On the foundations of statistics:

A frequentist approach

by

Frank Hampel

Research Report No. 85
June 1998

Seminar für Statistik
Eidgenössische Technische Hochschule (ETH)

CH-8092 Zürich

Switzerland
On the foundations of statistics:
A frequentist approach

Frank Hampel
ETH Zurich, Switzerland

Abstract: A limited but basic problem in the foundations of statistics is the following: Given a parametric model, given perhaps some observations from the model, but given no prior information about the parameters ("total ignorance"), what can we say about the occurrence of a specified event A under this model in the future (prediction problem)? Or, as probabilities are often described in terms of bets, how can we bet on A?

Bayesian solutions are internally consistent and fully conditional on the observed data, but their ties to the observed reality and their frequentist properties can be arbitrarily bad (unless, of course, the assumed prior distribution happens to be the true prior). Frequentist solutions are generally not possible with ordinary probabilities; but it is possible to define "successful bets" (using upper and lower probabilities), which even lead out of the state of total ignorance in an objective learning process converging to the true probability model. A special variant (successful bets on random parameter sets) provides a new and correct interpretation of the basic idea of Fisher's "fiducial probabilities."

Successful bets (which are only one-sided and not fully conditional) can be used for inference, but not for decisions which, as in Bayesian "fair bets," require ordinary probabilities. However, it is possible to define "best enforced fair bets" (or corresponding probability distributions) which solve the decision problem in a specific minimax sense. In as much as they are also Bayes solutions, they may be called "least unsuccessful Bayes solutions" (providing another candidate for "least informative priors"). On the other hand, among the successful bets we can select the "least unfair successful bets" which in a way come closest to Bayes solutions. Several (nontrivial) examples, mainly for two independent binomials, have already been worked out by R. Steiner and the author.

1 Introduction and overview

1.1 Motivation and outline

The motivation for the research presented here was that I was always interested in foundations, but dissatisfied with the situation in foundations of statistics, where two major schools, with several variants, fight each other back and forth over centuries without being able to find a common basis or to give satisfactory answers to the objections that can be raised against them.

As one of the clearest properties of a probability is its frequency interpretation, I tried to find out where and how far one gets with a basically frequentist approach. At the same time, I tried to see how one can learn, or make inference, from data. It turned out that the most common frequentist interpretation ("repeated sampling principle") does not really seem suitable for inference or learning, and that I had to develop a largely new frequentist approach.

In order to assess the quality of inferential procedures, both in principle and empirically, I found it much simpler and more natural to consider predictive inference, rather than inference about parameters. The derived probabilities are epistemic, which means they relate to our state of knowledge about Nature, and not aleatory, referring hypothetically to the true but unknown state of Nature. (I found this distinction more and more important.) At the same time, the probabilities are intersubjective, namely the same for all scientists with the same (scientific and experimental) background knowledge, and not subjective as in most other systems of inference. They can be empirically checked (up to a point) by considering sequences of independent different (not identical) experiments. As they are to describe incomplete knowledge about a stochastic system, I found it necessary to use upper and lower probabilities as a rule. These can be interpreted by means of one-sided bets, and their frequentist interpretation is formulated in the
concept of “successful bets.” Upper and lower probabilities can easily and naturally describe the “state of total ignorance” (about a parameter, given a parametric model), and the new theory allows also to get out of the state of total ignorance in an objective learning process, whose predictions converge to those under the true parametric model as more and more observations come in.

A rather special variant, namely inference about certain random parameter sets, contains what is correct about Fisher’s “fiducial probabilities” in particular situations, but it works also for discrete models, thus extending Fisher’s work and embedding it into a much more general theory.

When we consider decisions instead of inference (a distinction which is not always made), there are well known arguments that one should act as if one had proper probabilities, which can be described by two-sided bets. Best known are Bayesians “fair bets”; however, these are rarely “fair” in a frequentist sense except under the prior distribution used to derive them. Some potential expected losses are unavoidable in view of our partial ignorance about the true stochastic model. The minimax risk (in a rather strong sense) of these two-sided bets is provided by the “best enforced fair bets.” They usually are also specific Bayes solutions which may be called “least unsuccessful Bayes solutions,” and the corresponding priors are another candidate for “least informative priors.” In a sense, the best enforced fair bets are the fully conditional and coherent solutions closest to having a frequentist interpretation.

On the other hand, we can ask for the frequentist solutions, namely the successful bets, which are closest to being coherent and fully conditional. They are called “least unfair successful bets” and provide a distinguished solution to the inference problem.

The two risks (for successful bets under the two-sided betting paradigm — their risk under one-sided bets is zero —, and for fair bets under either betting paradigm) can be measured on the same scale and thus provide a unified framework for frequentist and Bayesian methods.

Some numerical examples for 2 independent binomials (one observed in the past, one in the future), starting with total ignorance about the parameter, have already been worked out.

There exists some work on extensions in different directions: successful bets for normal and other distributions and in the robustness context (Hampel 1996), start with partial knowledge about the parameter (bridging total ignorance and Bayesian priors) (Hampel 1993a), tests, and a related information theoretic concept (“cautious surprises”) (Hampel 1993a). However, this work can only be briefly mentioned in this paper.

1.2 Other approaches to the foundations of statistics

Work on foundations can be coarsely classified in two groups: frequentist and Bayesian. One of the main frequentist schools is the Neyman–Pearson theory (cf., e.g., Lehmann 1959), with Wald’s (1950) decision theory as a superstructure. The other main frequentist approach is the work by R. A. Fisher (cf. Fisher 1956), which is often rather close to the Neyman–Pearson theory, but which also contains two special branches: likelihood approach and fiducial probabilities, that are sometimes considered separate approaches to foundations.

Among Bayesians, we can crudely distinguish logical or “objective” Bayesians who try to prescribe a canonical prior distribution in the case of total ignorance about the parameter, and subjective or subjectivist Bayesians who allow a free choice of the prior. (In practical applications, we often find “pragmatic Bayesians,” who start with a not well defined “vague prior” in order to get reasonable results.) Among objective Bayesians are Bayes (1763) (as far as he was a Bayesian), Laplace (1812), and Jeffreys (1939). (Roughly in this direction, but rather remote from practical statistics, is the work by various philosophers, such as Carnap (1950) and Hempel (1966), on inductive logic.) A recent approach towards an objective Bayesian theory is given by Bernardo’s (1979) “reference priors.” Subjectivist Bayesians, presently the main Bayesian school, include de Finetti (1937), L. J. Savage (1954), and Lindley (1965).

I. J. Good (1983) calls himself a Bayesian, but he actually works with upper and lower probabilities. The latter are central for the “belief function theory” in various variants by Dempster (1968), Shafer
(1976), Smets (1991), and others. There are some formal similarities of belief functions with the "possibility theory" by Dubois and Prade (1988), and also with Zadeh’s (1965) “fuzzy set theory.”

Before I give some comments on these approaches from the viewpoint of philosophical foundations, I should hasten to emphasize that all these approaches, if used reasonably and judiciously, can give useful and reasonable results in practical applications. Fortunately for applied statistics, good data–analytic practice is possible despite the different schools.

