory set itself the goal of formalizing rational behavior in face of uncertainty. The starting point of that theory is the observation that in some situations we cannot predict, with any degree of certainty, the outcome of some events. The starting point of my theory is the observation that in some situations we seem to have all the relevant information concerning two experiments or observations, the information gives the same support for the same outcome in both cases, but experience shows that the outcomes are sometimes different. One of the main claims of this book is that the subjectivists' attempt at building a theory of rational behavior in face of uncertainty is a complete failure because there is very little that one can do if one does not assume any objective connection between a decision problem and the real world.

In my philosophy, the success of probability theory (that is, good predictions in the sense of (L5)) stems from objective information about the real world. My theory asserts implicitly that symmetries and physical independence are objective and that they can be

effectively used to make predictions. This categorical statement can be given the following weaker form: Any successful application of probability must be based on the (explicit or implicit) recognition of objective symmetries and physical independence, as in (L3) and (L4). If one adopts the philosophical position that objective symmetries do not exist or cannot be effectively recognized then the theory of probability becomes practically useless. I will later argue that the whole statistical and scientific practice shows that statisticians and scientists behave as if they believed in the objective symmetries and objective physical independence.

The rest of this section is devoted to a delicate question of philosophical nature. Although I have no doubt that (L1)-(L5) can be effectively applied in real life by real people (in fact they are), a careful philosophical scrutiny of (L3) and (L4) shows that they appear to be circular, or they lead to an infinite regress, and this seems to undermine their applicability.

One could rephrase (L3) as "If events are physically independent then they are mathematically independent," so to apply (L3), I have to know whether the two events are physically independent. How can I determine whether two events are physically independent or not? This knowledge can come from a long sequence of observations of similar events. If the occurrence of one of the events is not correlated with the occurrence of the other event, we may conclude that the events are physically unrelated. This seems to lead to a vicious circle of ideas—we can use (L3) and conclude that two events are independent only if we know that they are uncorrelated, that is independent.

The law (L4) applies to symmetric events, or, in other words, events invariant under a transformation. The concept of symmetry requires that we divide the properties of the two events into two classes. The first class contains the properties that are satisfied by both events, and the second class contains properties satisfied by only one event. Consider the simple example of tossing a deformed coin twice. Let A_1 be the event that the first toss results in heads and let A_2 denote heads on the second toss. The events must be different in some way or otherwise we would not be able to record two separate observations. In our example, A_1 and A_2 differ by the time of their occurrence. The two events have some properties in common, the most obvious being that the same coin is used in both cases. The law (L4) can be applied only if the properties that are different for the two events are physically unrelated to the outcome of the experiment or observation—such properties are needed to label the results. This brings us back to the discussion of (L3). It turns out that an effective application of (L4) requires an implicit application of (L3), and that law

seems to be circular.

A thorough and complete discussion of the problems outlined above would inevitably lead me into the depths of epistemology. I am neither inclined nor capable of fully analyzing the problem. Nevertheless, I will offer several arguments in defence of (L1)-(L5).

I am not aware of a probability theory that successfully avoids the problem of circularity of recognizing physically unrelated or symmetric events. If you want to apply the classical definition of probability, you have to recognize "cases equally possible." An application of the "principle of indifference" of the logical theory of probability presupposes the ability to recognize the situations when the principle applies, that is, those with some symmetries. The frequency theory is based on the notion of a "collective," involving a symmetry in an implicit way. If one cannot recognize a collective *a priori* then one will collect completely unrelated data. The subjective philosophy seems to be the only theory that successfully avoids the problem; the cost is a trifle—the subjective theory has nothing to report about the past observations, and makes no predictions.

Later on, I will show that the frequency theory and the subjective theory are meaningless without (L3) and (L4). Hence, if there is a genuine problem with these two laws, all philosophies of probability are severely affected.

The problem is not unique to the probability theory and (L1)-(L5). The ability to recognize events which are symmetric or physically unrelated is a fundamental element of any scientific activity. The need for this ability is considered so basic and self-evident that scientists almost never talk about it. Suppose a biologist wants to find out whether zebras are omnivorous. He has to go to Africa and observe a herd of zebras. This means finding a family of symmetric objects, characterized by black and white stripes. In particular, he must not mistake lions for zebras. Moreover, the zoologist must disregard any information that is unrelated to zebras, such as data on snowstorms in Siberia in the seventeenth century or car accidents in Brazil in the last decade. I will not try to enter into the discussion where the abilities to recognize independent or symmetric events come from (nature or nurture), what they really mean, and how reliable they are. The laws (L1)-(L5) are based on principles taken for granted elsewhere in science, if not in philosophy. All probabilistic, statistical, scientific and everyday experience tells me that the probability theory is very successful and hence even if a totally satisfactory resolution of this philosophical issue cannot be found, we should ignore this failure and not let it affect the way we formalize, apply and teach probability. I cannot prove beyond reasonable doubt that one can effectively recognize events that are disjoint, symmetric or physically unrelated. But I can prove that

if this cannot be achieved then the probability theory cannot be implemented with any degree of success.

The circularity of (L1)-(L5) discussed in this section resembles a bit a traditional philosophical view of the probability theory. According to that view, one can only generate probability values from some other probability values. In fact, this applies only to the mathematical theory. My laws (L1)-(L5) say that the primary material from which probabilities can be generated are not other probabilities but symmetries and physical independence in the real world.

5.4. Applications of (L1)-(L5): some examples.

Anyone who has ever had any contact with real science and its applications knows that (L1)-(L5) are a *de facto* scientific standard, just like Newton's laws of motion. I will give a few, mostly simple, examples. Some of them will be derived from real science, and some of them will be artificial, to illustrate some philosophical points.

First of all, (L3) and (L4) are used in probability is the same way as in the rest of science. Recall the example involving zebras in the last subsection. When a scientist wants to study some probabilistic phenomena, he often finds collections of symmetric objects. For example, if a doctor wants to study the coronary heart disease, he has to identify people, as opposed to animals, plants and rocks. This is considered so obvious that it is never mentioned in science. More realistically, physicians often study some human subpopulations, such as white males. This is a little more problematic because the definition of race is not clear-cut. Doctors apply (L3) by ignoring data on volcano eruptions on other planets and observations of moon eclipses.

Next, let us consider a truly probabilistic example of an application of (L3) and (L4). A popular model for random phenomena ranging from radioactive decay to telephone calls is a "Poisson process." This model is applied if we believe that the number of "arrivals" (for example, nuclear decays, telephone calls) in a given interval of time is unrelated to the number of arrivals in any other (disjoint) interval of time. The model can be applied only if we assume in addition a symmetry, specifically the "invariance under time shifts"—the number of arrivals in a time interval can depend on the length of the interval but not on its starting time. It can be proved, using the mathematical methods of probability, that the independence and symmetry described above completely specify the Poisson process except for its intensity, that is, the average number of arrivals in a unit amount of time.

Laws (L1)-(L5) are applied by all statisticians, classical and Bayesian. A typical statistical analysis starts with a "model," that is, a set of assumptions based on (L1)-(L5).

Here (L2), (L3) and (L4) are the most relevant laws, as in the example with the Poisson process. The laws (L1)-(L5) usually do not specify all (relations between) probabilities, such as the intensity in the case of the Poisson process. The intensity is considered by classical statisticians as an "unknown but fixed" parameter that has to be estimated using available data. Bayesian statisticians treat the unknown parameter as a random variable and give it a distribution known as a *prior*. The prior is not chosen according to (L1)-(L5). I will discuss the classical and Bayesian branches of statistics in much more detail later in the book. My point here is that (L1)-(L5) are used by both classical and Bayesian statisticians, and the application of these laws, especially (L3) and (L4), has nothing to do with the official philosophies adopted by the two branches of statistics. The classical statisticians apply (L3) and (L4) even if the available samples are small. The models (but not the priors) used by the Bayesian statisticians *de facto* follow the guidelines given in (L1)-(L5) and so they attract very little philosophical controversy. They can be and they are scrutinized only as much and in the same sense as any scientific model.

My next example has a different nature—I will show how the long run interpretation of probability fits into the framework of (L1)-(L5). I will later present a strong criticism of the von Mises theory but I cannot ignore this staple scientific application of probability.

To be concrete, consider a clinical test of a new drug. For simplicity, assume that the result of the test can be classified as a "success" or "failure" for each individual patient. Suppose now that you have a "large" number n of patients participating in the trial. There is an implicit other group of patients of size m, comprised of all people afflicted by the same malady in the general population. We apply the law (L4) to conclude that all $n+m$ patients form an "exchangeable" sequence. Choose an arbitrarily small number $\delta > 0$ describing your error tolerance and a probability p, arbitrarily close to 1, describing the level of confidence you desire. One can prove that for any numbers $\delta > 0$ and $p < 1$, one can find n_0 and m_0 such that for $n > n_0$ and $m > m_0$, the difference between the success rate of the drug among the patients in the clinical trials and the success rate in the general population will be smaller than δ with probability greater than p. One usually assumes that the general population is large and so $m > m_0$, whatever the value of m_0 might be. If the number of patients in the clinical trial is sufficiently large, that is, $n > n_0$, one can apply (L5) to treat the clinical trial results as the predictor of the future success rate of the drug.

In other applications of the idea of the long run frequency, the counterpart of the group of patients in the clinical trial may be a sequence of identical measurements of

an unknown constant. In such a case, the general population of patients has no explicit counterpart—this role is played by all future potential applications of the constant.

The next example is artificial although it is inspired by, and it somewhat resembles, a sequence of statistical problems and the corresponding recommendations made by a statistical consultant. The example is constructed to show that (L1)-(L5) are a much more faithful representation of what real statisticians (and other scientists) do than the representations offered by other philosophies of probability.

Let me start with an overview of the example. The idea is that one should generate a "random" sequence of zeroes and ones using a different experiment or mechanism for each of the numbers in the sequence. Some existing philosophies of probability can correctly analyze the example but only if all the numbers are generated in the same way. All philosophies fail if the experiments change with every number.

The simplest way of generating a sequence in question would be to use (i) a coin toss for the first number (heads gives a one), (ii) a roll of a die for the second number (an even number is recorded as one), (iii) drawing of a card from a deck (a red card yields a one), etc. This is a very good way of generating a sequence that I need, but one may soon run out of ideas for new experiments—there is a limited number of possibilities related to gambling, such as roulette, lottery machines, or darts. I guess one could construct some other contraptions, with some physical symmetry built in.

There is a different way of constructing the sequence, based on a well known idea of generating an "unbiased" sequence of zeroes and ones from a biased one. To generate a single number in the sequence, we will need two trials, each with two possible results that I will call a success (S) and failure (F). The trials have to be "identical" but the success does not have to have the same probability as the failure on a single trial. Such sequences of experiments, of arbitrary length, are known as Bernoulli trials. For example, one can toss a deformed coin. If the results of the trials are SF then we add a zero to our sequence of numbers, when the results are FS, we add a one, and otherwise we do not add any number to our sequence. This method of generating the sequence is based not on the physical symmetry in space (the ordinary coins are flat) but on the symmetry in time, that is, exchangeability. This method of generating the sequence is easier to implement because we are not tied to physically symmetric objects, usually man-made. Instead, we could use two pieces of data from different (exchangeable) data sets for different elements of the sequence, for example. There is an ample supply of exchangeable data sets collected by various scientists in the course of various unrelated projects.

Finally, I will describe a more sophisticated version of the last method of generating the sequence. There are two reasons for using this more complicated method. First, I want to avoid a potential problem in the analysis of the example, arising from the possibility that the labels "success" and "failure" are assigned randomly in the mind of the observer and the symmetry between SF and FS has more to do with the symmetry of our thoughts than with any real symmetry in the universe. Second, I will need the most complicated version to expose a problem with the logical theory of probability. My final method of generating the sequence requires that we perform three trials to generate just one 0 or 1. Zero is associated with SFF, one is recorded if the results are FFS, and any other result of the three trials generates no number.

For definiteness, suppose that we generate a sequence consisting of 500 zeroes and ones.

To put the philosophical analysis of the example into the proper context, I will start with a description of what real *people* (as opposed to abstract theories) would say. All statisticians and all scientists would agree that the probability of generating a zero is the same as that for a one at every stage of the procedure, by symmetry. The jargon used by different statisticians may depend on their philosophical preferences. Some will talk about the two or three trials in a single experiment as "i.i.d." and some will call them "exchangeable." Some statisticians would call the probabilities subjective, some would claim that they are objective, and some would express no such opinion. I cannot imagine that there would be any sizeable number of dissenters that would disagree with the claim that the numbers in the sequence are independent and the probability of a zero at a given place in the sequence is 1/2.

Hence, everybody would agree on all probabilities for all events related to our sequence. In the parlance of probability theory, all would agree that the number of zeros in the sequence would have the binomial distribution with parameters 500 and 1/2. This implies, for example, that the probability of observing 100 or fewer zeros (call this event A) is of order 10^{-43}. No matter which formal language a statistician or scientist may choose to express this finding, the practical meaning of this probabilistic assertion is clear—event A will not happen.

Note that (L1)-(L5) properly formalize the scientific analysis of this experiment, just as it would be done in real life. The symmetry and (L4) imply that the probability of a zero is the same as that of a one, for every element of the sequence. The lack of physical influence and (L3) imply that the elements of the sequence are (mathematically)

independent. Finally, (L5) implies that A is impossible—this is a testable prediction.

I remark parenthetically that the above analysis shows that (L4) is useful outside the context of very long sequences.

I will argue that no other philosophical theory of probability is equally successful in the analysis of this rather simple if slightly artificial example.

(i) It is not clear whether the classical definition of probability can be used to derive the statement that 0 and 1 have the same probability to appear at a given place in the sequence. In support of this conclusion, one may recall one of the crucial phrases from that definition: "cases equally possible." On the other hand, the definition talks about "all cases possible" and it is not unfair to interpret this as the limitation of that definition to those experiments in which all atoms of the sample space are equally probable. In our case, that would rule out most or all of the two-trial or three-trial experiments because the probability of a success in any one of them is not presumed to be 1/2.

Even if we assume that the classical definition actually implies that 0 and 1 have the same probability, another question surfaces—are the consecutive elements of our sequence of 0's and 1's independent, according to that definition? The definition seems to leave this in the hands of the person analyzing the model by talking about the "cases ... we may be equally undecided about ..." Hence, the question of whether the elements of the sequence are independent because the experiments are unrelated is swept under the rug. My general impression is that the classical definition fails to provide an adequate analysis of the example.

(ii) The logical theory correctly assigns identical probabilities to zeros and ones. However, the same theory makes also other, highly controversial, probability assignments. Assume that we use three-trial experiments to generate the sequence. There are several versions of the logical theory. According to the most important one, the probability of exactly one success in a sequence of three Bernoulli trials (with unknown probability of success) is 1/4. Hence, the logical theory says that the probability of generating a 0 or 1 in a single experiment is $1/4 \cdot 2/3 = 1/6$. This implies that a 0 or 1 will be generated in about 1/6 of all three-trial experiments. I do not think anybody would support this prediction and it is obvious that in some implementations of the example, the prediction would decisively fail.

Some other versions of the logical theory do not assign values to all probabilities so they might not generate false predictions.

(iii) The frequency theory has little to say about our example. If and when one

actually observes the sequence, the frequency of 0's in the sequence will be close to 50%. Hence, one can say that the probability of 0 in the sequence is 1/2. However, one does not need any philosophy to observe that the proportion of 0's in the sequence is close to 50%. The real question is whether the frequency theory can predict that the frequency of 0's in the sequence will be 50%. The answer is negative because no element of the example is a collective in the von Mises sense. A single three-trial experiment is too short to be a collective, and the consecutive three-trial experiments are not isomorphic so they also fail to form a collective.

(iv) The subjective theory makes no predictions whatsoever concerning the outcome of the experiments or the properties of the sequence of 0's and 1's. This is because the subjective theory claims that no verifiable ("objective") statements can be made about probabilities. The subjective theory is a normative theory which tells its followers what to do—"you have to be consistent." In the present case this only means that the probabilities assigned to various events have to obey the usual mathematical formulas for probabilities.

The frequency and subjective theories earned most respect among scientists yet, paradoxically, these theories have the least to say about our simple example.

