How To Remove the Ad Hoc Features of Statistical Inference within a Frequentist Paradigm

I. A. Kieseppä and Malcolm Forster

ABSTRACT

Our aim is to develop a frequentist theory of decision-making. The resulting unification of the seemingly unrelated theories of hypothesis testing and parameter estimation is based on a new definition of the *optimality* of a decision rule within an ensemble of token experiments. It is the introduction of ensembles that enables us to avoid the use of subjective Bayesian priors. We also consider three familiar problems with classical methods, the *arbitrary features of Neyman-Pearson tests*, the difficulties caused by *regression to the mean*, and the relevance of stopping rules, and show how these problems are solved in our extended and unified frequentist framework.

1 Introduction

Classical statistics, which was developed in the 1930's by Neyman and Pearson and by R. Fisher, is often referred to as 'frequentism'. This term refers to the philosophy underlying its apparently unrelated methods. According to this philosophy statistical methods should minimize the error that results from a statistical inference in a way that applies, ideally, to every possible state of the world. In a sense, it takes a *world-centric* viewpoint. On the opposing side there is the newer philosophy of Bayesian statistics, which takes a *person-centered* point of view — namely, the point of view that one should optimize statistical methods in the light of all the information available to the person who is making the inference.

The clashes between these two philosophies are as complicated and diverse as are the problems and methods that each camp uses. In particular, both kinds of methods can be applied to the seemingly unrelated problems of *hypothesis testing* and of *estimating the value of a parameter* from noisy data. In the first problem there are just two rival hypotheses, each of which is concerned with the probability distribution of the observed outcome, and the error to be minimized is simply the falsity of the accepted hypothesis. In contrast, the estimation of a parameter corresponds to a continuum of possible hypotheses, and in this case the error to be minimized is usually the squared difference between the inferred parameter value and the actual parameter value.

A further distinction can be drawn between the kinds of hypotheses being tested: there are tests *between two simple hypotheses*, each of which specifies *some particular* distribution of probability values over the space of observational outcomes, and there are *tests between composite hypotheses*, each of which is compatible with many different probability distributions. It is worth noting that among the traditional frequentist solutions to the problems of the above classification, the solution to the problem of the last type —that is, Neyman-Pearson tests between composite hypotheses—has had the most enduring success in statistics. This is partly because of the lack of competition until recent years. It is therefore the area in which the frequentist methodology is most solidly entrenched, and we believe that there are sound reasons for this fact. Certainly, the classical theory of estimation is also a major part of statistics. However, it has transformed itself into a cornerstone of Bayesian statistics, and it is also the

foundation of a Neo-Fisherian school of statisticians, called Likelihoodists, whose theory is concerned with the strength of evidence, instead of being a theory of statistical inference (Royall [1997]). In all cases the classical theory of estimation appears to have lost its frequentist foundations.

Below we shall look at Neyman-Pearson testing and parameter estimation from a new perspective. We introduce a new notion of *optimality*, which is different from Neyman and Pearson's notion of a best test, and argue for its advantages. The new notion of optimality is compatible with a Bayesian point of view. However, we give it a non-Bayesian interpretation, which leads to significant philosophical differences.

It will be seen how our framework succeeds in seeing merit in Neyman-Pearson statistics (Mayo [1996]) and in the Bayesian approach (Earman [1992]), although we wish to reject both the arbitrary features of classical methods and the extreme subjectivism of the Bayesian alternative. We propose a third philosophy of statistics, which is a new variant of frequentism, and which is grounded on an objective notion of optimality. We begin by recapitulating the basic ideas of the frequentist approaches to hypothesis testing and to estimation, and by presenting two familiar criticisms of them. Our extended frequentist framework not only answers these criticisms, but it sheds light on other many other problems as well. The relevance of stopping rules (section 11) is a particularly important example.

2 The Arbitrary Features of Neyman-Pearson Tests

After the seminal work of Neyman and Pearson in the 1930s, before the rise of Bayesianism, it would be accurate to say that hypothesis testing was a *universal* part of statistical practice. Moreover, classical hypothesis testing remains a central part of the way that science is practiced today (Mayo [1996])—although it is more solidly entrenched in some sciences than it is in others. The variety of examples to which it has been applied is enormous, yet all applications follow the same pattern of inference. We have chosen an example that does not presuppose any prior knowledge of a particular science—the coin tossing example. Yet this very simple example is sufficient to illustrate the pattern of inference and the philosophical issues that arise in every example. The reader should not misjudge the importance of these issues by the scientific

unimportance of this particular example. The philosophical problems are neither artifacts of an over-simplified example, nor issues of 'merely' historical interest.