1.3 Some comments on other approaches

The Neyman–Pearson school restricts itself explicitly to aleatory (ontological, physical) probabilities, that is, unknown and hypothetical “true” probabilities in Nature. This is at the same time a major strength and a major weakness. The advantage is that the theory is completely “rational” and “objective” in the style of pure mathematics (“if .... then ...”), very clear in its statements, and therefore easily palatable to mathematicians, and easily integrated into mathematics. The disadvantage is that its statements are often frustratingly weak for the users of statistics (who therefore tend to interpret things into them which are not there at all). Neyman (1957, 1971) explicitly said: “There is no inductive inference, only inductive behavior.” Hence no proper learning process is possible within the theory: the applied statistician does not “know” anything, he just acts. The rules given for inductive behavior, although they work in practice, appear rather arbitrary and ill–justified in many situations.

A more general point is that most deductions are within a given parametric or nonparametric setting; but this is true also for the other approaches, and the problems of model choice and robustness remain partly open on a higher level of foundations. They will not be discussed here further, but we should stay aware that they exist.

Fisher partly used aleatory probabilities as did Neyman and E. S. Pearson, but he also used epistemic probabilities, though he was perhaps not always clear about the distinction. I got the impression that Fisher’s early agreement and later disagreement with Neyman’s confidence intervals may have been due first to the common aleatory basis, and later to Fisher’s noticing that Neyman refused to allow any epistemic, “post hoc” interpretation of confidence intervals — something every “naive” user of statistics wants to have, at least subconsciously. He wants to be able to say, for example: “Within my present state of knowledge, I can bet 19:1 that this confidence interval contains the true parameter”; and indeed his bet will be “successful,” to use the new terminology; though not always fully conditional, as we shall see later. Very closely related are Fisher’s fiducial probabilities, though Fisher himself does not seem clear about their epistemic character; the step from single fiducial probabilities to full “fiducial probability distributions,” which then are easily though mistakenly treated like “ordinary” (aleatory) probability distributions, seems dangerous to me and needs a clear justification and description of the inherent limitations. In most situations, Fisher had no fiducial probabilities, but only likelihoods. In a sense, likelihoods carry all the information (given the model), but not in an interpretable way. They are an essential nucleus both for frequentist and Bayesian approaches, but in both approaches they need a further transformation, be it with a prior, or be it in such a way that they yield a frequentist statement.

The Bayesian paradigm would be ideal, were it not for the choice of the first prior distribution, which is rarely justified on frequentist grounds. Bayes (1763) himself described a purely physical model in most of his paper, and only in his famous Scholiwm did he speculate why the uniform prior might be the prior of choice also in the case of no physical background model and total ignorance about the parameter. Much in the Scholiwm is incomplete and between the lines. He argues that the continuous uniform for the parameter yields discrete uniform predictive distributions for all n. These are not changed by parameter transformations, thus evading the main criticism against choosing the uniform; moreover, as we know now (but what Bayes might only have felt or guessed), the predictive uniforms in their totality determine the continuous uniform for the parameter, and they are distinguished as the discrete distributions with the largest entropy. Thus, the choice of the uniform is justified in a deeper information theoretic sense. The only other distinguished priors I am aware of are Jeffreys’ prior, which is the uniform in the transformation of the parameter space.
in which the Fisher information is constant, and the uniform in the natural parametrization of an exponential family; furthermore, in dependence of the questions asked, Bernardo’s (1979) reference priors.

Bayes was looking for a solution in terms of proper probability distributions, and he might be interpreted as saying between the lines in the Scholium: “If I consider only ordinary probabilities, I see no other choice but ... .” He did not consider upper and lower probabilities, although they were not completely unknown at that time. Otherwise he might have arrived at very different solutions.

Neobayesianism started with de Finetti (cf., e.g., de Finetti 1937), who was a very deep and radical thinker. However, he carried subjectivism to the point that it lost any connection with the idea of an “objective” outside physical world and became fully solipsistic. Few scientists will follow him that far, otherwise they would not be scientists.

Less radical Bayesians point out that under mild conditions, the Bayes solutions converge asymptotically towards the true solutions, and that one should use rather flat (“vague”) priors in one of the customary parametrizations; but although there is evidence that this very often works in practice, there is usually no theoretical justification that the asymptotic argument works for a specific $n$, or that a specific prior can be considered “vague.”

Belief function theory, possibility theory, and also fuzzy set theory do not work with proper probabilities (except as special cases), but rather with something like upper and lower probabilities (although the rules used may be quite different). Contrary to Neobayesians, they all can easily describe total (and partial) ignorance formally, but in common with them, they all make also only subjective statements without any frequentist interpretation. The only bridge to aleatory probabilities I could find in lengthy discussions was the “Hacking principle”: if an aleatory probability is known to be $p$, then the corresponding subjective probability should also be taken as $p$. I was also told that the concepts of these theories have new meanings (different from aleatory probabilities) in the same way in which probability once was a new concept; but I could not find out an operational meaning of the specific numerical values given. Nevertheless, in a few examples I computed, the numerical values derived by belief function theory were surprisingly fairly close to the numbers obtained by my frequentist method (cf. Sec. 3.2), suggesting that they may be of some use in practice despite their unclear interpretation.

2 The inference problem and successful bets

2.1 The basic problem considered

One of the basic problems of statistics and inductive reasoning (but by no means the only one) is the following:

- Given a parametric model,
- possibly given an observation (or observation vector) from that model,
- starting either with the state of total ignorance or with some prior information (not necessarily a Bayesian prior) about the parameter,
- what can we infer about a future observation (or observation vector) from that model?
- Specifically, what can we say about its chances of falling into a prespecified set $A$?

A prototype example, occurring in many applications, are two independent binomial distributions (one observed in the past, one to be observed in the future) with the same parameter $p$.

For example, given that a novel medical treatment or operation, about whose quality nothing was known a priori, was successful in 3 out of 5 cases, what can we say about the chances that it will be successful in exactly 2, or in at least 2 out of 4 future cases?
We shall mainly consider predictions, rather than inference about parameters, for various reasons (cf. also Hampel 1997a):

- Ultimately, we want to predict.
- Predictions can (usually) be checked, parameter estimates and tests not (since we know the parameters at most approximately, or “asymptotically”).
- It is much clearer how to assess the quality of predictions (“right or wrong”) than that of estimation procedures (which criterion to use?).
- Prediction is connected with a specific predictive situation; it therefore allows and necessitates considerations of its quality which are not possible for an “all purpose” parameter estimate.

The question arises how to describe numerically the “chances” of a future event, and how to interpret these numbers. We shall use upper and lower probabilities, which can be described by means of one–sided bets, and which find an operational interpretation in the form of sequences of independent “successful” bets. But first we need some explanations.