5.5. Probability of a single event.

In some cases, such as tosses of a symmetric coin, laws (L1)-(L5) not only impose some relationships between probabilities but also determine probabilities of individual events. If an event has probability (close to) 0 or 1, this value can be verified or falsified by the observation of the event. An implicit message in (L1)-(L5) is that if an event has a probability (much) different from 0 or 1, this value cannot be verified or falsified. One has to ask then: Does this probability exist in an objective sense?

As long as probability assignments are not used to make a prediction (that is, to find an event of probability 0 or 1), they cannot be verified using (L1)-(L5). However, probability assignments which are not intended to make predictions by one person, may be used for that purpose by another person (the "other person" should be understood in a generalized sense, as a community, for example). Suppose someone thinks that if you toss a coin then the probability of heads is 1/3 and not 1/2. If that person tosses a coin only a few times in his lifetime, he will not be able to make a prediction related to the tosses and verify or falsify his belief about the probability of heads. Now suppose that there is a widespread belief in a certain community that the probability of heads is 1/3, and every individual member of the community tosses coins only a few times in his or her

lifetime. Then no individual in this population will be able to verify or falsify his beliefs, assuming that the members of the community do not discuss coin tosses with one another. Suppose an anthropologist visits this strange community, interviews the people about their probabilistic beliefs and collects the data on the results of coin tosses performed by various people in the community. She will see a great discrepancy between the aggregated results of coin tosses and the prevalent probabilistic beliefs. This artificial example is inspired by some real social phenomena. It has been observed that lotteries are more popular in poor communities than in affluent communities. There are many reasons why this is the case, but one of them might be the lack of understanding of probability among the poorer and supposedly less educated people. A single poor person is unlikely to be able to verify his beliefs about the probability of winning a lottery by observing his own winnings or losses (because winnings are very rare). But someone who has access to cumulative data on the lottery winnings and beliefs of gamblers may be able to detect that the members of poor communities overestimate the probability of winning.

A version of the above argument applies to a single person. Let us assume that for any event whose probability is 1/2 according to (L1)-(L5), one cannot prove in any way that assigning probability 1/3 to this event is incorrect. Suppose that someone plans to toss a coin 500 times, declares that the (future) results of the trials are independent and assigns probability 1/2 to heads on each trial. Since she knows that a single event does not really have a probability, she then changes the probability of heads on the first trial to 1/3, but retains the values of other probabilities as 1/2. It is impossible to detect in any way whether this new probability assignment is less correct than the original one, by the assumption we have made. She then changes the probability assigned to the heads on the second trial from 1/2 to 1/3, and so on. In the end, she assigns probability 1/3 to heads on all trials, and the argument given so far implies that there is no way in which one could prove that the end result of all the changes is incorrect. Of course, the results of the 500 coin tosses will clearly show that the probability of heads is 1/2, and this disproves the original assumption.

The above argument exploits the discrepancy between sharp transitions in idealized theoretical models and continuity of the real phenomena. It is often impossible to detect the effect of a single violation of (L1)-(L5) in probability assignments, because the effect is too small to be detected in any convincing way. But the combined effect of many incorrect probability assignments is readily detectable. This observation is not specific to probability—consider the following non-probabilistic example. Suppose that a ship can

carry x tonnes of cargo. It will surely be able to carry x tonnes and one extra atom of iron. This applies to every x, so either the ship cannot carry any cargo at all, or it can carry an unlimited amount of cargo. The paradox is easily solved—the ship's tonnage is not a well-defined number; the ship may float or sink depending on a number of circumstances. Similarly, a probabilistic prediction, in the sense of (L5), is not associated with a fixed cut-off probability close to 1.

The question of whether my arguments prove conclusively that (L1)-(L5) have to be applied to single (isolated) events, is left to the reader. There seems to be no evidence that an application of (L1)-(L5) ever resulted in adverse effects so one could apply here the following version of the principle of William of Occam—one should choose the simplest strategy among all strategies that give the same effect. Hence, one should apply (L1)-(L5) under all circumstances because no simpler strategy has been shown to be better.

5.6. On events that belong to two sequences.

A good way to test a philosophical or scientific theory is to see what it has to say about a well known problem. Suppose an event belongs to two exchangeable sequences. For example, we may be interested in the probability that a certain Mr. Winston, smoking cigarettes, will die of a heart attack. Suppose further that we know the relevant statistics for smokers of either sex, and also statistics for men (smokers and non-smokers combined), but there are no statistics for smoking men. If the long run frequencies are 60% and 50% in the two groups for which statistics are available, what are the chances of the death from a heart attack for Mr. Winston?

Laws (L1)-(L5) show that the question does not have a natural scientific answer. One needs symmetry to apply (L4), the most relevant law here. However, Mr. Winston is unique because we know something about him that we do not know about any other individual in the population. For all other individuals included in the data, we either do not know their sex or whether they smoke.

5.7. Symmetry and theories of probability.

Law (L4) is the core of the system (L1)-(L5), especially when it is applied to exchangeable events in the statistical context. The importance of exchangeable events has been recognized by each of the main philosophies of probability, under different names: "equally possible cases" in the classical theory, "principle of indifference" in the logical theory, "collective" in the frequency theory and "exchangeability" in the subjective theory. None of these philosophies got it right.

The classical theory of probability was based on symmetry although the term "symmetry" did not appear in the classical definition of probability. Since the definition used the words "all cases possible," it was applicable only in highly symmetric situations, where all atoms of the outcome space had the same probability. I think it would be stretching the reality to claim that the classical definition of probability could be used to derive the statement that in two tosses of a deformed coin, the events HT and TH have the same probability. The classical philosophy missed the important point that symmetry is useful even if not all elements of the outcome space have the same probability. Since the classical philosophy was not a conscious attempt to build a complete philosophical theory of probability but a byproduct of scientific investigation, one may interpret the shortcomings of the classical theory as incompleteness rather than as an error.

Law (L4) is built into the logical theory under the name of the "Principle of Indifference." This principle seems to apply to situations where there is inadequate knowledge, while (L4) must be applied only in situations when some relevant knowledge is available, and according to what we know, the events are symmetric. For example, we know that the ordering of the results of two tosses of a deformed coin does not affect the results. But we do not know how the asymmetry of the coin will affect the results. Hence, TH and HT have equal probabilities, but TT and HH do not. According to some versions of the logical theory, the probability of TT is $1/4$ or $1/3$. This leads to some clearly false predictions, such as in the last example of Section 5.4. The problem with these versions of the logical theory is that it extends the principle of indifference to situations with no known physical symmetry. The logical philosophy constructs a theory of probability within a formal language since natural languages are too chaotic. Hence, I believe that laws (L1)-(L5) can be properly presented in a suitable formal language and a correct logical theory of probability can be developed. I do not consider building such a formal theory useful from the scientific point of view so I will not make an attempt in this direction.

The frequency theory made the "collective" (population, sequence of events) its central concept. Collectives are infinite in theory and they are presumed to be very large in practice. Law (L4) is implicit in the definition of the collective because the collective seems to be no more than an awkward definition of an exchangeable sequence of events. To apply the frequency theory in practice, one has to be able to recognize long sequences invariant under permutations (that is, collectives or equivalently, exchangeable sequences), and so one has to use symmetry as in (L4). The frequency theory failed to recognize that (L4) is useful outside the context of collectives, that is, very long exchangeable sequences

(see Section 5.4).

The subjective theory's attitude towards (L4) is the most curious among all the theories. Exchangeability is clearly a central concept, perhaps *the* central concept, in de Finetti's system of thought, on the scientific side. These healthy scientific instincts of de Finetti gave way to his philosophical views, alas. His philosophical theory stresses absolute subjectivity of all probability statements and so deprives (L4) of any meaning beyond a free and arbitrary choice of (some) individuals. All Bayesian statisticians and other subjectivists use symmetries in their probability assignments just like everybody else. Yet the subjective theory of probability insists that none of these probability assignments can be proved to be correct in any objective sense.

5.8. Are coin tosses i.i.d. or exchangeable?

Consider tosses of a deformed coin. One may argue that they are independent (and so i.i.d., by symmetry and (L4)) because the result of any toss cannot physically influence any other result, and so (L3) applies. Note that (L1)-(L5) cannot be used to determine the probability of heads on a given toss.

An alternative view is that results of some tosses can give information about other results, so the coin tosses are not independent. For example, if you observe 90 heads in the first 100 tosses, you are likely to think that there will be more heads than tails in the next 100 tosses. The obvious symmetry and (L4) make the tosses exchangeable.

De Finetti's theorem (see Section 2.1.2) shows that both ways of thinking about coin tosses are equivalent in terms of putting mathematical restrictions on probabilities, so it does not matter whether one thinks about coin tosses as i.i.d. with unknown probability of heads or regards them as an exchangeable sequence.

6. DECISION MAKING

I will divide my discussion of decision making into several parts. Sections 6.1.1-6.1.5 will deal with decision making options for someone who uses (L1)-(L5) and has a sufficient understanding of the situation so that (L1)-(L5) determine the relevant probabilities. Section 6.2 will address the question of what to do when (L1)-(L5) do not specify probabilities needed to make a decision. Section 6.3 will be concerned, in a sense, with the decision of whether to adopt (L1)-(L5).

6.1. Decision making in the context of (L1)-(L5).

Recall the discussion of decision making from Section 4.1. Decision making is not a part of science. Science can (try to) predict the consequences of various decisions but it is not the role of science to tell people what they should do.

I will start with a semi-formal description of a simple but non-trivial probabilistic decision making problem. Suppose one has to choose between two decisions, D_1 and D_2. Suppose that if decision D_1 is made, the gain may take two values G_{11} and G_{12}, with probabilities p_{11} and p_{12}. Similarly, D_2 may result in rewards G_{21} and G_{22}, with probabilities p_{21} and p_{22}. Assume that $p_{11}+p_{12}=1$ and $p_{21}+p_{22}=1$, all four probabilities are strictly between 0 and 1, and $G_{11}<G_{21}<G_{22}<G_{12}$ so that there is no obvious reason why D_1 or D_2 is preferable to the other decision. Recall that, in this section, I assume that the four probabilities, p_{11}, p_{12}, p_{21} and p_{22}, are "known" in the sense that they are determined by (L1)-(L5). The question I am going to address is: Which of the decisions D_1 and D_2 is preferable? Actually, I will quickly move towards somewhat more realistic and interesting decision situations, but the above simple example sets the context for my discussion.

I will start by criticizing the most popular philosophy of decision making in face of uncertainty and then I will propose two other decision making philosophies.

6.1.1. Maximization of expected gain.

A standard decision making philosophy is to choose a decision which maximizes the expected gain. This decision making philosophy is quite intuitive but I will show that it has profound difficulties.

If we apply this decision making strategy to the example given above, we should make decision D_1 if $G_{11}p_{11}+G_{12}p_{12}>G_{21}p_{21}+G_{22}p_{22}$, and we should choose D_2 if the inequality goes the other way (the decisions are equally preferable if the expected values are equal).

First of all, the phrase "expected value" is misleading. Typically, the "expected value" is not expected at all. Everybody knows that the "expected number" of dots on a fair die is 3.5 but I sometimes wonder how many people subconsciously ignore this simple lesson. To emphasize the true nature of the "expected value", let me temporarily switch to an equivalent but much less suggestive term "first moment." Needless to say, "maximizing the first moment of the gain" sounds much less attractive than "maximizing the expected value of the gain." Why should one try to maximize the first moment of the gain and not minimize the third moment of the gain? I will address the question from both frequency and subjective points of view.

The frequency theory of probability identifies the probability of an event with the limit of relative frequencies of the event in an infinite sequence of i.i.d. events (that is, a collective). Similarly, the expected value (first moment) of a random variable is identified with the limit of averages in an infinite sequence of i.i.d. random variables, by the Strong Law of Large Numbers. If we want to use the frequency theory as a justification for maximizing of the first moment of the gain, we have to assume that we have a long sequence of i.i.d. decision problems and the same decision is made every time. Only in rare practical situations, one decision maker deals with an i.i.d. sequence of decision problems. A single decision maker usually faces decision problems that are not isomorphic; in everyday life, decision problems have often completely different structure, while in science and business, the form of the problems may sometimes remain the same but the information gained in the course of analyzing earlier problems may be applied in later problems and so the whole sequence is not invariant under permutations (it is not i.i.d.). The frequency theory of probability provides a direct justification for the practice of maximizing of the expected gain only on rare occasions.

Maximizing of the expected gain within the subjective theory of probability seems to be a reasonable strategy for the same reason as in the case of the frequency theory—linguistic. The subjective theory says that the only goal that can be achieved by choosing a consistent set of probabilities is the avoidance of the "Dutch book" situation. There are countless ways in which one can achieve consistency and none of them is any better than any other way in any objective sense, according to the subjective theory. The idea of "maximizing of expected gain" clearly exploits the subconscious associations of decision makers. They think that their gain will be large, if they choose a decision which maximizes the expected gain. The subjective theory says that the gain can be large or small (that is, it can be any number in the range of possible gains corresponding to a given decision)

but one cannot prove in any objective sense that the gain will be large. Moreover, the subjective theory teaches that when the gain is realized, its size cannot prove or disprove in the objective sense any claim about optimality or suboptimality of the decision that was made. Hence, maximizing the expected gain really means maximizing the subjective feelings about the gain. This sounds like a piece of advice from a "self-help" book rather than science.

Within the subjective philosophy, the idea of maximizing of the subjective gain is tautological. The prior distribution can be presented in various formal ways. One of them is to represent the prior as a set of beliefs containing conditional statements of the form "if the data turn out to be x then my preferred decision will be $D(x)$." Since in the subjective theory, probabilities and expectations are only a way of encoding consistent human preferences, an equivalent form of this statement is "given the data x, the decision $D(x)$ maximizes the expected gain." Hence the question of why you would like to maximize the expected gain is equivalent to the question of why you think that the prior distribution is what it is. In the subjective philosophy, it is not true that you should choose the decision which maximizes the expected gain; the decision that maximizes the expected gain was labelled so because you said you preferred it over all other decisions.

A different way to present the same criticism is to say that the idea of maximizing the expected gain within the framework of the subjective theory is circular. In that theory, one starts with a set of decisions satisfying certain intuitive rules. Probabilities are derived from the decisions using a mathematical procedure. Finally, when the posterior distribution is calculated, a decision is chosen that maximizes the expected value of the gain. Hence, the maximization of the expected value of the gain is nothing but a label given to the process of coordination of various decisions. The subjective theory does not have any advice on how to choose a consistent set of decisions, and in fact asserts that there is no objective way to choose a good consistent set of decisions.

6.1.2. Maximization of expected gain as an axiom.

Before I propose my own two alternative decision making philosophies, I have to mention an obvious, but repugnant to me, philosophical choice—one can adopt the maximization of the expected gain as an axiom. I have already argued that the choice of the decision strategy is not a part of the science of probability and so this axiom cannot be shown to be objectively correct or incorrect, except in some special situations. Hence, I am grudgingly willing to accept this choice of the decision philosophy, if anyone wants to make this choice. At the same time I strongly believe that the choice is based on a linguistic

illusion. If the same axiom were phrased as "one should maximize the first moment of the gain," most people would demand a good explanation for such a choice. And I have already shown that the justifications given by the frequency and subjective theories are insufficient.

The real answer to the question "Why is it a good idea to maximize the expected gain?" seems to be more technical than philosophical in nature. A very good technical reason to use expectations is that they are additive, that is, the expectation of the sum of two random variables is the sum of their expectations, no matter how dependent the random variables are. This is very convenient in many mathematical arguments. The second reason is that assigning a single value to each decision makes all decisions comparable, so one can always find the "best" decision. This is often an illusion based on a clever manipulation of the language, but many people demand answers, even poor answers, no matter what.

The maximization of the expected gain can be justified, at least in a limited way, within each of the two decision making philosophies proposed below. I find that approach much more palatable than the outright adoption of the expected gain maximization as an axiom.

6.1.3. Stochastic ordering of decisions.

The first of my own proposals for a decision philosophy is based on an idea that probability is the only quantity that distinguishes various events within the probability theory. I will use an analogy to clarify this point. Consider two samples of sulphur, one spherical and one cubic in shape. If they have the same mass, they are indistinguishable from the point of view of chemistry. Similarly, two balls made of different materials but with the same radii and the same density would be indistinguishable from the point of view of the gravitation theory. Consider two games, one involving a fair coin and the other involving a fair die. Suppose that you can win \$1 if the coin falls heads, and lose \$2 otherwise. You can win \$1 if the number of dots on the die is even, and otherwise you lose \$2. Since the probabilities are the only quantities that matter in this situation, one should be indifferent between the two games.

Now consider two games whose payoffs are known and suppose that they are stochastically ordered, that is, their payoffs G_1 and G_2 satisfy $P(G_1 \geq x) \geq P(G_2 \geq x)$ for all x. It is elementary to see that there exist two other games with payoffs H_1 and H_2 such that G_k has the same distribution as H_k for $k = 1, 2$, and $P(H_1 \geq H_2) = 1$. The game with payoff H_1 is obviously more desirable than the one with payoff H_2, and by the equiva-

lence described in the previous paragraph, the game with payoff G_1 is more desirable than the one with payoff G_2. In other words, the decision making philosophy proposed here says that a decision is preferable to another decision if and only if its payoff stochastically majorizes the payoff of the other decision.

Here are some properties of the proposed decision making recipe.

(i) Consider two decisions and suppose that each one can result in a gain of either $\$a$ or $\$b$. Then the gain distributions are comparable. In this simple case, the proposed decision algorithm agrees with the maximization of the expected gain.

(ii) Two decisions can be comparable even if their expected gains are infinite (that is equal to plus or minus infinity), or undefined.

(iii) If two decisions are comparable and the associated gains have finite expectations, a decision is preferable to another decisions if and only if the associated expected gain is larger than the analogous quantity for the other decision.

(iv) Consider a composite decision problem consisting of two decision problems, with two decisions available in each case. It may happen that the decisions are comparable in each problem and the first decision is the best in each case, but the aggregate of the first decisions is not comparable to the aggregate of the second decisions. This effect does not occur when the two decision problems in the composite problem are independent. From the technical point of view, this problem with the proposed decision making algorithm is disappointing. But the common sense dictates that in some cases it is crucial to see the "big picture"—one should make a decision only after taking into account all related decision problems. One should not insist on designing a decision making strategy that would be able to analyze every decision problem in isolation from other decision problems.

(v) One can justify the idea of maximizing of expected gain (under some circumstances) using the idea of stochastic ordering of decisions. Suppose that one has to deal with n decision problems, and each time one can choose between two decisions whose (random) gains are G_k^1 and G_k^2. If $EG_k^1 - EG_k^2 \geq 0$ for every k, the difference $EG_k^1 - EG_k^2$ is reasonably large, n is not too small, and the variances of G_k^j's are not too large then $G_1^1 + \ldots + G_n^1$ is either truly or approximately stochastically larger than $G_1^2 + \ldots + G_n^2$. Hence, it is beneficial to maximize the expected gain in every of the n decision problems. The conclusion requires a number of assumptions but they might be satisfied (at least in some approximate sense) in many situations. This justification of the idea of maximizing of the expected gain does not refer to the law of large numbers because it is not based on the approximate equality of $G_1^j + \ldots + G_n^j$ and its expectation. The number n of decision

problems does not have to be large at all—the justification works for moderate n but the cutoff value for n depends significantly on the joint distribution of G_k^1's and G_k^2's.

(vi) An obvious drawback of the proposed decision making philosophy is that not all decisions are comparable. Recall the utility function used by the subjective theory. I will make a reasonable assumption that all utility functions are non-decreasing. It is easy to show that two decisions are comparable if and only if one of the decisions has greater expected utility than the other for every utility function. Hence, the proposed ordering of decisions is consistent with the subjective philosophy in the following sense. In those situations in which the probabilities are undisputable, two decisions are comparable if and only if all decision makers, with arbitrary non-decreasing utility functions, would make the same choice. Let me use the last remark as a pretext to point out a weakness in the subjective theory. The comparability of all decisions in the subjective theory is an illusion because the ordering of decisions is strictly subjective, that is, it depends on an individual decision maker. He or she can change the ordering of decisions by fiat at any time, so the ordering has hardly any meaning. The subjective theory recommends avoiding the Dutch book situation but this restriction rarely puts meaningful bounds on real life choices.

6.1.4. Creating certainty.

My second proposal for a decision making strategy is better adapted to laws (L1)-(L5), especially (L5), than the "stochastic ordering" presented in the previous subsection.

The basic idea is quite old, it goes back to Cournot in the first half of the nineteenth century (quoted after Primas [4], p. 585):

> *If the probability of an event is sufficiently small, one should act in a way as if this event will not occur at a solitary realization.*

Note that (L5) is a mirror image of Cournot's statement in the sense that he refers to the future and (L5) refers to the past—(L5) is a way of summarizing the past observations, according to its minimalist interpretation presented in Section 5.2. Cournot's recommendation contains no explicit message concerning events which have probabilities different from 0 or 1. My proposal is to limit the probability-based decision making only to the cases covered by Cournot's assertion. In other words, I postulate that the probabilistic and statistical analysis should make certainty its goal. It is best to explain this abstract idea with an example of statistical flavor.

Suppose that one has to face a large number of independent decision problems, and at the k-th stage, one has a choice between decisions with payoffs G_k^1 and G_k^2, satisfying

$EG_k^1 = x_1$, $EG_k^2 = x_2 < x_1$, $\mathrm{Var}G_k^j \leq 1$. If one chooses the first decision every time, the average gain for the first n decisions will be approximately equal to x_1. Much of statistics is based on the Central Limit Theorem, which implies in our case that the average gain for the first n decisions will be of order $x_1 \pm 1/\sqrt{n}$. The proposed decision making philosophy has more to do with the Large Deviations Principle. A consequence of the Large Deviations Principle is that the probability $P(\sum_{k=1}^n G_k^1/n \geq (x_1 + x_2)/2)$ goes to 0 exponentially fast as n goes to infinity, and so it can be assumed to be zero for all practical purposes, even for moderately large n. This and a similar estimate for $P(\sum_{k=1}^n G_k^2/n \leq (x_1 + x_2)/2)$ imply that making the first decision n times yields a higher gain than the gain from making the second decision n times, with probability p_n very close to 1. Here, "very close" means that $1 - p_n$ is exponentially small in n. Such a fast rate of convergence is considered excellent in the present computer science-dominated intellectual climate.

We see that the traditional curse of statistics, the slow rate of convergence $(1/\sqrt{n})$, does not apply if we choose our goal appropriately. I conjecture that, subconsciously, decision makers believe only in these statements that are based on the Large Deviations Principle and pay only limited attention to those based on the Central Limit Theorem.

The proposed decision making strategy is partly based on the realization that in the course of real life we routinely ignore events of extremely small probability, such as being hit by a falling meteor. Acting otherwise would make life unbearable and anyhow would be doomed to failure, as nobody could possibly account for all events of extremely small probability. Hence, an application of the Large Deviations Principle can reduce the uncertainty to levels which are routinely ignored in normal life, out of necessity.

Clearly, the decision making strategy proposed in this section yields applicable advice in fewer situations than that proposed in the previous section. In my opinion, it is better to set goals for oneself that can be realistically and reliably attained rather than to deceive oneself into thinking that one can find a good recipe for success under any circumstances.

6.1.5. A new prisoner paradox.

This section contains an example, partly meant to illustrate the two decision making philosophies discussed in the last two subsections, and partly meant to be a respite from the dry philosophical arguments.

Imagine that you live in a medieval kingdom. Its ruler, King Seyab, is known for his love of mathematics and philosophy, and for cruelty. As a very young king, 40 years ago, he ordered a group of wise men to take an urn and fill it with 1000 white and black balls. The color of each ball was chosen by a coin flip, independently of other balls. There

is no reason to doubt wise men's honesty or accuracy in fulfilling king's order. The king examined the contents of the urn and filled another urn with 510 black and 490 white balls. The contents of the two urns is top secret and the subjects of King Seyab never discuss it.

The laws of the kingdom are very harsh, many ordinary crimes are punished by death, and the courts are encouraged to met out the capital punishment. On average, one person is sentenced to death each day. The people sentenced to death cannot appeal for mercy but are given a chance to survive by the following strange decree of the monarch. The prisoner on the death row can sample 999 balls from the original urn. He is told that the second urn contains 1000 balls, 490 of which are white. Then he can either take the last ball from the first urn or take a single random ball from the second urn. The prisoner's life will be spared if the ball turns out to be white. If the ball is white, one cannot be sentenced to death for the second time.

Now imagine that you have been falsely accused of squaring a circle and sentenced to death. You have sampled 999 balls from the first urn. The sample contains 479 white balls. You have been told that the second urn contains 490 white and 510 black balls. Will you take the last ball from the first urn or sample a single ball from the second one?

In view of how the balls were originally chosen for the first urn, the probability of the last ball being white is 0.50. The probability of sampling a white ball from the second urn is only 0.49. It seems that taking the last ball from the first urn is the optimal decision. However, you know that over 40 years, the survival rate for those who took the last ball from the first urn was either 48% or 47.9%. The survival rate for those who sampled from the second urn was about 49%. What would your decision be?

According to the "stochastic ordering" philosophy of decision making, you should take the last ball from the first urn. The decision philosophy based on "creating certainty" suggests that one should take a ball from the second urn, because the only meaningful event in this context, which has probability close to 1, is that the long run survival rates in the groups of prisoners taking balls from the first urn and second urn are about 48% and 49%, respectively. Of course, a prisoner may regard these long run frequencies as irrelevant to his own quandary

6.2. Events with no probabilities.

So far, my discussion of decision making was limited to situations where the probabilities were known. This section examines a decision maker options in the situation when (L1)-(L5) do not determine the relevant probabilities.

One of the great and undisputable victories of the subjectivist propaganda machine

is the widespread belief that there is a rational way to choose an action in any situation involving uncertainty. Many of the people who otherwise do not agree with the subjective theory of probability, seem to think that it is a genuine intellectual achievement of the subjective theory to provide a framework for making decisions in the absence of relevant and useful information.

What can other sciences offer in the absence of information or relevant theories? A physicist cannot give advice on how to build a plane flying at twice the speed of light or how to make a room temperature superconductor. Some things cannot be done because the laws of science prohibit them, and some things cannot be done because we have not learnt how to do them yet (and perhaps we never will). Nobody expects a physicist to give an "imperfect but adequate" advice in every situation (you cannot build a plane which "more or less" flies at twice the speed of light or make a superconductor which works at "more or less" room temperature). No such leniency is shown towards probabilists and statisticians by people who take the subjectivist ideology seriously—if probability is subjective then there is no situation in which you will lack anything to make probability assignments. And, moreover, if you are consistent, you cannot be wrong!

What should one do in a situation involving uncertainty if no relevant information is available? An honest and rather obvious answer is that there are situations in which the probability theory has no scientific advice to offer because no relevant probability laws or relations are known. This is not anything we, probabilists, should be ashamed of.

The form of laws (L1)-(L5) may shed some light on the problem. The laws do not give a recipe for assigning values to all probabilities. They only say that in some circumstances, the probabilities must satisfy some conditions. If no relevant relations, such as lack of physical influence or symmetry, are known then laws (L1)-(L5) are not applicable and any assignment of values to probabilities is arbitrary. Note that every event is involved in some relation listed in (L1)-(L5), for example, all events on Earth are physically unrelated to a supernova explosion in a distant galaxy (except for some astronomical observations). Hence, strictly speaking, (L1)-(L5) are always applicable but the point of the science of probability is to find sufficiently many relevant relations between events so that one can find useful events of very high probability and then apply (L5) to make a prediction.

One could argue that in a real life situation, one has to make a decision and hence one always (implicitly) assigns values to probabilities—in this limited sense, probability always exists. However, the same argument clearly fails to establish that "useful relations between events can be always found." By forcing someone to make a decision, you can

force that person to make implicit probability assignments. But you cannot force anyone to find useful relationships between probabilities which are the basis of (L1)-(L5). This reminds me of one of the known problems with torture (besides being inhumane): you can make every person to talk, but you cannot be sure that what you hear is true.

On the practical side of the matter, it is clear that people use a lot of science in their everyday lives in an intuitive or instinctive way. Whenever we walk, lift objects, pour water, etc., we use laws of physics, more often than not at a subconscious level. We are quite successful with these informal applications of science although not always so. The same applies to probability—a combination of intuition, instinct, and reasoning based on analogy and continuity can give very good practical results. This however cannot be taken as a proof that one can always assign values to all probabilities and attack every decision problem in a scientifically justified, rational way. As long as we stay in the realm of informal, intuitive science, we have to trim our expectations and accept whatever results our innate abilities might generate.

6.3. Law enforcement.

I think that the area of law enforcement provides excellent opportunities to document the total disconnection between the official philosophies of probability and the real applications of probability.

Consider the following two criminal cases. In the first case, a house was burgled and some time later, Mr. A.B., a suspect, was arrested. None of the stolen items were ever recovered but the police found a piece of evidence suggesting his involvement in the crime. The owners of the house kept a ten-letter code to their bank safe in their home safe. The home safe was broken into during the burglary. The search of Mr. A.B. yielded a piece of paper with the same ten letters as the code stored by the home owners. In court, Mr. A.B. maintained that he randomly scribbled the ten letters on a piece of paper, out of boredom, waiting at a bus stop. The prosecution based its case on the utter improbability of the coincidental agreement between the two ten-letter codes, especially since the safe code was generated randomly and so it did not contain any obvious elements such as a name.

The other case involved Mr. C.D. who shot and killed his neighbor, angered by a noisy party. In court, Mr. C.D. claimed that he just wanted to scare his neighbor with a gun. He admitted that he had pointed the gun at the neighbor from three feet and pulled the trigger but remarked that guns not always fire when the trigger is pulled, and the target is sometimes missed. Under questioning, Mr. C.D. admitted that he had had years of target practice, that his gun fired about 99.9% of time, and he missed the target about 1% of

time. Despite his experience with guns, Mr. C.D. estimated the chance of hurting the neighbor as 1 in a billion.

I am convinced that no court in the world would hesitate to convict both defendants (except possibly for some American juries—they delivered some truly amazing verdicts in recent history).

The conviction of both defendants would be based on the utter implausibility of their claims. Each of the defendants, though, could invoke one of the official philosophies of probability to strengthen his case. In the case of Mr. A.B., the frequency theory says that no probabilistic statements can be made because no long run of isomorphic observations ("collective") is involved. Specifically, a sequence of only ten letters cannot be called long. Likewise, the police could not find a long run of burglaries involving stolen codes. One could suggest running computer simulations of ten random letters, but Mr. A.B. would object—in his view, computer simulations are completely different from the workings of his brain, especially when he is "inspired."

Mr. C.D. could invoke the subjective theory of probability. No matter what his experience with guns had been, his assessment of the probability of killing the neighbor was as good as any other assessment, because probability is subjective. Hence, the killing of the neighbor should have been considered an "act of God" and not a first degree murder, according to Mr. C.D.

Needless to say, societies do not tolerate and cannot tolerate interpretations of probability presented above. People are required to recognize probabilities according to (L1)-(L5) and when they fail, or when they pretend that they fail, they are punished. A universal (although implicit) presumption is that (L1)-(L5) can be effectively implemented by members of the society. If you hit somebody on the head with a brick, it will not help you to claim that it was your opinion that the brick had the same weight as a feather. The society effectively assumes that weight is an objective quantity and requires its members to properly assess the weight. The society might not have explicitly proclaimed that probability is objective but it effectively treats the probability laws (L1)-(L5) as objective laws of science and enforces this implicit view on its members.

There are countless examples of views—scientific, philosophical, religious, political—that used to be almost universal at one time and changed completely at a later time. The universal recognition or implementation of some views does not prove that they are true. However, what is relevant to this discussion is not the universal enforcement of (L1)-(L5), but the fact that neither frequency theory supporters nor subjectivists object to this

situation in the least. I have no evidence that any statistician would have much sympathy for the probabilistic arguments brought up by the two defendants in my examples. The frequency and subjective probabilists use their philosophies only when they find them convenient and otherwise they use common sense—something I am trying to formalize as (L1)-(L5).