### 2.2 Upper and lower probabilities

Given a measurable space \((\Omega, \mathcal{A})\), we can define on the \(\sigma\)–algebra \(\mathcal{A}\) not only probabilities, but also, more generally, upper and lower probabilities, with values between 0 and 1. A corresponding upper probability \(m\) (which we shall denote by \(m\)) and lower probability \(\underline{m}\) are related by \(m(A) = 1 - \underline{m}(A^c)\) for all \(A \in \mathcal{A}\), hence it suffices to consider only one of them. While \(m\) has another, information–theoretic interpretation (which we shall not discuss here, cf. Hampel 1993a), so that we shall use both forms. It seems natural to require \(m(A) \geq \underline{m}(A)\) for all \(A \in \mathcal{A}\), and usually also \(\underline{m}(\Omega) = 1\). Otherwise there is much freedom in defining upper and lower probabilities.

More special ones can be defined as suprema and infima of sets of probability distributions. A Choquet capacity (Choquet 1953/54) is monotone in \(\mathcal{A}\) and fulfills some topological continuity conditions; if in addition it fulfills \(m(A \cup B) + m(A \cap B) \leq m(A) + m(B)\), for all \(A, B \in \mathcal{A}\), it is called 2–alternating (Choquet) capacity (and \(m\) is called 2–monotone capacity). These capacities of order 2 have nice properties and play a certain role in robust statistics (Huber and Strassen 1973, Huber 1973, 1976, 1981), but I do not think that all useful upper and lower probabilities have to be of this form. Even more beautiful and restrictive are Choquet capacities of infinite order which are used in belief function theory for finite spaces (Shafer 1976). My approach is that I deduce my upper and lower probabilities from their required properties of interpretation and do not impose any structure axiomatically. If (as it sometimes turns out) they happen to be Choquet capacities of infinite order, this is nice, but by no means necessary in my theory.

### 2.3 Probabilities and bets

One of the oldest and most natural interpretations of probabilities is in terms of bets, and betting odds. For example, if I bet 3:1 on an event \(A\) (with a stake \(s\), often taken to be \(= 1\)), I gain \(s \cdot 1/(3 + 1) = s \cdot 1/4\) if \(A\) occurs, and I lose \(s \cdot 3/(3 + 1) = s \cdot 3/4\) if \(A^c\) occurs. If \(P(A)\) is the true probability of \(A\), my expected gain \(E[G] = (P(A) - 3P(A^c)) \cdot s/4\). This is \(\geq 0\) iff \(P(A) \geq 3/4\). Hence, if I don’t want a negative expected gain, I shall offer this bet voluntarily if and only if I believe or know that the true probability is at least 3/4.

Proper probabilities correspond to 2–sided bets: if I am willing to bet 3 : 1 on \(A\), I am also willing to bet 1 : 3 on \(A^c\), and I shall offer both bets exactly if I believe or know that the true probability \(P(A) = 3/4\). My expected gain is then \(= 0\) for either bet, if my belief (or knowledge) is correct; otherwise one of the two bets will have a negative expected gain. In terms of the discussion in the previous paragraph: If both bets, 3 : 1 on \(A\) and 1 : 3 on \(A^c\), are acceptable to me, then I
must be convinced that \( P(A) \geq 3/4 \) and \( P(A^c) \geq 1/4 \), hence that \( P(A) = 3/4 \) and \( P(A^c) = 1/4 \). (You may note that these numbers refer to my personal convictions — which we call “epistemic probabilities” —, and thus it is gratifying to see that in this situation they behave like an ordinary probability in that their sum is equal to one.)

The picture looks simpler if we start with a known objectively true (“aleatory”) probability \( P(A) \) (which we naturally — by “Hacking’s principle” — adopt as our personal conviction). Then we can bet \( P(A)/(1 - P(A)) \) on \( A \), and \( (1 - P(A))/P(A) \) on \( A^c \), and both bets have expected gain zero. But in practice we often do not know the objective probability exactly.

Lower (and hence the equivalent upper) probabilities correspond to one–sided bets: if my zero. But in practice we often do not know the objective probability exactly. \( P \) can bet \( (which we naturally — by ”Hacking’s principle” — adopt as our personal conviction). Then we

The picture looks simpler if we start with a known objectively true (“aleatory”) probability \( P(A) \) (which we naturally — by “Hacking’s principle” — adopt as our personal conviction). Then we can bet \( P(A)/(1 - P(A)) \) on \( A \), and \( (1 - P(A))/P(A) \) on \( A^c \), and both bets have expected gain zero. But in practice we often do not know the objective probability exactly.

Lower (and hence the equivalent upper) probabilities correspond to one–sided bets: if my zero. But in practice we often do not know the objective probability exactly. \( P \) can bet \( (which we naturally — by ”Hacking’s principle” — adopt as our personal conviction). Then we

Please note that me offering a bet 1 : 3 on \( A \) does not imply that I also offer a bet 3 : 1 on \( A^c \), as the Bayesians require. (Perhaps I offer a bet 2 : 1 or only 1 : 3 on \( A^c \); I confess partial ignorance about my knowledge of \( P(A) \).) The Bayesian requirement is justified in a decision situation, as Bayesians have shown, but not in inference, namely a description of the knowledge I have. (Some statisticians, such as Neyman and perhaps some Neobayesians, find inference irrelevant or nonexisting or being identical to decisions; but in the sequel we shall keep the distinction between inference and decisions.) Bayesians argue correctly that it would be foolish to offer bets 3 : 1 on \( A \) and simultaneously 3 : 1 on \( A^c \) (in general: to have the sum of the lower subjective probabilities of \( A \) and \( A^c > 1 \)), because if both bets are accepted simultaneously, one would incur a sure loss (and if one bet would be accepted at random with probability 1/2, one would have a negative overall expected gain, no matter what the true \( P(A) \)). But in the context of inference, I found in Neobayesian writings only a spurious symmetry argument why the sum of the lower subjective probabilities should not be < 1 either, and hence should be = 1. (Perhaps there was also some subconscious confusion with the idea of a “fair bet.”)

### 2.4 Bets with partners

I have been asked, especially by Bayesians, with whom I would bet with my one–sided bets. But the same question can be asked of Neobayesians with their (pairs of) “fair bets.” Bayesians (in this context I always mean subjectivist Neobayesians) can rely on any other Bayesian to be willing to bet against them (if he has the money and time), but not on any other statistician. (If a Bayesian is willing to bet 1 : 3 on \( A \) and on \( A^c \), and my lower probabilities are both only 1/4, I shall not accept either bet, because I know too little. I would select one bet, though, if my lower probabilities were 2/3 and 1/6, for example.) Bayesians with an unwilling partner then would try to enforce acceptance of one of the two bets offered, but this is enforcing a decision, and NOT eliciting the state of knowledge of the partner. Bayesians never allow (and are never allowed) to say: “I don’t know.”

Even if Bayesians bet against Bayesians, the situation is more complicated than usually described. Bayesians seem happy and content to evaluate their expected gain with their own subjective prior probability distribution, and then of course their expected gain is always zero, no matter against whom they bet and which bet the partner selects. But under the true objective probability (a concept apparently rather foreign to Bayesians) they are often bound to lose in the average.