6.4. Identification of decisions and probabilities.

I will reiterate some thoughts on decisions and probabilities made in Section 6.1.1, taking a slightly different angle. The subjective theory of probability identifies decisions and probabilities. I will ignore an element of that identification, the "utility" function, because it is irrelevant to my discussion in this section. Every set of consistent decisions corresponds to a probability distribution, that is, a consistent (probabilistic) view of the world, and vice versa, any probability distribution defines a consistent set of decisions. This suggests that the whole discussion of decision making is redundant. This is the case only if we assume that objective probabilities do not exist. If objective probabilities (or relations between probabilities) exist then the identification of probabilities and decisions is simply not true. If objective probabilities exist, decision makers can use them in various ways. The subjectivist claim that your decisions uniquely determine your probabilities is nothing more than a way of encoding your decisions, of giving them labels. In principle, these labels may have nothing to do with objective probabilities.

7. FREQUENCY THEORY OF PROBABILITY

This chapter is devoted to a detailed critique of the frequency theory. I will show that the frequency theory fails to account for a number of common scientific uses of probability and that it is meaningless without the ability to recognize symmetries. Hence, in view of (L1)-(L5), the concept of a collective is redundant.

7.1. Probability does not rely on i.i.d. sequences.

I will present two classes of examples where the frequency theory fails to provide a foundation for established scientific methods. A large number of sequences of random variables encountered in scientific practice and real life applications are not i.i.d. or exchangeable but form "stochastic processes." Some of the best known classes of stochastic processes are Markov processes, stationary processes and Gaussian processes. Markov processes represent randomly evolving systems with short or no memory. Stationary processes are invariant under time shifts, that is, if we start observations of the process today, the sequence of observations will have the same probabilistic characteristics as is we started yesterday or tomorrow. Gaussian processes are harder to explain because their definition is somewhat technical. They are closely related to the Gaussian (normal) distribution which arises in the Central Limit Theorem and has the characteristic bell shape. One can make excellent predictions based on a single trajectory of any of these processes. Predictions may be based on various mathematical results such as the "ergodic" theorem or the extreme value theory. In some cases, one can transform the process mathematically into a sequence of i.i.d. random variables. However, even in cases when this is possible, this purely mathematical procedure is artificial and has little to do with von Mises' collectives. The frequency theory is useless as a scientific theory in such cases and it does not supply any valuable philosophical interpretation either.

Another class of examples when the frequency theory is miles apart from the real science are situations involving very small probabilities. Suppose someone invites you to play the following game. He writes a 20-digit number on a piece of paper, without showing it to you. You have to pay him \$10 for an opportunity to win \$1,000, if you guess the number. Anyone who has even a basic knowledge of probability would decline to play the game because the probability of winning is a meager 10^{-20}. According to the frequency theory, we cannot talk about the probability of winning as long as there is no long run of identical games. The frequency theory has no advice to offer here although no scientist would have any problems with making a rational choice. More practical examples involve

all kinds of very unlikely events, for example, natural disasters. Some dams are built in the US to withstand floods that may occur once every 500 hundred years. We would have to wait many thousands of years to observe a reasonably long sequence of such floods. In that time, several new civilizations might succeed ours. According to the frequency theory, it makes no sense to talk about the probability that dams will withstand floods for the next 100 years. There are many events that are not proven to be impossible but have probability so small that they are considered impossible in practice, and they do not fit into any reasonable long run of events. It is generally believed that yeti does not exist, that there is no life on Venus and that the US banks will not collapse in our lifetime. If we take the frequency theory seriously, we cannot make any assertions about probabilities of these events—this is a sure recipe for the total paralysis of life as we know it.

7.2. Definition of a collective.

The concept of a "collective" invented by von Mises is just an awkward version of "exchangeability," favored by de Finetti. The idea is very well understood by scientists, because they deal with populations, repeated observations, and sequences of identical experiments. One can try to formalize this idea in several different ways. On the mathematical side, there seem to be only two options: an exchangeable sequence or an "i.i.d." (independent identically distributed) sequence. Exchangeability is a form of symmetry—any rearrangement of a possible sequence of results is equally probable as the original sequence. The idea of an i.i.d. sequence stresses the independence (lack of physical influence) of one experiment in a series from another experiment in the same series, given the information about the probabilities of various results for a single experiment. In interesting practical applications, this information is missing, and then, by de Finetti's theorem, an i.i.d. sequence can be treated as an exchangeable sequence.

The definition of a collective is really mathematical, not scientific, in nature. The definition requires that for a given event, the relative frequency of that event in the sequence (collective) converges, and the same is true for "every" subsequence of the collective (the limit must be always the same). Here "every" is limited to subsequences chosen without clairvoyant powers, because otherwise we would have to account for the subsequence consisting only of those times when the event occurred, and similarly for the subsequence consisting of those times when the event did not occur. The limits along these subsequences are 1 and 0, of course. I think that the modern probability theory provides excellent technical tools to express this idea—the limit must be the same along any sequence of stopping times, with probability 1 (but not simultaneously along every such subsequence, with prob-

ability 1). This mathematical development comes too late, though. Some quite interesting mathematical approaches to this problem were proposed in the past but they do not seem to have any significance beyond pure philosophy.

The requirement that the relative frequencies have same limits along "all" subsequences is especially hard to interpret if one has a finite (but possibly long) sequence. In this case, we necessarily have limits 1 and 0 along some subsequences, and it is hard to find a good justification for eliminating these subsequences from our considerations. The purpose of the requirement that the limit is the same along all subsequences is to disallow sequences that contain patterns, such as seasonal or daily patterns. For example, temperatures at a given location show strong daily and seasonal patterns so temperature readings do not qualify as a collective. Surprisingly, this seemingly philosophically intractable aspect of the definition of a collective turned out to be tractable in practice in quite a reasonable way. One of the important tools used by modern statistics and science are random number generators. These are either clever algebraic algorithms (generating "pseudo-random" numbers) or, more and more popular, electronic devices generating random numbers (from thermal noise, for example). From the practical point of view, it is crucial to check that a given random number generator does not produce numbers that contain patterns, and so there is a field of science devoted to the analysis of random number generators. The results seem to be very satisfactory in that most statisticians and scientists can find a random number generator sufficiently devoid of patterns to meet their needs. In this special sense, von Mises is vindicated—it is possible to check in practice if a sequence is a collective.

7.3. Collectives and symmetry.

The problem with the collectives lies somewhere else. A scientific theory has to be applicable in the sense that its laws have to be formulated using terms that correspond to real objects and quantities observable in some reasonable sense. There is more than one way to translate the theory of collectives into an implementable theory. If we use a collective as an observable, we will impose a heavy burden on all scientists, because they will have to check for the lack of patterns in all potential collectives. This is done for random number generators out of necessity and in some other practical situations when the provenance of a sequence is not fully understood. But to impose this requirement on all potential collectives would halt the science. An alternative way is to identify a collective with an exchangeable sequence. The invariance under permutations (that is, the defining feature of an exchangeable sequence) can be ascertained in a direct way in many practical situations—this eliminates the need for testing for patterns. This approach is based on a

symmetry, and so it implicitly refers to (L4) and more generally, to (L1)-(L5). I conclude that either it is impossible to implement the concept of a collective or the concept is redundant.

There is another, closely related, reason why the concept of a collective is almost useless without (L1)-(L5). Typically, when a scientist determines a probability by performing a large number of experiments or collecting a large number of observations, she wants to apply this knowledge in some other context; in a sense this is the essence of science. If we base probability theory on the concept of a collective, we will have to apply knowledge acquired by examining one collective to some other collective. A possible way to do that would be to combine the two collectives into one sequence and check if it is a collective. This theoretical possibility can be implemented in practice in two ways. First, one could apply a series of tests to see if the combined sequence is a collective—this would be a solid but highly impractical approach, because of its high cost in terms of labor. The other possibility is to decide that the combined sequence is a collective (an exchangeable sequence) using (L4), that is, to recognize the invariance of the combined sequence under permutations. This is a cost-efficient method but since it is based on (L4), it makes the concept of the collective redundant.

A different way to present the same argument is the following. Mathematics is used in science to reduce the number of measurements or to make predictions. A scientist makes a few measurements and then uses mathematical formulas appropriate for a given science to find values of some other quantities. If we adopt the frequency view of probability, the only predictions offered by this theory are the predictions involving limits of long run relative frequencies. In the case of a single collective, the theory makes only a single, rather weak, prediction, that the limit exists. Moreover, one has to recognize the collective in the first place, and this is either hard, as indicated in the last paragraph, or uses (L4). The mathematical theory of probability is very rich but all the interesting results involve more than two probabilities, often by making some general assumptions on families of probabilities (recall the example of the Poisson process from Section 5.4). In most practical situations, one deals either with one collective (for example, a single sequence of measurements of a physical constant), or two collectives (for example, a group of patients undergoing clinical trials, and the remaining patients in the country). The situations when one considers a larger number of collectives are extremely rare. Hence, the frequency view of the probability theory as a calculus for certain classes of infinite sequences is purely abstract and has no real applications.

7.4. Imaginary collectives.

If a real collective does not exist or cannot be created in a reasonable way, can one at least imagine a collective containing a given event and apply the theory of collectives in this way?

There are several problems with imaginary collectives. Since we do not have direct access to anyone's mind, imaginary collectives have no operational meaning. In other words, we cannot check whether anyone actually imagines any collectives. Hence, we can use imagination in our own research or decision making but our imagined collectives cannot be a part of a meaningful scientific theory. Contemporary computers coupled with robots equipped with sensors can do practically everything that humans can do (at least in principle) except for mimicking human mind functions. In other words, we can program a computer or robot to collect data, analyze them, make a decision and implement it. We cannot program a computer to imagine collectives and it is irrelevant whether we will be ever able to build computers with an imagination—the imagination would not make them any more useful in this context.

A different problem with imagined collectives is that in many (perhaps all) cases one can imagine more than one collective containing a given event. In many such cases, the probability of the event is different in the two imagined collectives. Consider a single toss of a deformed coin. This single event can be imagined to be a part of the collective of tosses of the same deformed coin, or a part of a collective of experiments consisting of deforming different coins and tossing each one of them once. Both collectives are quite natural and one can easily perform both types of experiments.

8. CLASSICAL STATISTICS

It is natural to suppose that the classical statistics is (implicitly?) based on the frequency philosophy of probability, although this assumption is not quite obvious, because classical statisticians do not stress any particular philosophy in their science, unlike (some) Bayesian statisticians. The assertion that the classical statistics is associated with the frequency theory can be derived by the method of elimination: this branch of statistics is not based on the classical philosophy of probability because that philosophy covers only a handful of highly symmetric models; it is not based on the logical or propensity theories because these are hardly known to statisticians and mathematicians; and it is not based on the subjective theory because that theory is the basis of the other major branch of statistics—the Bayesian statistics. This leaves the frequency theory as the only candidate for the philosophical foundations of the classical statistics. One could argue, though, that the classical statistics is not associated with any philosophy of probability whatsoever. I do not take this view here—I believe that the classical statistics adopted implicitly the frequency philosophy as its justification for the use of the expected value in various contexts. I will argue that, in fact, the classical statistics is based on (L1)-(L5).

8.1. Classical models.

None of the main aspects of the classical statistics is based on the frequency theory of probability, except for a few trivial examples. I will start my analysis with the models of classical statistics.

A typical classical statistical analysis starts with a model with some unknown parameters. For example, if you have a deformed coin, the results of its tosses will be represented by an i.i.d. (independent identically distributed) sequence of heads and tails. The probability of heads on a given toss is assumed to be an unknown constant (parameter) θ and the goal of the analysis is to find a good estimate of the true value of θ. If n tosses were performed and k of them resulted in heads, one can take k/n as an estimate of θ. This estimator (that is, a function generating an estimate from the data) is unbiased in the sense that the expected value of the estimate is the true value of θ.

The number of tosses n in the above example or similar models can be large in some practical situations and then the theory of collectives is applicable, at least in a vague sense. However, classical statisticians do not insist that the estimators similar to the one presented above must be applied only to very large data sets. The accuracy of the estimator k/n depends on the number of trials n, and, of course, the larger n, the better

the accuracy. This is never taken to mean that the estimator cannot be used for small n. It is left to the end user of statistical methods to decide whether the estimator is useful for any particular value of n in a given situation.

Classical statisticians do not hesitate to analyze models which are far from von Mises' collectives, for example, stationary processes. The defining property of a stationary process is that its distribution does not depend on the choice of the origin of time, in an appropriate sense. Many important stationary processes are characterized by strong dependence between their values at different times and so they are far from collectives. Some subclasses of stationary processes can be parameterized, that is, distinct members of a family of stationary processes may be labelled by real numbers θ, for example. There is nothing that would prevent classical statisticians from estimating parameter θ of a process in this family. In general, there is no way of representing a stationary process as a collective in a natural way. Many other statistical models used by classical statisticians are hard or impossible to represent as collectives.

8.2. Interpretation of statistical analysis results.

The interpretation of the results of the classical statistical analysis is not based on the frequency theory either, in most practical applications. Some of the typical results of the classical statistical analysis are an unbiased estimate of an unknown parameter, a decision to reject or accept a hypothesis, or a confidence interval. An estimator is called unbiased if its expectation is equal to the value of the estimated (unknown) parameter. The long run interpretation of this statement would necessarily involve a sequence of identical and independent data sets, and the corresponding sequence of estimates. Usually, statisticians work with only one data set at a time and they make no assumption that identical data sets will be collected in the future. Even if multiple data sets related to the same physical phenomenon are collected, they often differ by their size—this alone disqualifies a sequence of data sets from being a collective in the sense of von Mises. If one and the same statistician collects a large number of isomorphic data sets to estimate the same unknown parameter, it is natural for her to combine all the data sets into one large data set and generate a single accurate estimate of the parameter. There might be some practical situations when she has to derive repeatedly a large number of isomorphic estimates of the same unknown constant, each one based on a different data set. Although one cannot decisively rule out such a situation, it is clearly far from common.

The results of hypothesis testing and confidence intervals are expressed in terms of probabilities and they defy the long run interpretation for the same reason as estimators. I

conclude that the long run interpretation of probability and expected value is not applied to the results of the classical statistical analysis.

8.3. Does classical statistics need the frequency theory?

Why do classical statisticians need the frequency philosophy of probability, if they need it at all? They seem to need it, for two reasons. First, some of the most elementary techniques of the classical statistics agree very well with the theory of collectives. If you have a deformed coin with an unknown probability of heads, you may toss it a large number of times and record the relative frequency of heads. This relative frequency represents probability in the von Mises theory, and it is perhaps the most popular unbiased estimator in the whole classical statistics. Moreover, a slight variation of the procedure is the most used algorithm for finding values of scientific constants. In that context, the averaging procedure is supposed to limit the influence of the measurement errors on the results of experiments. From the perspective of a professional statistician, the estimation procedure described here is only a "baby" example for undergraduates. From the point of view of some scientists, the long run frequency approach is the essence of probability because it is used extremely often.

Classical statisticians use the frequency theory in an implicit way to justify the use of expectation in their analysis. It is generally recognized that an estimator is good if it is "unbiased," that is, if its expected value is equal to the true value of the unknown parameter. People subconsciously like the idea that the expected value of the estimator is equal to the true value of the parameter even if they know that the mathematical "expected value" is not expected at all in most cases. An implicit philosophical justification for unbiased estimators is that the expected value is the long run average, so even if our estimator is not quite equal to the true value of the unknown parameter, at least this is true "on average." The problem here, swept under the rug, is that, in a typical case, there is no long run average to talk about, that is, the process of estimation of the same unknown constant is not repeated on a long sequence of isomorphic data sets.

8.4. Hypotheses testing and (L5).

Readers familiar with statistical methods must have noticed a similarity between my interpretation of (L5) and testing of statistical hypotheses, a popular method among classical statisticians. Despite unquestionable similarities, there are some subtle philosophical differences—I will outline them in this section.

Testing a statistical hypothesis often involves a parametric model, that is, some prob-

ability relations are taken for granted, such as exchangeability of the data, and some parameters, such as the expected value of a single observation, are considered unknown. The hypothesis to be tested usually refers to the parameter, whose value is considered "unknown." In my interpretation of (L5), the failure of a prediction would invalidate some assumptions adopted on the basis of (L2)-(L4), that is the model, which under normal circumstances is considered "known."

The standard mathematical model for hypothesis testing involves not only the "null hypothesis," which is often slated for rejection, but also an alternative hypothesis. When a prediction made on the basis of (L1)-(L5) fails, and so one has to reject the model built using (L2)-(L4), there is no alternative model lurking in the background. This is in agreement with general scientific practices—a failed scientific model is not always immediately replaced with an alternative model; some phenomena lack good scientific models, at least temporarily.