For example, it is easy to describe situations with two Bayesians betting against each other, and whoever offers his bet first has a negative expected gain. (Let \( a \) and \( b \) be the subjective probabilities of partners \( A \) and \( B \) for an event considered, and \( c \) the true probability, with \( a < c < b \). If \( A \) offers his bets first, \( B \) will select the one he thinks is advantageous; it is not as good as he thinks, but it still will give him an advantage.) And naturally, if a Bayesian bets against someone who knows the true probabilities, he is bound to lose on the average, unless he also knows the true probabilities.

If I offer a one–sided bet, not every Bayesians will be willing to accept it. But if the distribution of subjective prior distributions is as large as is permitted by de Finetti and others, I could always
find some Bayesians who think accepting my bet is fair or even to their advantage. Even if I bet 0 : 1 on some event \( A \) (that is, if I pay nothing if \( A' \), and gain everything if \( A \) happens), I could find Bayesians willing to accept the bet, namely those who believe in \( A' \) with probability 1.

But the main point is not whether and where I could find actual betting partners. The bets are only a tool to describe my state of knowledge, and it suffices to consider fictive bets. Since I want to make actual claims about the quality of my bets (namely that my expected gain, and hence in the long run approximately my average gain, is nonnegative), I need a neutral (fictive) betting partner who accepts all my bets; this neutral partner may be Nature, my own scientific conscience, or a “statistician’s stooge” (I. J. Good, orally). If my partner were allowed to decline the bets that seem disadvantageous to him (and if he might guess or know something about the true probabilities), the desired numerical claims would be impossible. This holds, of course, also for any objective claims possibly made by Bayesians.

### 2.5 Aleatory and epistemic probabilities

Before we come to the definition of “successful bets,” let us briefly return to the possible interpretations of the concept of “probability.” During my work on foundations, I found it more and more important to distinguish between aleatory (ontological, physical) probabilities (namely the ones objectively in Nature, in our random mechanism or experiment, but normally unknown to us) and epistemic probabilities (which describe what we know, or believe to know, about the true unknown aleatory probabilities).

As we have seen, epistemic proper probabilities can be quite naturally described by two–sided bets (pairs of bets, which Bayesians call “fair bets”), and epistemic lower probabilities can be simply described by one–sided bets.

Epistemic probabilities can be subjective (as in Neobayesian theory) or intersubjective, which means the same for different scientists with the same observations. Intersubjective probabilities can be logical (derived from some prescribed rules, which have no direct frequentist interpretation, as in Jeffreys’ approach, but also in Bayes’ Scholium and in Bernardo’s reference priors, as well as in the work of inductive logicians, such as Carnap) or frequentist (allowing a direct frequentist interpretation, as Fisher did in part, and as we shall do).

It has been asked what the basic probability space for epistemic probabilities is. Some further formal clarification or development of a suitable notation may be needed here, but a less formal description shall be attempted now. Epistemic probabilities typically refer to some class of random statements or claims about Nature, and are derived from the aleatory distribution(s) of some random variable \( X \), which also helps to select the statement.

For example, we start with a “universal” aleatory probability statement \( P(\theta \geq X - c) = \alpha \) \((c, \alpha, \text{and of course } \theta \text{ fixed})\) which holds simultaneously for all \( \theta \) permitted. (This is a typical situation leading to “fiducial probabilities.”) Then \( P(\text{the statement } S := “\theta \geq X - c” \text{ is true}) = \alpha \), with the random statement depending on \( X \). Observing \( X = x \), and not knowing more about \( \theta \), the statement \( S_x := “\theta \geq x - c” \) (e.g. “\( \theta \geq 8) is true with probability \( \alpha \), if \( X \) was allowed to take on all possible values according to its true (though unknown) distribution under \( \theta \), no matter what \( \theta \). There is nothing random about \( \theta \); but the 8 might have been, for example, a 12 or a \(-1,1\), and hence the whole statement is randomly selected. We don’t know the probability (or density) that \( X - c = 8 \), but we know that the statement selected by \( X \) in this way has probability \( \alpha \) of being true. Hence we can bet on it (“fairly”) at odds \( \alpha : (1 - \alpha) \). Of course, the actual statement “\( \theta \geq 8 \)” is either true or false, that is, the (trivial) aleatory probability of it being true is either 1 or 0, but the point is that we do not know this aleatory probability, while we can embed the statement into a larger class of statements of which we know the probabilities of them being true. The statements are conditional statements, given \( X = x \); but their contents are absolute claims (“\( \theta \geq 8 \)”), while most results on aleatory probabilities are only conditional results (“if \( \theta = \ldots \), then \( \ldots \)”) which do not allow any absolute claims about any facts that might have been learned from the data.
2.6 Successful bets

We are now close to the definition of a “successful bet.” Briefly and simply, it is a one–sided bet (and corresponding lower and upper probability) with expected gain \( \geq 0 \), no matter what the true \( \theta \) is (within its permitted range). However, there are some more technical details to be discussed, namely on what to bet, and what the prior knowledge used for the bet can be.

We assume that a parametric model \((\Omega, A, \{F_\theta, \theta \in \Theta\})\) is given, and in this paper we usually assume that we start with the “state of total ignorance” about the parameter (symbolically denoted by \( M_0 \equiv 1 \)), namely that nothing is known a priori about \( \theta \). (More generally, we can start with some limited set of prior distributions for \( \theta \) considered possible, up to the extreme of a single Bayesian prior, cf. Hampel 1993a.) In most cases, we then have observed a (generally vector–valued) random variable \( X \) on our probability space which provides us with some information on \( \theta \). Typically, we now consider another random variable \( Y \) on the same space to be observed in the future (\( X \) and \( Y \) do not have to be independent, though they often are), and we bet on the event that \( Y \) falls into some fixed set \( A \) in its range. The bet depends on the observed \( X = x \). A variant of the set–up is that the event \( A \) also depends on \( X \) (as with prediction intervals). Another variant is that the set \( A = A(X) \) is not in the range of some future \( Y \), but is a subset of the parameter space (as with confidence intervals and fiducial probabilities); in this case, as it turns out, \( A \) has to be random for nontrivial inference to be possible.

Bayesians have criticized the concept of “state of total ignorance” about a parameter as fictive, since there always were some knowledge available. Elsewhere (e.g., in Hampel 1995) I have given two examples where most Bayesians would be hard pressed to proclaim any genuine knowledge: the probability that presently a camel walks on the main street of Almaty, and the probability of encountering a tree kangaroo in a certain time–space region of Papua New Guinea. (A Bayesian actually procured a number in the latter situation, but this was an arbitrary decision based on no prior knowledge whatsoever, with no introspection allowing more insight, and with the many (?) books on tree kangaroos in PNG just not being available in the situation.) But my main point is that the “state of total ignorance” is often a simple and useful approximation to reality, as good an approximation as a Bayesian prior in other situations (and even on the conservative side, while a Bayesian prior by the same token claims more knowledge than is available). And often our prior knowledge will be somewhere in between the two extremes; then inference based on a start with total ignorance is trustworthy even if wasting a bit of information, while inference based on a Bayesian prior is shaky and unreliable by an unknown amount.

With no prior knowledge and no past observation \( X \), in general only the trivial bet \( 0 : 1 \) can be successful. If an \( X = x \) has been observed, it can always be a very unlikely and misleading observation, and if we evaluate our expected (nontrivial) gain conditionally on this \( x \), it will be negative. Hence we have to be able to balance the few misleading \( x \)–values by averaging the expected gain over the full distribution of \( X \). This definition works and gives nontrivial and informative results, as we shall see.

Starting with total ignorance about \( \theta \), and denoting the random gain of the bet by \( G \), we can summarize the previous discussions and results by writing formally:

**Definition:** A bet on \( A \) (more precisely a class of bets, depending on \( x \)), described by \( m \), is called successful iff

\[
E_\theta G := E_\theta^X E_\theta^Y G \geq 0 \quad \text{for all } \theta, \text{ or equivalently iff } \\
E_\theta^X m(Y \in A|X) \leq P_\theta(A) \quad \text{for all } \theta.
\]

(The superscript denotes the random variable with respect to whose distribution the expectation is taken. — If we want to incorporate the start with total ignorance formally into \( m \), we can write \( m(Y \in A|X, M_0 \equiv 1) \).

An operational interpretation of successful bets is the following. Consider a sequence of independent different (!) experiments, and one successful bet for each experiment. Then under some mild conditions on the sequence of distributions of the random gains (the stakes must neither explode nor implode too quickly, and the distributions must not tend to be singular too quickly, so that a
— sufficient, not always necessary — nondegenerate law of large numbers for the form $\sum c_i X_i$, $c_i$ fixed, $X_i$ random, can still work), given any $\epsilon > 0$ and $\delta > 0$, the average gain will be $\geq -\epsilon$ with probability $1 - \delta$ if the number of experiments $n$ is sufficiently large.

### 2.7 Examples for successful bets.

Classical examples for successful bets, though somewhat atypical in our framework (since $A$ depends on $X$ and may even be in the parameter space), are provided by (conservative and exact) prediction and confidence intervals and fiducial probabilities; in fact, the latter, as far as they are valid, are nothing but a very special case in our framework (cf. Hampel 1993a), but they have been grossly over– and misinterpreted. Other examples, starting not from total ignorance but from the other extreme, a Bayesian prior, are all Bayes solutions. Examples for the normal, exponential with shift, and other continuous distributions (with fixed $A$) are given in Hampel (1996), including a first sketch of robustness considerations (“What happens if the parametric model is only approximately true?”). We shall here now consider some of the “simplest” and most basic discrete examples, as indicated in Section 2.1.

Let $X \sim B(n, p)$ (past observation) and $Y \sim B(k, p)$ (future observation) be two independent binomial random variables with the same success parameter $p$ ($0 \leq p \leq 1$) about which no prior knowledge is available. As alluded to before, if $n = 0$ (no past observation), the only successful bets on $Y \in A$ for any $k$ and any atomic (one–point) $A$ (and “most” other $A$’s) are the trivial ones, namely $0 : 1$ (since $P(Y \in A) = 0$ is a possibility).

If $n = k = 1$ (one Bernoulli trial each in past and future), we can without loss of generality consider $A = \{1\}$. How can we bet successfully on “$Y = 1$,” given $X = x$?

If $X = 0$, we must bet $0 : 1$, for $P(X = 0) = 1$ and $P(Y = 1) = 0$ is a possibility (it happens with $p = 0$). Thus $m(Y = 1|X = 0, M_0 \equiv 1) = 0$, with the conditional expected gain (given $X = 0$) $E^X Y[G(X = 0) = p \cdot 1 - q \cdot 0 = p \geq 0$, where $q := 1 - p$. Now take any $c$ ($0 \leq c < 1$) and bet $c : (1 - c)$ on $A$ if $X = 1$. Then the conditional expected gain (given $X = 1$) is $E^X Y[G(X = 1) = p(1 - c) - qc = p - c$ which is $\geq 0$ if $p \geq c$. Since $p = 0$ is possible, the only successful conditional bet is again the trivial bet $0 : 1$, which shows I cannot get out of the state of total ignorance with fully conditional bets. However, if I am allowed to average the bets over the distribution of $X$, I obtain $E^X E^Y G = p(p - c) + qp = p(1 - c) \geq 0$, no matter what $p$ is; and this for any $c$. Hence my second conditional bet is completely arbitrary; I can even offer the “bold” bet $1 : 0$ on $Y = 1$ if $X = 1$, and I am still successful (with $EG \equiv 0$).

Thus we have a problem of nonuniqueness, of having to select a particular successful solution. Intuitively, neither the “bold” solution nor the “ignorant” solution $0 : 1$ seems satisfactory (the latter not at all, since we would have learned nothing from $x$); rather, we might want to select a solution in between, perhaps closer to the “bold” one. But a formal proposal will have to wait until we have discussed the decision problem and “enforced fair bets” (cf. Sec. 4.2).

If $n = k = 2$ (two Bernoulli trials each in past and future), there are already 6 nontrivial sets $A$ asking (because of the symmetries $x \to 2 - x$ etc.) for 4 nontrivial solutions, each one depending on 3 possible values of $x$. Let us now switch to the (equivalent, but in some sense more natural) notation for the upper probabilities $m(Y \in A|X = x)$ for 4 nontrivial solutions, each one depending on $X$.

For example, if we fix $A = \{1\}$ and write $m(\{1\}|x) =: m(x)$ ($x = 0, 1, 2$), we have to find all solution vectors ($m(0), m(1), m(2)$) with $0 \leq m(x) \leq 1$ which fulfill $P_Y(\{1\}) = 2pq \leq q^2 m(0) + 2pq m(1) + p^2 m(2)$ for all $p$ and $q = 1 - p$. This is the type of problem we have to solve, if necessary numerically by brute force (though there may be also shortcuts and approximations, and even formulas). We shall restrict ourselves to symmetric solutions, namely with $m(0) = m(2)$, which seems most natural, though there are also nontrivial asymmetric solutions, as Steiner (1995) has shown. A symmetric vector of $m$–values is successful iff $m(0) + m(1) \geq 1$, as can be checked with the above inequality. The solutions on the boundary $m(0) + m(1) = 1$ may be called “admissible” successful solutions, in analogy with decision functions, since they cannot be
“improved” any further; but this does not imply they make sense intuitively. Among the admissible successful vectors \((m(0), m(1), m(2))\) are the trivial and uninformative solution \((1/2, 1/2, 1/2)\), the extreme solution \((0, 1, 0)\) which is in some sense minimax \((\sup_p(E_p^X m(X) - P_p(\{1\})) = 0)\), the “paradoxical” extreme solution \((1, 0, 1)\), and many others, including \((1/4, 3/4, 1/4)\). The latter is of the following form: let \(m_1(x)\) be the probability of \(A\) under the maximum likelihood estimator \(\hat{p}\) of \(p\) based on \(x\) (namely \(\hat{p} = 0/2 = 0, 1/2, 1)\), hence \(m_1(x) = 2\hat{p}(1 - \hat{p})\); then \(m(x) = \min((m_1(x) + c), 1)\) with \(c = 1/4\) being the smallest constant which makes the \(m-\) functions of this form successful. This solution looks nice also in other ways (cf. Hampel 1993a); we shall return to it after discussing how to select an “optimal” successful solution.