Traditionally, the significance level used in hypotheses testing is taken to be 5% or 1%. Roughly speaking, people are willing to tolerate error probabilities as high as 5%. Anyone wishing to apply (L5) has to choose a number playing a similar role. A prediction, in the sense of (L5), is an event whose probability is very close to 1. My personal preference is to limit predictions to events whose probabilities are much closer to 1 than 5% or even 1%. This choice limits the domain of applicability of the probability theory but makes it very reliable. In general, hypotheses testing is not meant to achieve extremely high levels of reliability but to offer a good practical method for making decisions, especially in the context of a typical scientific practice that involves testing large numbers of various hypotheses over long periods of time.

One can wonder whether my interpretation of (L5) can be formalized using the concept of hypotheses testing. It might, but doing so would inevitably lead to a vicious circle of ideas, on the philosophical side. Hypotheses testing needs a scientific interpretation of probability and so it must be based on (L1)-(L5) or a similar system. My observations of scientific practice indicate that (L5) is applied without any attempt at formalization, and with great success. In any science, the basic building blocks have to remain at an informal level, or otherwise one would have to deal with an infinite regress of ideas. Law (L5) does not need any further formalization, in my opinion.

8.5. Classical statistics and (L1)-(L5).

I am quite comfortable with most of the methods of classical statistics because they are based on (L1)-(L5), except that I do not like the universal use of the expected value as

the basis for presenting and evaluating the results of the analysis. My own inclination is to adopt in statistics the philosophy of decision making presented in Section 6.1.4, based on the Large Deviation Principle, because it fits best with (L1)-(L5), especially with (L5). Hence, I would be happy to see the theory of statistics reworked so that it is based on this decision making philosophy. Needless to say, there is little chance for this happening just because of the publication of this book. Likewise, I do not expect the classical statisticians to rush and reshape their theory so that it matches my second choice for the decision making philosophy, the one presented in Section 6.1.3, based on the idea of "stochastic ordering." My guess is that the expected value will retain its central position in statistical thinking, because of its unquestionable technical convenience. I have argued in Sections 4.1 and 6.1.2 that decision making is not a part of science and so one can adopt the maximization of the expected gain as an axiom. I have also argued that this is a dubious intellectual choice, based mostly on a linguistic trick. My hope is that it will be recognized at some point that the continued use of the expected value does not have a solid scientific or philosophical justification; a partial justification can be based on ideas presented in Sections 6.1.3 and 6.1.4.

9. SUBJECTIVE THEORY OF PROBABILITY

"Motion does not exist" Zeno of Elea (c. 450 B.C.)

"Probability does not exist" Bruno de Finetti (c. 1950 A.D.)

Alan Sokal published a paper "Transgressing the Boundaries: Toward a Transformative Hermeneutics of Quantum Gravity" in *Social Text* in 1996 ([6]). The paper turned out to be a parody of modern pseudo-science, to the embarrassment of the editors of the journal. Sokal has this to say, among other things, in a follow-up paper [7]:

> *Rather, my concern is explicitly political: to combat a currently fashionable postmodernist/poststructuralist/social-constructivist discourse—and more generally a penchant for subjectivism—which is, I believe, inimical to the values and future of the Left.*

I do not want to stretch things too far but it is hard to miss the central position of the word "subjectivism" in the quote. In view of Sokal's hoax, it is not entirely unthinkable that the subjective theory of probability was also a hoax, the greatest hoax of the twentieth century. Since this hoax has not been exposed yet, I will try to do it myself. The main claims of my argument will be the following.

(i) The subjective theory of probability has no scientific content. It is a purely philosophical theory and as such, it should not be applied as science.

(ii) If the subjective theory of probability is treated as science, it is demonstrably false.

(iii) The subjective theory and its standard scientific implementation involving the Bayes theorem are logically inconsistent, that is, the combination of the two contains a contradiction.

9.1. Subjective theory of probability is not science.

It is clear from the writings of de Finetti and all Bayesian practice that the subjective theory of probability is meant to be used as science. Hence, the subjectivists cannot claim that their theory is not science because it was never intended to be a science.

The theory fails the basic test for a scientific theory—it has no scientific content because it does not report on any facts or patterns observed in the past. I will illustrate this claim with some examples from physics and probability.

Facts and patterns can be classified according to their generality. Consider the following facts and patterns.

(A1) John Brown cut a branch of a tree on May 17, 1963, and noticed that the saw was very warm when he finished the task.

(A2) Whenever a saw is used to cut wood, its temperature increases.

(A3) Friction generates heat.

(A4) Mechanical energy can be transformed into heat energy.

(A5) Energy is always preserved.

I might have skipped a few levels of generality but the reader should get the picture. Here are probabilistic counterparts of these facts and patterns.

(B1) John Brown flipped a coin on May 17, 1963. It fell heads up.

(B2) About 50% of coin flips in America in 1963 resulted in heads.

(B3) Symmetries in an experiment such as coin tossing or in a piece of equipment such as a lottery machine are usually reflected by symmetries in the relative frequencies of events.

(B4) Probabilities of symmetric events, such as these in (B3) and in time-exchangeable sequences, are identical.

I consider the omission of (B4) from the subjective theory to be its fatal flaw that destroys its claim to be a scientific theory. I will examine possible excuses for the omission.

Some sciences, such as paleontology, report individual facts at the same level of generality as (A1) or (B1) but I have to admit that the theory of probability cannot do that. One of the reasons is that the number of individual facts relevant to probability is so large that they cannot be reported in any usable way, and even if we could find such a way, the current technology does not provide tools to analyze all the data ever collected by the humanity.

The omission of (B2) by the subjective theory is harder to understand but it can be explained. It is obvious that most people consider this type of information useful and relevant. However, truly scientific examples at the level of generality of (B2) would not deal with coin tosses but with repeated measurements of scientific constants, for example. It is a legitimate claim that observed patterns at this level of generality belong to various fields of science such as chemistry, biology, physics, etc. They are in fact reported by scientists working in these fields and so there is no need to incorporate them into the theory of probability. One could even say that such patterns do not belong to the probability theory because they belong to some other sciences.

Finally, we come to (B3) and (B4). Clearly, these patterns do not belong to any science such as biology or chemistry. If the science of probability does not report these patterns, who will?

If you roll a die, the probability that the number of dots is less than 3 is 1/3; this is a concise summary of some observed patterns. Every theory of probability reported this finding in some way, except for the subjective theory. Needless to say, de Finetti did not omit such statements from his theory because he was not aware of them—the omission was a conscious choice. De Finetti's choice can be easily explained. If he reported any probabilistic patterns, such as the apparent stability of relative frequencies in long runs of experiments, his account would take the form of "scientific laws." Scientific laws need to be verified (or need to be falsifiable, in Popper's version of the same idea). Stating any scientific laws of probability would have completely destroyed de Finetti's philosophical theory. The undeniable strength of his theory is that it avoids in a very simple way the thorny question of verifiability of probabilistic statements—it denies that there are any, in the objective sense. The same feature that is a philosophical strength, is a scientific weakness. No matter how attractive the subjective theory may appear to the philosophically minded people, it has nothing to offer on the scientific side.

9.2. Subjective science of probability is false.

Although the subjective theory of probability has no scientific content, it is used as a scientific theory. I will show that if we treat it this way, it is demonstrably incorrect.

9.2.1. Creating something out of nothing.

One of the most extraordinary claims ever made in science and philosophy is that consistency alone is the sufficient basis for a science, specifically, for the science of Bayesian statistics. I feel that people who support this claim lack imagination. I will try to help them by presenting an example of what may happen when consistency is indeed taken as the only basis for making probability assignments.

Dyslexia is a mild disability which makes people misinterpret written words, for example, by rearranging their letters, as in "tow" and "two." Let us consider the case of Mr. P. Di Es, an individual suffering from a probabilistic counterpart of dyslexia, a "Probabilistic Dysfunctionality Syndrome." Mr. P. Di Es cannot recognize events which are disjoint, physically unrelated or invariant under symmetries, and the last two categories are especially challenging for him. Hence, Mr. P. Di Es cannot apply (L1)-(L5) to make decisions. Here are some examples of Mr. P. Di Es' perceptions. He thinks that the event that a bird

comes to the bird feeder in his yard tomorrow is not physically unrelated to the event that a new war breaks out in Africa next year. At the same time, Mr. P. Di Es does not see any relationship between a cloudy sky in the morning and rain in the afternoon. Similarly, Mr. P. Di Es has problems with sorting out which sequences are exchangeable. When he reads a newspaper, he thinks that all digits printed in a given issue form an exchangeable sequence, including those in the weather section and stocks analysis. Mr. P. Di Es buys bread at a local bakery and is shortchanged by a dishonest baker about 50% of the time. He is unhappy every time he discovers that he was cheated but he does not realize that the sequence of bread purchases in the same bakery can be considered exchangeable and so he goes to the bakery with the same trusting attitude every day.

Some mental disabilities are almost miraculously compensated in some other extraordinary way, for example, some autistic children have exceptional artistic talents. Mr. P. Di Es is similarly talented in a very special way—he is absolutely consistent in his opinions, in the sense of de Finetti.

Needless to say, a person impaired as severely as Mr. P. Di Es is as helpless as a baby. The ability of Mr. P. Di Es to assign probabilities to events in a consistent way has no discernible effect on his life.

The example is clearly artificial—there are very few, if any, people with this particular combination of disabilities and abilities. This is probably the reason why so many people do not notice that consistency alone is totally useless. Consistency is never applied without (L1)-(L5) in real life. It is amazing that the subjective theory, and implicitly the consistency idea, claim all the credit for the unquestionable achievements of the Bayesian statistics.

9.2.2. Searching for the essence of probability.

I will formalize the example given in the last section. First, it will be convenient to talk about "agents" rather than people. An agent may be a person or a computer program. It might be easier to imagine an imperfect or faulty computer program, rather than a human being, acting just as Mr. P. Di Es does.

Consider four agents, applying different strategies in face of uncertainty.

(A1) Agent A1 assigns probabilities to events without using the mathematics of probability, without using consistency and without using (L1)-(L5). He does not use any other guiding principle in his choices of probability values.

(A2) Agent A2 is consistent but does not use (L1)-(L5). In other words, he acts as Mr. P. Di Es does.

(A3) Agent A3 uses (L1)-(L5) in his probability assignments but does not use the mathematical rules for manipulating probability values.

(A4) Agent A4 applies both (L1)-(L5) and the mathematical theory of probability (in particular, he is "consistent").

Let me make a digression. I guess that agent A3 is a good representation for a sizeable proportion of the human population. I believe that (L1)-(L5) are at least partly instinctive and so they are common to most people but the mathematical rules of probability are not easy to apply at the instinctive level and they are mostly inaccessible to people lacking education. Whether my guess is correct is inessential since I will focus on agents A1, A2 and A4.

Before I compare the four agents, I want to make a comment on the interpretation of the laws of science. Every law contains an implicit assertion that the elements of reality not mentioned explicitly in the law do not matter. Consider the following example. One of the Newton's laws of motion says that the acceleration of a body is proportional to the force acting on the body but inversely proportional to the mass of the body. An implicit message is that if the body is green and we paint it red, that will not change the acceleration of the body. (This interpretation is not universally accepted—some young people buy red cars and replace ordinary mufflers with noise-making mufflers in the hope that red color and noise will improve the acceleration of the car.)

It is quite clear that agents A1 and A4 lie at the two ends of spectrum when it comes to the success in ordinary life, but even more so in science. Where should we place agent A2? I have no doubt that A2 would have no more than 1% greater success than A1. In other words, consistency can account for less than 1% of the overall success of probability theory. I guess that A3 would be about half-way between A1 and A4, but such a speculation is not needed for my arguments.

Now I am ready to argue that the subjective theory of probability is false, if we consider it a scientific theory. The theory claims that probability is subjective, there is no objective probability, and you have to be consistent. An implicit message is that if you assign equal probabilities to symmetric events, as in (L4), you will not gain anything, just like you cannot increase the acceleration of a body by painting it red. Similarly, the subjective theory claims that using (L3) cannot improve your performance. In other words, the subjective theory asserts that agent A2 will do in life as well as agent A4. I consider this assertion absurd.

De Finetti failed in a spectacular way by formalizing only this part of the probabilistic

experience which explains less than 1% of the success of probability—he formalized only the consistency, that is, the necessity of applying the mathematical rules of probability.

I do not see any way in which the subjective *science* of probability can be repaired. It faces the following alternative: either it insists that (L1)-(L5) can give no extra advantage to people who are consistent, and thus makes itself ridiculous by advocating Mr. P. Di Es-style behavior; or it admits that (L1)-(L5) indeed provide an extra advantage, but then it collapses into ashes. If (L1)-(L5) provide an extra advantage, it means that there exists a link between the real universe and good probability assignments, so the subjective philosophy is false.

9.3. Inconsistent theory of consistency.

It is not unfair to call the subjective theory a "consistency" theory because this concept is one of its fundamental ideas. Another central idea is the updating of the subjective probability views after collecting new data. The Bayes theorem provides the technical tool for doing that. I will show that the Dutch book argument is inconsistent with the standard applications of the Bayes theorem. This is a striking failure for a theory that put consistency on its banners. My argument is based on an idea of Ryder [5] (see also Gillies [2]).

The standard Bayesian procedure can be described as follows. Start with a prior distribution describing your subjective opinions about unknown events. Then collect the data. Finally, derive the posterior distribution from the prior distribution and the data using the Bayes theorem. Why should one derive the posterior distribution from the prior distribution? Why isn't it a good idea to abandon the prior distribution and come up with a completely new posterior distribution, not necessarily related in any way to the prior one? Intellectual consistency seems to be a good reason for coordinating prior and posterior views but it is far from clear whether consistency can have any desirable results because the subjective theory denies the existence of objective probabilities. Hence, the Dutch book argument has to be invoked in support of using the Bayes theorem to coordinate prior and posterior views. In all non-trivial situations, the posterior distribution is different from the prior distribution. Hence, for some future event A, the posterior distribution assigns a different value to the probability of A than the prior distribution does. If the decision maker has a chance to act on both believes about A, prior and posterior, in presence of appropriate betting opportunities, a Dutch book will be formed against him. It follows that Bayesian statisticians must not change their priors at all when the data arrive. Note that if one does not modify the prior distribution using the Bayes theorem but keeps it

unchanged after collecting the data, this will result in a consistent betting strategy, that is, the bets made before and after collecting of the data will not form a Dutch book against the statistician.

The abstract argument given above can be illustrated by the following simple example. Suppose that Susan is shown two urns, the first one with two white balls and one black ball, and the other with two black balls and one white ball. Someone tosses a coin without showing it to Susan and notes the result on a piece of paper. It is natural for Susan to assume that the result of the coin toss is heads with probability 1/2. Susan is offered and accepts the following bet (Bet I) on the result of the coin toss. She will collect \$8 if the result is tails; otherwise she will lose \$7. Then someone looks at the result of the coin toss and samples a ball from the first urn if the result is heads; he samples a ball from the other urn otherwise. Suppose Susan is shown a white ball but she is not told which urn the ball came from. The Bayes theorem implies that the posterior probability of heads is 2/3. Susan is now offered and accepts a bet (Bet II) that pays \$6 if the result of the coin toss is heads; otherwise she loses \$9. A Dutch book has been formed against Susan because she accepted both bets—no matter what the result of the coin toss was, she will lose \$1 once the result is revealed. A simple way for Susan to avoid a Dutch book would have been to take 1/2 as the probability of heads, before and after observing the color of the sampled ball.

I will now return to the abstract level of argument to give another, this time graphical, presentation of the argument. The crucial elements of the example are ordered in time as follows.

$$\text{Bet I} \longrightarrow \text{Arrival of data} \longrightarrow \text{Bet II} \longrightarrow \text{Dutch book (?)}$$

$$\longrightarrow \text{Observation of the event and payoff of bets}$$

Bets I and II are offered before the relevant event (coin toss result) can be observed—otherwise they would not be bets. The sole purpose of the probability theory, according to the subjectivists, is to coordinate bets so that no Dutch book is ever formed, that is, the bets are consistent. The subjective theory claims that this goal is indeed attainable. The theory goes on to say that after the arrival of data, the Bayes theorem must be used to coordinate Bets I and II to achieve the goal of consistency, that is, to avoid a Dutch book. In my example, a Dutch book is formed if the Bayes theorem is applied. This proves that an application of the Bayes theorem yields a contradiction in the subjective theory.

Here is an "explanation" of the logical error made by the subjectivists. The Dutch book argument in support of consistency (and, therefore, in support of using probability to express one's opinions) is static in nature—the probabilities are assumed to be immutable. Statistics is a naturally dynamic science—the assessment of probabilities keeps changing as the data accumulate.

If this had been a mathematical book, it would have been only three pages long. A single inconsistency, such as the one presented in this section, would be enough to annihilate any theory in mathematics; there would be no need for any other form of the critique. I am far from saying that philosophy tolerates contradictions. However, a philosophical idea may be formalized in many ways and one could imagine a version of the subjective theory avoiding the inconsistency pointed out in this section. I do not see a way out of this quagmire, though. The simple example given above shows that a Dutch book can be formed against a Bayesian statistician and nothing in this argument depends on the assumptions of the subjective probability. Moreover, the Bayesian statistician could avoid the danger of the Dutch book in a simple way, by not changing her prior distribution. I do not see how one can justify the "irrational" behavior of changing the prior distribution to the posterior distribution, and so exposing oneself to the Dutch book, without resorting to some "objective" argument. If we assume that there is no objective probability, why is it rational or beneficial for Susan to accept both bets?

9.4. Science, probability and subjectivism.

De Finetti has this to say about the fact that beliefs in some probability statements are common to all scientists (quote after [2], page 70):

> *Our point of view remains in all cases the same: to show that there are rather profound psychological reasons which make the exact or approximate agreement that is observed between the opinions of different individuals very natural, but there are no reasons, rational, positive, or metaphysical, that can give this fact any meaning beyond that of a simple agreement of subjective opinions.*

This is a purely philosophical statement. The fact that all scientists agree on the probability that the spin of an electron will be positive under some experimental conditions is not subjective or objective—this agreement is the essence of science. The question of whether this agreement has any objective meaning can be safely left to philosophers because it has no effect on science. No branch of "deterministic" science has anything to offer besides the "simple agreement of subjective opinions" of scientists. Nobody knows the objective truth,

unless he or she has a direct line to God—even Newton's physics proved to be wrong, or at least inaccurate. The agreement of probabilistic opinions held by various scientists is as valuable in practice as their agreement on deterministic facts and patterns.

De Finetti correctly noticed (just like everybody else) that the evidence in support of probabilistic laws, such as (L1)-(L5), is less convincing than that in support of deterministic laws (but I would argue that this is true only in the purely philosophical sense). Hence, the users of probability have the right to treat the laws of probability with greater caution than the laws of the deterministic science. However, I see no evidence that they exercise this right; laws (L1)-(L5) are slavishly followed by even most avowed supporters of the subjectivist viewpoint.

De Finetti did not distinguish between the account of the accumulated knowledge and the application of the same knowledge. Science has to summarize the available information the best it can, so the science of probability must consist of some laws such as (L1)-(L5). The same science of probability must honestly explain how the laws were arrived at. A user of probability may choose to consider all probabilistic statements subjective, as proposed by de Finetti, but there is nothing peculiar about the probability theory here—the quantum physics and even Newton's laws of motion can be considered subjective as well, because one cannot prove beyond any doubt that any given law is objectively true, as long as its formulation comes from humans.

9.5. A word with a thousand meanings.

One of the reasons why the subjective theory of probability is so successful is because the word "subjective" has numerous meanings and every objection to the subjective theory or to the Bayesian statistics can be countered by appealing to that meaning of "subjectivity" that fits the current argument. I will review some of the meanings of the word "subjective" in the hope that this will help the discussions surrounding the subjective theory—one cannot expect a substantial convergence of opposing philosophical views if their holders use the same word in different ways.

Dictionaries contain long lists of different meanings of the word "subjective" but many of those meanings are not relevant to our discussion, and vice versa, some meanings used in the specialized probabilistic context cannot be found in the dictionaries.

I start my review by repeating verbatim four possible interpretations of the statement that "probability is subjective" and their discussion from Section 3.5.

(i) "Although most people think that coin tosses and similar long run experiments displayed some patterns in the past, scientists determined that those patterns were fig-

ments of imagination, just like optical illusions."

(ii) "Coin tosses and similar long run experiments displayed some patterns in the past but those patterns are irrelevant for the prediction of any future event."

(iii) "The results of coin tosses will follow the pattern I choose, that is, if I think that the probability of heads is 0.7 then I will observe roughly 70% of heads in a long run of coin tosses."

(iv) "Opinions about coin tosses vary widely among people."

Each one of the above interpretations is false in the sense that it is not what de Finetti said or what he was trying to say. The first interpretation involves "patterns" that can be understood in both objective and subjective sense. De Finetti never questioned the fact that some people noticed some (subjective) patterns in the past random experiments. De Finetti argued that people should be "consistent" in their probability assignments and that recommendation never included a suggestion that the (subjective) patterns observed in the past should be ignored in making one's own subjective predictions of the future, so (ii) is not a correct interpretation of de Finetti's ideas either. Clearly, de Finetti never claimed that one can affect future events just by thinking about them, as suggested by (iii). We know that de Finetti was aware of the clustering of people's opinions about some events, especially those in science, because he addressed this issue in his writings, so again (iv) is a false interpretation of the basic tenets of the subjective theory.

(v) I continue the review with the meaning that was given to the word "subjective" by de Finetti. According to him, a probability statement cannot be proved or disproved, verified or falsified. This suggests that "probability" does not refer to anything objective in the real universe. From the philosophical point of view, it is not obvious at all that if some statement cannot be verified or falsified in any way, then it has to be subjective. People who believe in God do not consider God's attributes subjective just because one cannot prove, in any ordinary sense of the word, any statement about those attributes. The current view of physics is that we will never be able to verify, in any sense accepted by physics, any statement about the interior of a black hole. I do not think that anyone interprets this as saying that the processes in the interior of the black hole are subjective—I think that a better way to describe the current scientific views is to say that those processes are "objective but unknowable." All these philosophical subtleties are rather irrelevant when it comes to scientific practice. If one accepts the claim that probability statements are not verifiable then it will make no harm to think about probability as subjective. A probability that is objective but unknowable is not any more useful than a subjective probability.

(vi) The next meaning on my list is related to the fact that different people have different information and, as is recognized in different ways by all theories, the probability of an event depends on the information possessed by the probability assessor. One can deduce from this that probability is necessarily subjective, because one cannot imagine a realistic situation in which two people have identical knowledge. This interpretation of the word "subjective" contradicts in a fundamental way the spirit of the subjective theory. The main idea of the subjective theory is that two rational people with access to the same information can differ in their assessment of probabilities. If the differences in the probability assessments were attributable to the differences in the knowledge, one could try to reconcile the differences by exchanging the information. No such possibility is suggested by the subjective theory of probability, because that would imply that probabilities are a unique function of the information, and in this sense they are objective. De Finetti was not trying to say that the impossibility of the perfect communication between people is the only obstacle preventing us from finding objective probabilities.

(vii) In order to implement (L1)-(L5) in practice, one has to recognize events that are disjoint, independent or symmetric. This may be hard for several reasons. One of them is that no pair of events is perfectly symmetric, just like no real wheel is a perfect circle. Hence, one has to use a "subjective" judgment to decide whether any particular pair of events is symmetric or not. Even if we assume that some events are perfectly symmetric, the imperfect nature of our observations makes it impossible to discern such events and, therefore, any attempt at application of (L1)-(L5) must be subjective in nature. This interpretation of the word subjective is as far from de Finetti's definition as the previous interpretation. In de Finetti's theory, real world symmetries are totally irrelevant when it comes to the assignment of probabilities. In his theory, probability is subjective in the sense that numbers representing probabilities are not linked in any way to the observable world. Probability values chosen using symmetries are not verifiable, just like any other probability values, so symmetry considerations have no role to play in the subjective theory.

(viii) "Subjective" opinion can mean "arbitrary" or "chaotic" in the sense that nobody, including the holder of the opinion, can give any rational explanation or guiding principle for the choice of the opinion. This meaning of subjectivity is likely to be attributed to subjectivists by their critics. In some sense, this critical attitude is justified by the subjective theory—as long as the theory does not explicitly specify how to choose a consistent opinion about the world, you never know what a given person might do. I do not think that de Finetti understood subjectivity in this way. It seems to me that he believed

that an individual may have a clear, well organized view of the world. De Finetti argued that it is a good idea to make your views consistent, but he also argued that nothing can validate any specific set of such views in a scientific way.

(ix) "Subjective" can mean "objectively true" or "objectively valuable" but "varying from person to person." For example, my appreciation of Thai food is subjective because not all people share the same taste in food. However, my culinary preferences are objective in another sense. Although my inner feeling of satisfaction when I leave a Thai restaurant is not directly accessible to any other person, an observer could record my facial expressions, verbal utterances and restaurant choices to confirm in quite an objective way that Thai food indeed gives me pleasure and is among my favorite choices. There is no evidence that this interpretation of the word "subjective" has anything to do with de Finetti's theory. In many situations, such as scientific research, the consequences of various decisions are directly observable by all interested people and there is a universal agreement on their significance. In such cases, a result of a decision cannot be "good" or "true" for one person but not for some other person.

9.6. Apples and oranges.

Can you mix objective and subjective probabilities in one theory? The reader might have noticed that many of my arguments were based on the same categorical assumption that was made by de Finetti, that no objective probability can exist whatsoever. It may seem unfair to find a weak point in a theory and to exploit it to the limit. One could expect that if this one hole is patched in some way, the rest of the theory might appear quite reasonable. Although I disagree with de Finetti on almost everything, I totally agree with him on one point—it is not possible to mix objective and subjective probabilities in a single philosophical theory. This was not a fanatical position of de Finetti, but a profound understanding of the philosophical problems that would be faced by anyone trying to create a hybrid theory. I will explain what these problems are in just a moment. First, let me say that the idea that some probabilities are subjective and some are objective goes back at least to Ramsey, the other co-inventor of the subjective theory, in 1920's. Carnap, the most prominent representative of the logical theory of probability, talked about two kinds of probability in his theory. And even now, Gillies [2] advocates a dual approach to probability. Other people made similar suggestions but all this remains in the realm of pure heuristics.

If you assume that both objective and subjective probabilities exist, your theory will be a Frankenstein monster uniting all the philosophical problems of both theories and creating

some problems of its own. You will have to answer the following questions, among others.

(i) If some probabilities are objective, how do you verify objective probability statements? Since the subjective probability statements cannot be objectively verified, do they have the same value as the objective statements or are they inferior? If the two kinds of probability are equally valuable, why bother to verify objective probability statements if one can use the subjective probabilities? If the two kinds of probability are not equally valuable, how do you define and measure the degree of inferiority of subjective statements?

(ii) If you multiply an objective probability by a subjective probability, is the result objective or subjective? The same question applies to the result of an addition of two probabilities of different kinds.

(iii) Are all probabilities totally objective or totally subjective, or can a probability be, say, 70% subjective? If so, question (ii) has to be modified: If you multiply a 30% objective probability by a 60% objective probability, to what degree is the result objective? How do you measure the degree to which a probability is objective?

There is no point in continuing the list—I am not aware of any theory that would be able to give even remotely convincing answers to (i)-(iii).

9.7. Imagination and probability.

A common criticism of the frequency theory coming from the subjectivist camp is that the frequency theory applies only to long sequences of i.i.d. events; in other words, it does not apply to individual events. Ironically, subjectivists fail to notice that a very similar criticism applies to their theory, because the subjective theory is meaningful only if one has to make at least two distinct decisions—the Dutch book argument is vacuous otherwise. Both theories have problems explaining the common practice of assigning a probability to a unique event in the context of a single decision. Actually, the subjective theory is meaningless even in the context of a complex situation involving many events, as long as only one decision is to be made. Typically, for any given event, its probability can be any number in the interval from 0 to 1, for some consistent set of opinions about all future events, that is, for some probability distribution. If a single decision is to be made and it depends on the assessment of the probability of such an event, the subjective theory has no advice to offer. There are spheres of human activity, such as business, investment and warfare, where multiple decisions have to be coordinated for the optimal result. However, there are plenty of probabilistic situations, both in everyday life and scientific practice when

only isolated decisions are made. For example, a task performed often by scientists is to measure or estimate some quantity on the basis of repeated experiments or observations. The scientist often reports a single number (the estimate) in a scientific journal, and she may also include some information about the error (the standard deviation). The whole process involves only one decision—the publication of the findings in a journal. The results may be later found to be false in some sense (objective, for the believers in the objective probability; subjectivists may find the results extremely different from their own opinions). However, even if the scientist's published results prove to be false in either sense, no Dutch book is ever formed. In cases like this, the subjective theory does not provide any guidelines for distinguishing bad science from the good science.

Here is the summary of the above argument. One of the claims of the subjective theory is that it can deal with individual events, unlike the frequency theory. All that the subjective theory can say about an individual event is that its probability is between 0 and 1—a totally useless piece of advice.

The subjective theory of probability suffers from the dependence on the imagined entities, just like the frequency theory. In the case of the frequency theory, one has to imagine non-existent collectives; in the case of the subjective theory of probability, one has to imagine non-existent collections of decisions (subjective probabilities are just a way of encoding consistent choices between various imaginary decisions). The philosophical and practical problems arising here are very similar in both theories. On the philosophical side, it is hard to see why the imagined sequences of experiments or collections of decisions are necessary to apply the probability theory to real life events. Why no other scientific theory insists that people use their imagination? On the practical side, imagined entities differ from person to person, so a science based on imagination cannot generate reliable advice.

9.8. An enemy within.

This section is devoted to some purely philosophical remarks on the objective elements in the subjective theory. I will make a few arguments designed to show that the subjective theory is not completely devoid of objective components. This should not be confused with the arguments presented in the next chapter, where I show that Bayesian statisticians assign probabilities in a way that can be only called objective, that is, in agreement with (L1)-(L5).

My first argument will analyze the possibility of changing one's consistent set of probabilistic opinions, given some external incentive. First, consider the following deterministic example. Recall that one of the Newton's laws of motion states that the acceleration is

proportional to the force but inversely proportional to the mass of an object. Suppose that you want to buy a car and the only feature that you care about is its acceleration. You are negotiating a deal with a car dealer who has two cars of the same model, one red and one blue. According to Newton's laws, the color does not make any difference when it comes to the acceleration. Suppose you are indifferent between the two colors and assume further that the dealer is willing to sell the blue car for $500.00 less than the red car, because red cars are popular with other buyers. The only rational decision for you is to buy the blue car and save $500.00.

Now suppose that you have to make a decision in a situation involving uncertainty. According to the subjective theory, you have to use a consistent set of opinions about unknown events, that is, you have to come up with a probability distribution encoding your preferences. The subjective theory maintains that there is no objective probability and so one cannot compare different consistent views of the world—none of them is more "true" than any other, and no amount of information collected in the future can show that any consistent view is "better" than any other view. This leaves an enormous freedom to the decision maker, assuming that the current decision problem is unrelated to the decisions made in the past, and so there is no danger of falling into the Dutch book trap.

A decision maker may be offered an incentive to change his consistent set of opinions, as in the case of a car buyer above. Recall that, according to the subjective theory, if you change your consistent set of opinions to another set of consistent opinions, the change cannot be shown to be wrong in any objective sense. Hence, given the slightest financial incentive, you should be willing to change your consistent set of opinions. Suppose that D_1 is the best decision assuming your current consistent set of opinions P_1. Let D_2 be the best decision assuming some other probability distribution P_2. Next, use the current distribution P_1 to calculate the expected gains corresponding to D_1 and D_2. The difference between the two expectations represents your current estimate of the loss incurred due to switching from P_1 to P_2. If the difference is greater than the external incentive offered to switch from P_1 to P_2 then you should reject the incentive and keep the distribution P_1. However, you will never be able to see that retaining P_1 was the right decision, according to the subjective theory. If you accept the incentive, you will be able to enjoy the undeniable advantage of having the incentive. No outcome of currently unknown events will be able prove that the decision to change the consistent set of opinions from P_1 to P_2 was wrong or false or suboptimal in any sense, according to the subjective theory. We have arrived at a contradiction—the subjective theory implies that you should and should not accept the

external incentive to switch from P_1 to P_2. It is clear from the general tone of examples given by subjectivists that one is expected to use his or her "own" subjective probability, and so the subjective probability distribution is objectively valuable in some sense.