For other sets \(A\), there is only one admissible successful solution: for \(A = \{0\}\) : \((m(0), m(1), m(2)) = (1, 0, 0)\); for \(A = \{0, 1\}\) : \((1, 1, 0)\); for \(A = \{0, 2\}\) : \((1, 0, 1)\); the remainder by symmetry. The solution \(\min((m_1(x) + 1/4), 1)\) is also successful, though not admissible, with values \((1, 1/2, 1/4), (1, 1, 1/4), \) and \((1, 3/4, 1)\) for the same sets \(A\), respectively.

Thus we have seen that it is possible to derive nontrivial exact successful solutions, even for a fixed \(A\) and for discrete, asymmetric and noninvariant distributions. It can also be shown that some solutions (both upper and lower probabilities) converge to the true probability of the set \(A\) as \(n \to \infty\) (for fixed \(k\)), using confidence intervals for \(p\) with increasing levels. Before we can return to the selection problem, we have to make an apparent detour and study the decision problem.

### 3 The decision problem and enforced fair bets.

#### 3.1 Best enforced fair bets and probability distributions.

Bayesians (and others) have convincingly shown that if one wants to make optimal decisions, one has to act as if one has a proper probability distribution. We shall not go into these arguments again, but we note that in one form of the belief function theory, there is a “pignistic transformation” from lower probabilities describing beliefs to proper probabilities used for decisions and actions (Smets 1990). And the updating, given new observations or information, is not done on the probabilities level via Bayes’ theorem, but on the belief functions (lower probabilities) level, without incurring a “sure loss” so dreaded by Bayesians outside their own theory (Smets 1993). Even if we confess total ignorance about the probabilities of \(A\) and \(A^c\) we may be forced by circumstances to allocate one part of a stake on \(A\) and the remainder on \(A^c\), and we want to minimize somehow our expected loss. The question arises how to obtain a proper probability distribution as a counterpart in a decision situation to successful bets. (Cf. Hampel 1993b.)

We remember that proper probabilities correspond to two-sided bets: if Peter bets \(p : q\) on \(A\), he is also willing to bet \(q : p\) on \(A^c\). If Peter has proper prior probability \(p/(p + q)\) for \(A\), his subjectively expected gain for either bet is zero, and Bayesians call this pair of bets a “fair bet.” Peter’s betting adversary Paul — who may even know the true \(P(A)\) — is allowed to choose either of the single bets, but as long as Peter’s prior is correct, both have expected gain zero. If not, Peter may (or even will) lose on the average, of course (depending on Paul’s knowledge).

Now we note that in our paradigm, a one-sided bet is actually a whole class of bets, depending on the past observation \(x\). The strongest form of transforming it into a two-sided bet is to allow Paul to change sides or not as a function of the past observation \(x\). That is, Paul is not only allowed to accept the whole bundle of bets offered to him by Peter, or else to offer it himself to Peter in return; rather he can consider every conditional bet, given \(x\), separately and decide which side to take.

Since Paul may know or guess the true probability distribution, Peter can in general not bet successfully if he offers two-sided bets (unless he knows the true probability distribution himself). All he can try to do is to minimize his expected loss in some sense. A strong (but feasible) requirement is to minimize the maximum expected loss over all probability distributions allowed and over all choices of sides by Paul. Again, as with successful bets, if Peter minimizes his conditional expected loss, given \(x\), he cannot learn from \(X\). (He cannot utilize the fact that under
some parameter, some $X$’s are much more likely than others, and always has to assume the worst possible situation that could give rise to $x$, no matter how unlikely; that is, he cannot get out of the state of total ignorance.) Hence the expectation has to be taken also over the distribution of $X$. If Paul knows the true distribution (the worst possible case for Peter), he will for every $x$ select the bet favorable for him.

In detail: if Peter entertains an $m(A)$ for some event $A$, he offers the bets $(1 - m) : m$ on $A^c$ and, since now $m$ is a probability (an enforced two–sided bet), also $m : (1 - m)$ on $A$. With stake 1, Peter’s expected gain from the latter bet is $(1 - m)P(A) - m(1 - P(A)) = P(A) - m$, and from the former bet $m - P(A)$. If Paul knows $P(A)$, he will select the bet to make sure that Peter’s expected gain is $-|m - P(A)|$. Now, if $m = m(A|X)$ depends on a past observation $X$, Peter’s overall expected loss is $E^X[m(A|X) - P(A)]$. Introducing the parameter $\theta$ (unknown to Peter), we see that Peter’s task is to find a function $m(A|X)$ (depending on $x$, but not on $\theta$) which minimizes $\sup_\theta E^X[m(A|X) - P_\theta(A)]$. Thus we arrive at the

Definition: $w = w(A|x)$ (a function of $x$) is a “best enforced fair bet” (starting with total ignorance about the parameter) iff

$$w = \arg \min \sup_{m, \theta} E^X |m(Y \in A|X) - P_\theta(A)|. $$

We see that $w$ yields for every $x$ a probability distribution on the trivial $\sigma$–algebra consisting only of $A$ and $A^c$ (besides $\Omega$ and $\emptyset$), and that $\hat{w}(A'|X) := 1 - w(A|X)$ yields the identical distributions starting from $A^c$.

If we want to consider bets not only on $A$ and $A^c$, but on all events $A_i$ in some larger $\sigma$–algebra, we can use the following

Definition: $w = w(A_i|x)$ (as a function of $x$ and the $A_i$) is a “best enforced probability distribution” on the $A_i$ (starting with total ignorance) iff

$$w = \arg \min \sup_{m, A_i, \theta} E^X |m(Y \in A_i|X) - P_\theta(A_i)|. $$

Depending on the bets allowed (that is, on the future events of interest), we thus obtain “Bayesian priors” for a start with total ignorance which have a well–defined optimality property. Interestingly — and gratifyingly —, they share their dependence on the event $A$ considered with Bernardo’s (1979) reference priors.

It is easily possible to replace the start with total ignorance about the parameter by a start with partial knowledge, up to the other extreme of a Bayesian prior (which in turn may even be a single point mass). Compare Hampel (1993a) for a broader basic paradigm.

3.2 Examples for best enforced fair bets and probability distributions

Let us consider again 2 independent binomials $B(n, p)$ (past) and $B(k, p)$ (future), and start with total ignorance about $p$.

The first example is the famous case $n = 0, k = 1$: how should I assign a probability to an event $A$ about whose potential of occurrence I know absolutely nothing?

Let us derive the best enforced fair bet. If Peter bets $c : (1 - c)$ on $A$ ($0 \leq c \leq 1$) and if $P(A) = p$, then his expected gain is $E_p G = (1 - c)p - c(1 - p) = p - c$ if his partner Paul does not switch sides, and $= c - p$ otherwise. If Paul knows $p$ (or guesses it correctly), he will switch sides iff $c \leq p$, hence in this least favorable case Peter’s expected gain is $E_p G = -|c - p|$ and $\inf_p E_p G = -\max(c, 1 - c) \leq -1/2$. Now Peter can choose $c$ such that he minimizes his maximum expected loss, and this happens exactly for $c = 1/2$ with minimax $R_2 := \min_c \max_p (-E_p G) = 1/2$.