The freedom to choose any consistent set of opinions leads to an intriguing practical possibility. One could choose a set of opinions that leads to the least demanding calculations, thus saving time and expense. Recall that in the Bayesian approach to statistics, one has to calculate the posterior distribution using the Bayes theorem. This can be a very hard thing to do in some specific cases. If one can find a consistent set of opinions that would minimize such computations, this could become a good rational guiding principle for choosing one's prior distribution. No such trend is obvious in statistics. Perhaps the most optimal prior distribution is the one that makes all relevant events independent, if the situation allows for that (deformed coin tossing is a good example of such a situation). If such a prior is adopted, one can ignore the data altogether. There is no doubt that this approach to the selection of the prior distribution will never be popular. Bayesian statisticians must believe that some inconvenient prior distributions are objectively valuable, since they do not reject them for purely technical reasons.

My next argument belongs to pure philosophy and is more removed from practice than the previous ones. I claim that subjectivists have to recognize the fundamental role of symmetry (and the ability of humans to effectively recognize symmetries) to apply their theory in practice. Consider an experiment consisting of a toss of a deformed coin. A subjectivist would assign a subjective probability to the outcome "heads." It is natural to suppose that the mental description of this outcome, on any level from the verbal utterance to the neural network representation, is rather vague and does not specify the exact position of the coin. When the experiment is actually performed, the subjectivist has to identify the result as "heads" or "tails." This requires an application of symmetry, that is, the subjectivist has to identify the visual image of the coin with one of the two vague mental images considered in the past, "heads" and "tails." The procedure requires identification of those aspects that are relevant and identical in the visual and past mental images, and identification of those aspects of the image that are not relevant here. In other words, the subjectivist has to perform the same mental operations that are the basis of (L4). If this procedure is deemed to be subjective then the Bayes theorem is applied in a subjective way, that is, the passage from the prior to the posterior is arbitrary.

9.9. Free market of subjective probabilities.

I will argue that the Bayesian statistics treats consistent sets of opinions in a different way than the subjective theory does. This sends a confusing message to users of statistics.

I start with the observations similar to those in the last section. The subjective theory argues that it is a good idea to use a consistent set of opinions to make decisions but at the same time it asserts that no probability statement can be verified. A decision maker could use a consistent set of opinions that she arrived at herself, or she could use her friend's consistent set of opinions. The subjective theory states that neither set of consistent opinions can be proved to be better in an objective way.

A decision maker usually wants to base her decisions on the information about the past events and this information may vary from one person to another. Although it is obvious that the total information in possession of any person is unlike any other personal information, I would find it hard to believe that this is the case of the *relevant* information. In many practical situations in science and business, people can communicate quite effectively and the idea that they have the same or very similar information relevant to a particular decision problem is quite reasonable. There are good reasons to use a consistent set of opinions borrowed from some other person or generated by a computer. One obvious reason is efficiency—some people have more important things to do in life than to find a consistent set of opinions. Delegating decisions is common in business and politics, as long as there exists a sufficient level of trust in the capabilities and intentions of the other person. Hence, one can imagine a market for consistent views of the world, generated by specialists, and supported by buyers who are willing to pay for someone to make decisions for them. Such a market in fact exists—it goes by the name of Bayesian statistics. Bayesian statistics is not only a set of mathematical methods related to the subjective theory, but also a large collection of examples. Most users of the Bayesian statistics do not bother to find their own subjective views of the world but use generic models and priors found in Bayesian textbooks.

The identification of probabilities with decisions was originally invented to determine the "true" probabilistic beliefs of a person. It was argued that a questionnaire asking someone to list his personal probabilities would not yield results as convincing and reliable as a questionnaire based on betting preferences. This sounds perfectly reasonable but this suggests that one's own consistent view of the world is in some way more valuable than any other set of consistent set of opinions.

Standard textbooks on Bayesian statistics contain many examples of models and priors, implicitly offered to the readers for their own use. At the same time, textbooks shy

from saying explicitly that one can use the examples with the same degree of success as when using one's own subjective priors. Hence, the message is confused—the inclusion of many examples suggests that they are as good as truly personal consistent opinions but the ostensible support for the subjective theory prevents the authors from saying so.

10. BAYESIAN STATISTICS

Recall the structure of the Bayesian analysis from Section 2.3. One of the elements of the initial setup is a "prior," that is, a prior probability distribution, a consistent view of the world. The data from an experiment or observations are the second element. A Bayesian statistician then applies the Bayes theorem to derive the "posterior," that is the posterior probability distribution, a new consistent view of the world. The posterior can be used to make decisions—one has to find the expected value of the gain associated with every possible decision and make the decision that maximizes this expectation.

This simple and clear scheme conceals a rather important aspect of applied Bayesian methods. Almost always, the prior distribution is specified in two steps. First, a "model" is found. The model involves some unknown numbers that are called "parameters" by the classical statisticians. The term "prior" refers only to the unknown distribution of the "parameters." This is best explained using a coin tossing example. Consider a sequence of a deformed coin tosses. Usually, the results are represented mathematically as an exchangeable sequence. According to de Finetti's theorem, an exchangeable sequence is equivalent (mathematically) to a mixture of "i.i.d." sequences. Here, an "i.i.d. sequence" refers to a sequences of independent tosses of a coin with a fixed probability of heads. The assumption that the sequence of tosses is exchangeable is a "model." This model does not uniquely specify which i.i.d. sequences enter the mixture and with what weights. The mixing distribution (that is, the information which i.i.d. sequences are a part of the mixture, and with what weights), and only this distribution, is customarily referred to as a "prior."

10.1. Models.

Models are treated as objective representations of the reality in the Bayesian analysis. One of the common misunderstandings about the meaning of the word "subjective" comes here to play. Bayesian statisticians may differ in their opinions about a particular model that would fit a particular real life situation—in this sense, their views are subjective. For example, some of them may think that the distribution of a given random variable is symmetric, and some others may not adopt this view. This kind of subjectivity has nothing to do with de Finetti's subjectivity—according to his theory, it does not matter whether a given random variable is assumed to be symmetric or not. The symmetry in the real world, even if it is objective, is not linked in any way to probabilities, because probability values cannot be verified in any way in the subjective theory. Hence, according

to the subjective theory, differences in views between Bayesian statisticians on a particular model are totally irrelevant from the point of view of the future success of the statistical analysis—no matter what happens, nothing will prove that any particular model is right or wrong. I do not find even a shade of this attitude among the Bayesians. The importance of matching the model to the real world is taken as seriously in Bayesian statistics as in the classical statistics. Bayesian statisticians clearly think that it is a good idea to make the mathematical model symmetric if the corresponding real phenomenon is symmetric. In other words, they act as if they believed in some objective probability relations.

10.2. Priors.

One could expect that of all the elements of the Bayesian method, the prior distribution would be the most subjective one. Recall that in practice, the term "prior distribution" refers only to the opinion about the "unknown parameters," that is, that part of the model which is not determined by (L1)-(L5).

Surprisingly, Bayesian statisticians discuss the merits of different prior distributions. This strongly indicates that they do not believe in the subjectivity of priors. If a prior is subjective in the personal sense, that is, if it reflects one's own opinion, then there is nothing to discuss—the prior is what it is. Moreover, deliberations of various properties of priors suggest that priors may have some demonstrably good properties—this contradicts the spirit and the letter of the subjective theory.

Since no subjective prior can be shown to be more true than any other prior, according to the subjective theory, one could try to derive some benefit by simplifying the prior. Many priors can save money and time by reducing the computational complexity of a problem, for example, by making the future events independent from the data (one does not have to apply the Bayes theorem then). In the context of coin tossing, a very convenient prior is the one that makes the sequence of coin tosses i.i.d. with some fixed probability of heads—this subjective opinion would never require updating. Needless to say, priors are never chosen just on the basis of their technical complexity.

I have argued that priors are not expected to reflect personal opinions and they are not chosen to be computationally efficient. So how are they chosen? Let us have a look at the standard example of a deformed coin tossing. The usual choice for the prior in this case is "unform," that is, it is assumed that the sequence of tosses is a mixture of i.i.d. sequences, each i.i.d. sequence has success probability p, and p is a random variable which lies in any subinterval $[a, b]$ of $[0, 1]$ with probability $b - a$. The reason why this prior is popular is that this and similar priors are the only priors that yield posterior distributions that agree

with our intuition. If you tossed a coin n times and you want to predict the result of the next toss, it is natural to take the relative frequency of heads in the first n trials as a good estimate of the probability of heads on the $(n+1)$-st toss. The subjective philosophy does not support this choice of the estimate in any way, contrary to the widespread scientific practices and common sense.

The above remarks need a clarification because posterior distributions are also used as prior distributions. Consider again coin-tossing. After the first n observations are made, the posterior distribution based on these first n trials may be considered to be the prior distribution for the next m trials. If the original prior was uniform and n is large, this intermediate or double-role distribution is typically quite different from the uniform distribution. Since it also plays the role of a prior distribution, this seems to contradict my remarks about the almost universal use of the uniform prior. The uniform prior is used when no relevant data are available.

In Bayesian practice, the following two criteria seem to determine the choice of priors. First, priors are chosen to generate posteriors that agree with the intuition. Second, technical considerations are involved, for example, good priors give quick convergence to the desirable result in a small number of steps, if the data are collected sequentially. In this sense, the choice of priors is dictated by the end users of probability and it is not justified by the subjective theory. Priors are considered subjective because they cannot be derived using (L1)-(L5). Hence, they are subjective in the sense of being "arbitrary," not personal. However, the fact that priors are not determined by (L1)-(L5) does not mean that statisticians are not looking for some other good properties of priors. There is no fatalistic attitude among the statisticians that the "prior is what it is."

To see the true role of the prior, consider the following (rather unrealistic) situation. Suppose that a statistician faces a large number of unrelated decision problems. She chooses a prior for each of the problems and collects the data. When she is finished, the computer memory fails, she loses all the data and she has no time to collect any more data before making decisions. She has to make all the decisions on the basis of her priors. The official philosophical status of the prior is that it represent the best course of action in the absence of any additional information. In reality, nobody seems to be trying to choose priors taking into account a potential disaster described above. If anything like this ever happened, there would be no expectation that the priors chosen to fit with the whole statistical process (including data collection and analysis) would be useful in any sense in the absence of data.

10.3. Posteriors.

The posterior has the least subjective status of all elements of the Bayesian statistics, mainly because of the reality of the society. Business people, scientists, and ordinary people would have nothing to do with a theory that emphasized the subjective nature of its advice. Hence, the subjectivity of the prior may be mentioned in many circumstances but the posterior is implicitly advertised as objective. Take, for example, the title of a well known textbook on Bayesian statistics by DeGroot, "Optimal Statistical Decisions" ([1]). Optimal? According to the subjective theory of probability, your opinions can be either consistent or inconsistent, they cannot be true or false, and hence your decisions cannot be optimal or suboptimal. Of course, you may consider your own decisions optimal, but this does not say anything beyond that you have not found any inconsistency in your views—the optimality of your decisions is tautological. Decisions may be also optimal in some purely mathematical sense, but I doubt that that was the intention of DeGroot when he chose a title for his book—I have not doubt that the title was chosen, consciously or subconsciously, to suggest some objective optimality of Bayesian decisions.

From time to time, somebody expresses an opinion that the successes of the Bayesian statistics vindicate the claims of the subjective theory. The irony is that according to the subjective theory itself, nothing can confirm any probabilistic claims—the only successes that the Bayesians could claim are consistency and absence of Dutch book situations—this alone would hardly make any impression on anyone.

10.4. Bayesian statistics as an iterative method.

The usual association of the Bayesian statistics with the subjective theory of probability is no more than a superficial coincidence. Both the subjective theory and the Bayesian statistics use the Bayes theorem as their main technical (mathematical) tools. Bayesian statisticians choose a model using (L1)-(L5), believing, consciously or subconsciously, that the model should reflect the reality, in terms of symmetries and independence. Next they choose a prior with good technical properties, that is, such that it yields a desirable posterior distribution in combination with the smallest possible amount of data.

The prior resembles a seed in many iterative methods of mathematics and numerical analysis. Suppose that one wishes to find a function that solves a differential equation. An iterative method starts with a seed S_1, that is a function. Then one has to specify a transformation that takes S_1 into a function S_2. Usually, the same transformation is taken to map S_2 onto S_3, and so on. The method works if one can prove that the sequence

$S_1, S_2, S_3, \ldots$ converges to the desirable limit, that is, the solution of the equation. This, however, does not mean that S_1 is a solution to that equation or that it is even close to such a solution. Not all seeds will generate a sequence of S_k's converging fast to the desirable limit, and the number of iterations needed for a good approximation to the limit depends both on the problem itself and on the seed. The choice of the transformation $S_k \rightarrow S_{k+1}$ and the seed S_1 is a non-trivial problem with no general solution—the answer depends on the specific situation.

The seed is not expected to represent anything real—it is a purely mathematical entity. Recall that when a mathematician attempts to solve a differential equation using an iterative method, he may start with a function (the seed) that does not in the least resemble a solution of the equation. The treatment of the prior in the Bayesian literature clearly indicates that it is considered a seed with no subjective or objective meaning, whose properties have to be optimized from the purely technical point of view. A useful data set has to be reasonably large and one can imagine that the probability distribution is updated every time a new piece of data arrives (the final result is the same as if one used the whole data set once to derive the posterior distribution directly from the prior). The sequence of thus generated distributions plays the role of successive approximations $S_1, S_2, S_3, \ldots$ mentioned above.

The posterior distribution is the result of combining the prior and the model with the data. Since the posterior is not based on (L1)-(L5) alone, it is not always true that the posterior probability assignments are correct, in the sense of predictions, as in (L5). The Bayesian predictions may be far off the mark if the model is wrong or the prior is very unusual. The weakest point of the philosophical foundations of the Bayesian statistics is that they do not require any proof that the posterior distribution has desirable properties. The subjective philosophy not only fails to make such a recommendation but asserts that this cannot be done at all. Needless to say, Bayesian statisticians routinely ignore this part of the subjectivist philosophy and verify the validity of their models, priors and posteriors in various ways.

An important lesson from the representation of the Bayesian theory as an iterative method is that Bayesian algorithms yield little useful information if the data set is not large. What this really means depends, of course, on the specific situation. It is clear that most people assume, at least implicitly, that the value of the posterior is almost negligible when the model is not based on (L1)-(L5) or the data set is small. Nevertheless, the subjective philosophy makes no distinction whatsoever between probability values arrived

at in various ways—they are all equally subjective and unverifiable.

10.5. Who needs subjectivism?

There are (at least) two reasons why Bayesian statisticians embrace the subjective theory. One is the mistaken belief that there is no scientific justification for the use of the prior distribution except that it represents the subjective views of the decision maker, and as such it is justified by the subjective theory. I have argued that in fact the prior is a seed of an iterative method and requires only a purely technical (mathematical) justification.

Another reason for the popularity of the subjective theory among Bayesians is that the subjective theory provides an excellent excuse for using the expected value of the gain as the only determinant of the usefulness of a decision. As I argued in Sections 6.1.1 and 6.4, this is an illusion based on a clever manipulation of words—the identification of decisions and probabilities is true only by a philosopher's fiat. If probabilities are derived from decisions, there is no reason to think that they represent anything in the real world. The argument in support of using the expected value is circular—probabilities are used to encode a rational choice of decisions and then decisions are justified by appealing to thus generated probabilities.

11. TEACHING PROBABILITY

The way probability is taught at different levels illustrates well the disconnection between the frequency and subjective philosophies on one hand and the real science of probability on the other.

At the undergraduate college level and at schools, the teaching of probability starts with combinatorial models using coins, dice, playing cards, etc., as real life examples. The models are implicitly based on (L1)-(L5) and are clearly designed to imbue (L1)-(L5) into the minds of students (of course, (L1)-(L5) are not explicitly stated in contemporary textbooks in the form given in this book). Many textbooks and teachers present Kolmogorov's axioms at this point but this can do more harm than good. If we ignore the countable additivity (a purely technical and irrelevant condition at this level of instruction), the only message that axioms contain is that probability is finitely additive, that is, the probability of the union of disjoint events is the sum of their probabilities. An implicit message in this system of axioms is that independence is not worth mentioning at the most fundamental level. This is awfully confusing to students (assuming they are paying any attention) because probability without independence is useless. If the countable additivity of probability is mentioned, this can only cause even more confusion because the property cannot be understood without some mathematical training and maturity, and it is rarely, if ever, used at this level.

The frequency and subjective theories enter the picture in their pristine philosophical attire. They are used to explain what probability "really is." A teacher who likes the frequency theory may say that the proper understanding of the statement "probability of heads is 1/2" is that if you toss a coin many times, the relative frequency of heads will be close to 1/2. Teachers who like the subjective philosophy may give examples of other nature, such as the probability that your friend will invite you to her party, to show that probability may be given a subjective meaning. In either case, it is clear from the context that the frequency and subjective "definitions" of probability are meant to be only philosophical interpretations and one must not try to implement them in real life. I will illustrate the last point with the following example, resembling textbook problems of combinatorial nature. A class consists of 36 students; 20 of them are women. The professor randomly divides the class into 6 groups of 6 students, so that they can collaborate in small groups on a project. What is the probability that every group will contain at least one woman? The frequency theory suggests that the "probability" in the question makes sense only if the professor divides the same class repeatedly very many times, in an independent

way. Needless to say, such an assumption is unrealistic, and students have no problem understanding that the frequency interpretation does not apply here, except in some purely philosophical sense, as a mental model. Students quickly learn that they have to use the classical definition of probability to solve combinatorial problems such as the one presented here. A subjectivist instructor will not allow them to have any consistent view of the world, but only one of those views that conform to the accepted standards.

The teaching of probability suffers from schizophrenia—students are taught two separate theories of probability. On the scientific side, they are taught (L1)-(L5) by example, without ever stating these laws in an explicit way. On the philosophical side, they are often presented with two most popular philosophical views of probability, but with an implicit message that these philosophies must not be implemented in real life.

At the graduate level, the situation is equally strange. A graduate course in probability theory often identifies the science of probability with the mathematical theory based on Kolmogorov's axioms. In other words, no distinction seems to be made between mathematical and scientific aspects of probability. It is left to students to learn how one can match mathematical formulas and observations.

A graduate course in classical statistics implicitly introduces the long run interpretation of probability when students learn their first unbiased estimators of the mean, based on long sequences of i.i.d. random variables. On the other hand, nobody invokes collectives and von Mises' interpretation of probability when it comes to the meaning of confidence intervals.

A graduate course in the Bayesian statistics may start with an axiomatic system for decision making. The axioms and the elementary deductions from them are sufficiently boring to make an impression of a solid mathematical theory. The only really important elements of the Bayesian statistics, the model and the prior, are then taught by example. The official line seems to be "you are free to have any subjective and consistent set of opinions" but "all reasonable people would agree on exchangeability of deformed coin tosses." Students (sometimes) waste their time learning the useless axiomatic system and then have to learn the only meaningful part of the Bayesian statistics from examples.

12. ABUSE OF LANGUAGE

Much of the confusion surrounding probability can be attributed to the abuse of language. Some ordinary words were adopted by statistics, just like by any other science. In principle, every such word should acquire a meaning consistent with the scientific theory using it. In fact, the words often retain much of the original colloquial meaning. This is sometimes used in dubious philosophical arguments. More often, the practice exploits subconscious associations of the users of statistics. The questionable terms often contain hidden promises with no solid justification. I will give a very short review of some of objectionable terms that I encountered. Since the topic is tangential to the main body of the book, I will not go into details.

(i) *Expected value.* The "expected value" of the number of dots on a fair die is 3.5 and so this value is not expected at all. In practice, the "expected value" is hardly ever expected.

(ii) *Standard deviation.* The "standard deviation" of the number of dots on a fair die is about 1.7. This number is not among possible deviations from the mean: 0.5, 1.5 and 2.5.

(iii) *Subjective opinions.* See Section 9.5 for a list of nine different meanings of "subjectivity" in the probabilistic context. Only one of them fits well with the philosophical theory invented by de Finetti. This special meaning is rarely, if ever, invoked by statisticians and users of statistics.

(vi) *Bayesian learning theory.* A branch of science calling itself a "learning theory" is based on Bayesian, and so implicitly subjective, ideas. The theory studies, among other things, the evolution of subject's probabilistic opinions. The expression "Bayesian learning" is an oxymoron because it suggests a process of acquiring knowledge, perhaps even objective knowledge. Nothing of the kind is ever happening, according to the subjective theory, because there are no objective probabilities. Maintaining consistency in the subjectivist sense does not have the right to be called "learning."

(v) *Optimal Bayesian decisions.* Bayesian decisions cannot be optimal, contrary to the implicit assertion contained in the title of [1]. Decisions can be consistent or inconsistent according to the subjective theory. One can artificially add some criteria of optimality to the subjective philosophy, but no such criteria emanate naturally from the theory itself.

(vi) *Confidence intervals.* Confidence intervals are used by classical statisticians. The word "confidence" is hard to comprehend in the frequency context. It would make much more sense in the subjectivist theory and practice.

(vii) *Significant difference.* When a statistical hypothesis is tested by a classical statistician, a decision to reject or accept the hypothesis is based on a number called a "significance level." The word "significant" means in this context "detectable by statistical methods using the available data." This does not necessarily mean "significant" in the ordinary sense. For example, suppose that the smoking rates in two countries are 48.5% and 48.7%. This may be statistically significant, in the sense that a statistician may be able to detect the difference, but the difference itself may be insignificant from the point of view of social life, health care system, etc.

13. CONCLUDING REMARKS

This chapter contains a handful of general comments that did not fit well anywhere else.

13.1. Does science have to be rational?

Science is the antithesis of subjectivity. How is it possible that the grotesque subjective theory of probability is taken seriously by scores of otherwise rational people? I think that the blame should be assigned to the quantum theory and the relativity theory, or rather to their popular representations. These two greatest achievements of the twentieth century physics demand that we revise most of our standard intuitive notions of time, space, relation between events, etc. A popular image of modern physics is that it is "absurd but nevertheless true." Nothing can be further from the truth. The power of the relativity theory and quantum physics derives from the fact that their predictions are much more reliable and accurate than the predictions of any theory based on Newton's ideas. The predictions of modern physics are not absurd at all—they perfectly agree with our usual intuitive concepts. Einstein's relativity theory was accepted because it explained the trajectory of Mercury better than any other theory, among other things. Every CD player contains transistors and a laser, both based on quantum effects, but the music that we hear when the CD player is turned on is not an illusion any more than the sounds coming from a piano. Quantum physicists bend their minds only because this is the only way to generate reliable predictions that actually agree with our intuition.

My guess is that most Bayesian statisticians think that the subjective theory of probability is just like quantum mechanics. They believe that one has to start with some totally counterintuitive theory to be able to generate reliable predictions that match very well our usual intuition. Perhaps one day somebody will invent a philosophical theory representing this scheme of thinking but de Finetti's theory is miles away from implementing this idea. According to de Finetti, all probability is subjective so there are no reliable predictions whatsoever—you cannot verify or falsify any probability statement. I have argued in earlier chapters that any attempt to build reliable predictions into de Finetti's theory would completely destroy it.

13.2. Common elements in frequency and subjective theories.

Despite enormous differences between the frequency and subjective theories of probability, there are some similarities between them. One of the common elements that affected

both theories in a negative way was a conviction that in any scientific theory, the probability should be a physical quantity measurable in the same way as mass, charge or length. In other words, one should have an effective way of measuring the probability of any event. Von Mises defined probability in an operational way, as a result of a specific measurement procedure—the observation of the limiting relative frequency of an event in an infinite (or very long) sequence of isomorphic experiments. He unnecessarily denied the existence of probability in other settings. De Finetti could not think about any scientific way to achieve the goal of measuring probability with perfect accuracy so he settled for an unscientific measurement. In the subjective theory, the measurement of probability is straightforward and perfect—all you have to do is to ask yourself what you think about an event. The incredibly high standard for the measurement of probability set by von Mises and de Finetti has no parallel in science. Take, for example, temperature. A convenient and reliable way to measure temperature is to use a thermometer. However, if the temperature is defined as the result of this specific measurement procedure, we have to conclude that there is no temperature in the center of the sun. At this point, it seems that we will never be able to design a thermometer capable of withstanding temperatures found in the center of our star. Needless to say, physicists do not question the existence of the temperature at the center of the sun. Its value may be predicted using known theories. The value can be experimentally verified by combining observations of the mass, radius, luminosity, and other properties of the sun with physical theories. Von Mises and de Finetti failed for the same reason—they set an unattainable goal for themselves.

13.3. Philosophical sources of failure.

The classical philosophy of probability did not fail—it was never pushed far enough. It was a good attempt to codify the scientific knowledge of probability available at the time.

The logical theory of probability failed because its focus was the formal side of the theory. The development of the logical theory was premature, the project was mostly an exercise in pure logic. The "principle of indifference" adopted by the formal logic and its concrete implementations show the lack of full scientific understanding of probability.

The frequency theory insists that probability exists only if we can observe it in the direct way, by performing a specific type of measurement—a long run of isomorphic experiments or observations. The theory seems to be based on a philosophical assumption that every meaningful physical quantity can be effectively measured. Modern physics, specifically the quantum theory and the relativity theory, show that this need not be the case.

There is no point in denying the objective reality of the wave function of a single electron (except in purely philosophical arguments) but the function cannot be effectively measured for various reasons. Similarly, we cannot peek inside the black hole but it would make no scientific sense to claim that there is nothing inside the black hole. One could blame von Mises' mistake on his inadequate understanding of modern physics, born at about the same time as his theory. However, his philosophical error has in fact little to do with modern physics. There are easy examples, based on the classical physics, showing that we may have no effective way of measuring a quantity whose existence is unquestioned and unquestionable. For instance, we cannot check directly what is inside the earth because we lack the appropriate technology. Similarly, we cannot directly observe a human brain at work because cutting someone's head is considered unethical. Von Mises failed to notice that some physical quantities that cannot be observed directly are meaningful because we can observe them in an indirect way.

The sources of de Finetti's mistakes seem to be twofold. First, he wanted to achieve absolute certainty in the measurement of probability and failing that, he invented a theory that completely circumvented the problem. This mistake is similar to a well known philosophical trick—if you do not know how to answer a question, deny that the question makes any sense. The trick seems to have been invented by Zeno two thousand years ago; he denied the existence of motion because he could not solve a paradox.

De Finetti's second mistake was more subtle. De Finetti did not sufficiently appreciate the fact that the symmetry is one of the foundations of science. Science is meaningful only if knowledge learnt about some objects can be applied to other objects, related to ("symmetric" with) the ones that we started with. Otherwise, we would have a useless collection of unrelated facts. For example, if we measure the mass of an electron, we assume that the mass of other electrons is identical. The question of how we identify symmetric objects is a non-trivial philosophical problem but if we deny the usefulness of symmetry or if we question our ability to discern symmetries in the universe, nothing remains of science. De Finetti treated events in his philosophical theory as unique entities, unrelated to other events. This is a striking contradiction with his scientific views, based on the exchangeable events, that is, a form of symmetry. It is a paradox that de Finetti clearly thought of some families of events as objectively symmetric and so "exchangeable" but his philosophical theory could not accommodate this sound and intuitive idea—he did not find a way to include any objective statements into his philosophical theory.

A common philosophical view is that, since the mathematical theory of probabil-

ity relates probabilities to other probabilities, the role of the science, or the philosophy of probability, is to provide the initial assignment for some probability values. The frequency theory found one way of doing this—via observations of infinitely long or very long sequences. This is often impractical and in general it is philosophically unsound. The subjective theory proposed assigning initial values to probabilities using subjective opinions. This is as convincing as saying that astronomers should assign subjectively a value to the mass of the moon. Both philosophies failed to observe that people can recognize in a more or less direct way and with great reliability events which are disjoint, physically unrelated and symmetric. Moreover, people can observe whether an event occurred or not. In other words, people can apply (L1)-(L5) in a very reliable way, and these laws do not require that the initial probability values are prescribed. The main philosophies of probability failed to notice that observables in the scientific probability theory are different from probabilities. The classical and logical theories of probability put more stress on symmetry than the frequency and subjective theories and in this respect they are closer to the science of probability.

13.4. On popularity of ideologies.

A good way to tell the difference between science and ideology is to see how its proponents react to gaps in the theory, inconsistencies with observations, and similar problems. Natural sciences go to a great length to fill all the remaining gaps and in general have little tolerance for inconsistencies. Political and religious ideologies lie at the other extreme. Their followers focus on a few aspects of the ideology that they like and either ignore inconvenient facts or even deny their existence. This is often combined with some "unconventional" logic. The frequency and subjective ideologies are popular for good reasons. To this day, one of the best and most reliable ways of measuring any quantity in natural sciences, including but not limited to probability, is to perform a sequence of identical experiments. Hence, the frequency interpretation of probability is taken for granted by some physicists. It is hard to believe that they are not aware of situations when long run frequencies are not realistic, but they choose to ignore them, consciously or subconsciously.

The subjective theory seems to be popular for at least four reasons. One is that Bayesian statisticians mistakenly think that their methods are based on the subjective probability. The second one is that the subjectivist vision of statistics and all decision making can be summarized in one simple recipe, covering all simple and complicated situations alike—start with your prior, observe the data, and use the Bayes theorem to

derive the posterior, which should be taken as the basis of your decisions. The third reason for the popularity of the subjective theory is that the subjective probability framework seems to be a reasonable approximation of the human (but not only human) learning process, at least in some cases. This leads to a logical mistake that is best explained using an analogy. The human mind seems to hold an image of the universe that is consistent with Newtonian physics. This cannot be taken as the proof that Newtonian physics is a better theory than that of Einstein, but a similar reasoning is at least implicit in the arguments supporting the subjective theory. The fourth reason for the popularity of the subjective theory is its system of axioms. There is a feeling, going beyond the subjective community, that this unique feature among all philosophies of probability makes the subjective theory respectable or even unassailable. Actually, the logical theory of probability has also a formal structure. Many, perhaps most, followers of the subjective theory, failed to notice that the axiomatic system of the subjective theory is nothing more than a complicated way to say that you should use the mathematical theory of probability in your probability assignments.

My guess is that the frequency and subjective theories of probability will stay popular as long as probability users remain focused on their narrow field or some specific method. A success of a narrow statistical algorithm is often taken as a proof of the validity of a whole philosophical theory related to that algorithm.

13.5. On peaceful coexistence.

One of the philosophical views of probability tries to reconcile various philosophies by assigning different domains of applicability to them. Sometimes it is suggested that the frequency theory is appropriate for natural sciences such as physics and the subjective theory is more appropriate for other sciences, such as economics, and everyday life. I strongly disagree with this opinion. The frequency theory is extremely narrow in its scope—it fails to account for a number of common scientific uses of probability and hence it fails to be a good representation of probability in general, not only in areas far removed from the natural sciences. The subjective theory is nothing but a failed attempt to create something out of nothing, that is, to provide guidelines for rational behavior in situations where there is no relevant information available to a decision maker.

14. REFERENCES

[1] M.H. DeGroot *Optimal Statistical Decisions.* McGraw-Hill Book Co., New York-London-Sydney, 1970.

[2] D. Gillies, *Philosophical Theories of Probability*, Routledge, London, 2000.

[3] I. Hacking, *Logic of statistical inference*, University Press, Cambridge, 1965.

[4] H. Primas (1999) Basic elements and problems of probability theory *Journal of Scientific Exploration* **13**, 579–613.

[5] J.M. Ryder (1981) Consequences of a simple extension of the Dutch book argument. *British J. of the Philosophy of Science* **32**, 164–7.

[6] A. Sokal, Transgressing the Boundaries: Toward a Transformative Hermeneutics of Quantum Gravity *Social Text* **46/47**, 217–252 (spring/summer 1996). See also http://www.physics.nyu.edu/faculty/sokal/

[7] A. Sokal, Transgressing the Boundaries: An Afterword *Dissent* **43** (4), 93–99 (Fall 1996).

[8] R. Weatherford, *Philosophical foundations of probability theory*, Routledge & K. Paul, London, 1982.

Department of Mathematics
University of Washington
Box 354350
Seattle, WA 98195-4350, USA

Email: burdzy@math.washington.edu
http://www.math.washington.edu/~burdzy/