Hence his best enforced fair bet on $A$ (and likewise on $A^c$) is $1 : 1$, or “fifty to fifty.”

This result looks very familiar. It is also the result of the “principle of insufficient reason,” of any symmetric Bayesian prior, and of the pignistic transformation of the vacuous belief function.
(describing total ignorance). However, the point is that this result has been derived in four different ways; and in other situations these ways lead to different results, as we shall see.

Already for \( n = 0, k = 2 \) (a “biased coin,” about which nothing else is known, is to be tossed twice independently), we obtain a variety of different results. If we bet only on some \( A \) and its complement, we obtain for \( A = 0 \), in obvious notation: \( w(0) = w(\{1, 2\}) = 1/2; \) correspondingly for \( A = 2 : w(2) = w(\{0, 1\}) = 1/2; \) but for \( A = 1 : w(1) = 1/4 \) and \( w(\{0, 2\}) = 3/4 \).

On the other hand, if we allow bets on all six nontrivial subsets simultaneously, we obtain the “best enforced probability distribution” given by \( w(0) = w(2) = 1/2 \) (and \( w(1) = 0 \) etc., by additivity), with \( R_e := \min_w \max_A \max_{p}(−E_p G) = 1/2 \). This is a Bayes solution exactly for the prior \( P(p = 0) = P(p = 1) = 1/2 \), but for none of the common priors, such as uniform, Jeffreys’, or any “vague” prior. The pignistic transformation of the vacuous belief function (Smets 1990) yields \( P(0) = P(2) = 5/12 \) and \( P(1) = 1/6 \) which is different from our solution, but interestingly numerically somewhat similar.

For \( n = 0, k = 3 \) and bets on all subsets simultaneously, we obtain \( w(0) = w(3) = 5/12 \) and \( w(1) + w(2) = 1/6 \), that is, the solution is not unique. For \( w(1) = 0 \) it cannot be derived from any Bayesian prior, and for symmetric \( w \) the corresponding Bayesian priors are different from the ones for \( n = 0, k = 2 \). The result of the pignistic transformation is again different, though numerically somewhat similar.

Perhaps of more practical interest is the case \( n = 1, k = 1 \) (a biased coin to be tossed once again, after one toss). Here the best enforced fair bets \( w(A|x) \) are: \( w(1|1) = 3/4 \) with \( w(0|1) = 1/4 \), and \( w(1|0) = 1/4 \) with \( w(0|0) = 3/4 \), and the minimax risk \( R_e = 1/4 \). This solution is also a Bayes solution, namely for, e.g., Jeffreys’ prior; we may call it a “least unsuccessful Bayes solution.”

For comparison: the “bold” bets \( m(1|1) = m(0|0) = 1, m(0|1) = m(1|0) = 0 \), which are successful as one-sided bets, have sup risk = 1/2. The trivial constant bets (with \( m \equiv 1/2 \)), which are optimal if one is not allowed to average over the distribution of \( X \), and which are fully conditional on \( x \), but allow no learning and no leaving the state of total ignorance, also have sup risk = 1/2. And Laplace’s “rule of succession” based on the uniform prior, with \( m(1|1) = m(0|0) = 2/3, m(1|0) = m(0|1) = 1/3 \), has sup risk = 1/3.

4 A bridge between frequentist and Bayes approaches.

4.1 Least unfair successful bets.

There are two desirable aims for statistical procedures for inference and decisions:

(i) They should have a frequentist interpretation: the long–run properties (of independent claims for different experiments or situations — not necessarily “repeated sampling” from the same population without any learning effect) should be correct;

(ii) They should be “coherent” (cf. Walley 1991); the claims should be internally fully consistent (not only avoiding sure loss, as frequentist procedures can also do, but not even wasting any information, contrary to, e.g., randomization in the design of experiment), and in particular they should be fully conditional.

Bayesian statistics fulfills (ii), but not (i) (unless the assumed prior happens to be the true aleatory (objective) prior).

Very little of traditional “frequentist” statistics fulfills (i) for unconditional statements (mainly confidence and prediction intervals, and fiducial probabilities, if interpreted correctly), but the theory of successful bets does, as we have seen. On the other hand, all frequentist methods rarely fulfill (ii).

In general, it does not seem possible to achieve both (i) and (ii) (except if some artificial randomization is built into the procedure), because this would usually imply full knowledge of the
probabilistic model. The latter is in fact assumed in Bayesian statistics, but it makes statistics in a sense rather trivial by reducing it to mere conditioning on new data.

However, within the class of procedures fulfilling one desideratum, we can try to come as close as possible to fulfilling the other desideratum (after defining suitable measures of closeness, of course).

Starting with epistemic proper probability distributions or systems of “fair bets” which fulfill (ii) (and which, as we noted, strictly include, but probably are almost identical with all Bayes procedures), we asked for the distributions closest to a frequentist claim (in the worst possible case), namely the “best enforced fair bets” (including the “least unsuccessful Bayes solutions”), and their discrepancy to being “successful” is measured by the minimax risk $R_e$.

Correspondingly, we can now consider all “successful” solutions (which fulfill (i)) and ask for the ones which come closest to coherence in the sense that their maximum risk ($\leq 0$, as long as we consider only one–sided bets) is minimized if we now allow arbitrary switching of sides and hence the most general form of two–sided bets. This leads us to the

**Definition:** $w = w(A|x)$ is a “least unfair successful bet” (starting with total ignorance) iff

$$w = \arg \min_{m: \text{m successful}} \sup_{\theta} E_\theta [m(Y \in A|X) - P_\theta(A)]$$

(and analogously for a system $\{A_i\}$, with minimax risk $R_f$).

This definition works and gives nontrivial and intuitively reasonable results, as we shall see, and it finally solves our problem of selecting a particular successful bet among the many candidates in an appealing way.

Obviously, $0 \leq R_e \leq R_f$, since we have restricted the class of $m$–functions allowed in the competition. $R_f = 0$ means full conditioning (a.a.x) of the corresponding $m =: m_f$ and gives successful and fair pairs of bets (a.a.x). This happens for certain predictions for random events ($A = A(X)$) with prediction intervals and confidence intervals (and fiducial probabilities), and — not starting from total ignorance — in the Bayesian case of a known aleatory true prior. In this latter case, the Bayes solution is the least unfair successful bet (and the best enforced fair bet and least unsuccessful Bayes solution — there is only one Bayes solution permitted).

Taking a more general view, given a fixed $\sigma$–algebra of $A_i$ (typically with $A$ and $A^c$ as nontrivial members), given a one–sided collection of bets by $m(A_i|x)$, we may define the one–sided risk $r_1(m) := \sup_{A_i} \sup_{\theta} E_\theta [P_\theta(A_i) - m(A_i|X)] \leq 0$ for collections of successful bets, and $\geq 0$ for $m$ being probability distributions, and the two–sided risk $r_2(m) := \sup_{A_i} \sup_{\theta} E_\theta [P_\theta(A_i) - m(A_i|X)]$. (For a more general framework which does not only start from total ignorance and which also includes Bayes solutions, cf. Hampel 1993a.) The vectors $(r_1(m), r_2(m))$ describe an accessible region $R$ in the plane as $m$ ranges over all possible one–sided bets. We know that $R$ is bounded by $r_2(m) = R_e$ and by $(0, R_f) \in R$ from below, and also that $(R_e, R_e) \in R$ (since the best one–sided version of an $m$ yielding $R_e$ can at most be better), and in the future we may perhaps want to investigate the “admissible” part of the boundary of $R$ between $(0, R_f)$ and $(R_e, R_e)$, in analogy with decision theory, as a class of optimal compromises between successful bets (described by upper and lower probabilities) and least unsuccessful proper probability distributions.

### 4.2 Examples for least unfair successful bets.

We now consider some examples for least unfair successful bets, again in the case of two independent binomials $B(n, p)$ (past) and $B(k, p)$ (future) and starting from total ignorance about $p$. Most results were obtained by R. Steiner (1995), compare also Hampel (1997b).

For $n = 0, k = 1$ (and more generally $n = 0$ and all $A$ whose probability can range from 0 to 1), the only successful, hence also the least unfair successful bets are the trivial ones $0:1$, with $R_f = 1$.

For $n = k = 1$, we obtain the first interesting result. We remember (cf. Section 2.7) that in order to be successful, if $X = 1$, any bet $c : (1 - c)$ on $A = \{1\}$ is allowed, while with $X = 0$ we
have to bet $0 : 1$ on $A$. Now which $c$ is least unfair? Assume our betting partner Paul is allowed to switch sides, and he knows $p$. If $X = 0$, he will always switch sides, making my conditional expected gain $= -p$. If $X = 1$, he will switch iff $p_{(>)} c$, making my conditional expected gain $= -p - c$. My overall expected gain is then $-pq - |p - c|$, namely $= -p(1 - c)$ if $p_{(>)} c$ and $= p(1 - c) - 2pq$ if $p_{(<)} c$. The $c$ which minimizes the maximum expected loss over all $p$ is given by $1 - c = 6 - 4\sqrt{2} \approx 0.3431$, hence $c \approx 0.6569$ (closer to 1 than to 0, as intuitively expected, cf. Section 2.7), with minimax risk $R_f = 6 - 4\sqrt{2} \approx 0.3431$. The corresponding $m$–function for $A = \{0\}$ is thus $(m(A|0), m(A|1)) = (1, 0.3431)$. For comparison: the best enforced fair bet (cf. Section 3.2) on $A = \{0\}$ was $(0.75, 0.25)$ with $R_e = 0.25$.

For $n = k = 2$, the vectors $(m(A|0), m(A|1), m(A|2))$ of the $m$–functions for least unfair successful bets are: $(0.25, 0.75, 0.25)$ with $R_f = 0.25$ for $A = \{1\}$, $(0.2256, 0.4640, 1)$ with $R_f = 0.3065$ for $A = \{2\}$, $(1, 4 - 2\sqrt{3} = 0.5359, 1)$ with $R_f = 0.2679$ for $A = \{0, 2\}$, and $(1, 1, 6 - 4\sqrt{2} = 0.3431)$ with $R_f = 0.3431$ for $A = \{0, 1\}$ (the remainder by symmetry).

These solutions are remarkably close to (and in one case identical with) the likelihood–based successful bets of the form $\min(m_k + 1/4, 1)$ (cf. Section 2.7), giving some encouragement to the hope that solutions of the latter type may provide simple and usable approximations in more complicated situations when computation of the exact solutions is out of the question.

For comparison: M. Wolbers (personal communication) has done preliminary computations of the best enforced fair bets and probability distributions for $n = k = 2$. According to his results, the two types of best enforced fair bets have risks $R_e = 0.1716$ (for $A = \{1\}$) and $= 0.2198$ (for $A = \{0\}$), respectively, and the best enforced probability distribution has risk $R_e = 0.24148$, which is not much lower than the values above. It is to be expected that for $n$ increasing (and fixed $k$), the differences between the optimal successful and optimal fair solutions and between their risks will become smaller and smaller.

## 4.3 Summary

We have seen that it is possible to build a statistical theory and to define a learning paradigm that fulfills the strict frequentist criterion of “successful bets” and even leads out of the state of total ignorance. This theory has connections to virtually all other statistical theories, but is quite novel in its approach. It distinguishes between aleatory and epistemic probabilities, and it uses upper and lower probabilities interpreted as one–sided bets for the description of our incomplete knowledge. In passing, it provides an interpretation for fiducial probabilities, as far as they are valid, as a very special case of the theory and embeds them into a much broader framework (which also allows a continuous transition from successful bets in discrete models to fiducial probabilities in continuous models, cf. Hampel 1993a). Solutions are given for the inference problem with two independent binomials (one in the past, one in the future), hence also for what K. Pearson (1920, 1921) called “the fundamental problem of practical statistics.” For other solutions, see Hampel (1996).

While successful bets are most appropriate for inference (to describe what we know), for decisions (namely how to act on the basis of our incomplete knowledge) we need “fair” bets and proper probability distributions which contain (and are almost identical with) the Bayes solutions. The solutions closest to the frequentist criterion of being successful in a (rather strict) minimax sense are called “best enforced fair bets” and contain the “least unsuccessful Bayes solutions” with another candidate class for “least informative priors.” Again, we have looked at several numerical solutions.

On the other hand, we may search for the successful bets which minimize the maximum expected loss in the decision (two–sided betting) situation, called “least unfair successful bets.” Again, we have derived and given some numerical results.

The least unfair successful bets appear to be the most appropriate choice for purposes of inference, while the best enforced fair bets (including the least unsuccessful Bayes solutions) appear to be most suitable for decisions. It may be possible to find optimal “compromises” between these two
solutions, though their practical purpose is at present not clear. More important would it be to find simple asymptotic approximations for these two solutions, perhaps based on likelihoods and maximum likelihood estimators, as some examples might faintly indicate.

For any Bayes procedure, we can assess the maximum expected loss \((r_1 \geq 0)\) incurred because in general it does not have a frequentist interpretation; and for any successful bet \((r_1 \leq 0)\) we can assess the (even larger) maximum expected loss \((r_2 \geq R_f \geq R_e)\) incurred in the two-sided betting situation (for which it was not designed primarily) because in general it is not fully conditional. We thus have reached a certain symmetry between frequentist and Bayes solutions: we can measure their main strengths and defects by just 2 numbers on the same scale.

References


Smets, P. (1993). No Dutch Book can be built against the TBM even though update is not obtained by Bayes rule of conditioning, Technical Report TR/IRIDIA/93–9, Institut de Recherches Interdisciplinaires et de Développements en Intelligence Artificielle, Bruxelles.