A Bernoulli trial is the result of a chance process in which one of two possible outcomes occurs with probability θ or $1-\theta$, respectively, with the additional requirement that the outcomes of repeated trials are probabilistically independent. We shall consider a situation in which there are just two rival hypotheses about the correct value of the Bernoulli parameter, θ. For example, suppose that a coin is taken at random from a box containing two coins, such that the selected coin is tossed repeatedly, and the other coin is left in the box. The problem is to decide which coin was taken from the box from the observation of the coin tosses. More specifically, we shall assume that one of the coins is characterized by the value $\theta = 1/3$, and the other one by the value $\theta = 3/4$.

If we denote the result "heads up" by H and the result "tails up" by T, the result of each coin toss will be either H or T. If the chosen coin is tossed N times, the corresponding observations can be represented with a sequence of the letters H and T of length N. For example, when $N = 2$, the observations are represented by one of the sequences HH, HT, TH, and TT. The two possible values of the Bernoulli parameter, 1/3 and 3/4, correspond to different probability distributions on the space of these alternatives.

In the standard procedure, due to Neyman and Pearson, the considered probability distributions are divided into two families, the first of which usually contains only one distribution, which is called the null hypothesis. The null hypothesis, h_0, is tested against an alternative h_1 which states that one of the other considered probability distributions is the actual one. The hypothesis h_1 is, of course, composite whenever two or more alternative probability distributions are under consideration. A Neyman-Pearson test is a procedure whose aim is to provide good reasons for believing that the null hypothesis h_0 is false.

When the aim of a researcher is to falsify some particular simple hypothesis (like the hypothesis that two random variables are independent), that hypothesis is *the* obvious choice for the null hypothesis (Forster [2000]), but in other cases, like in our coin flipping example, the choice of a null hypothesis is necessarily quite arbitrary. One of two hypotheses under consideration in our example states that the true distribution corresponds to $\theta = 1/3$, and the

other one states that it corresponds to $\theta = 3/4$. We shall take the former of these hypotheses to be the null hypothesis.

The *size* of a test between the null hypothesis h_0 and its alternative h_1 is, by definition, the probability of erroneously rejecting the null hypothesis when it is true, *i.e.* of choosing h_1 when h_0 is true. If the other hypothesis h_1 is simple, the *power* of the test can be defined to be the probability of (correctly) rejecting h_0 when h_1 is the actual probability distribution. In this case a *best test* is a test which has the largest power among all the tests of its size (see *e.g.*, Hogg and Craig [1978], p. 261). In classical hypothesis testing one normally fixes the size of the test *by convention*, finds the best test of the given size, and makes use of that test.

We shall illustrate the contents of these definitions in the case in which a coin is chosen with the procedure mentioned above and in which *just one* coin flip is observed, so that $N = 1$. Now there are just two possible observed outcomes, H and T, and it is clearly the case that $\Pr(\mathrm{H}|\theta = 1/3) = \frac{1}{3}$, $\Pr(\mathrm{T}|\theta = 1/3) = \frac{2}{3}$, $\Pr(\mathrm{H}|\theta = 3/4) = \frac{3}{4}$, and $\Pr(\mathrm{T}|\theta = 3/4) = \frac{1}{4}$. A test is an 'acceptance' procedure in which one observes an outcome — which has to be either H or T — and, given the outcome, then chooses either the null hypothesis $\theta = 1/3$ or its alternative $\theta = 3/4$.

There appear to be just four 'acceptance' procedures of this kind. Firstly, there are the two *a priori* procedures in which one chooses either $\theta = 1/3$ or $\theta = 3/4$ independently of the evidence and, secondly, there are two procedures in which one chooses $\theta = 1/3$ for one of the two possible outcomes H and T, and $\theta = 3/4$ for the other one. Below, the procedure in which $\theta = 1/3$ is chosen in both cases will be called T_1, and the procedure in which $\theta = 3/4$ is chosen in both cases will be called T_2. One of the two remaining procedures consists in choosing $\theta = 1/3$ when H is observed and $\theta = 3/4$ when T is observed.. We shall call this procedure T_3. Intuitively, T_3 appears to be an unreasonable way to proceed, because in this procedure one accepts the hypothesis that 'goes against' the evidence. Finally, there is procedure in which $\theta = 3/4$ is chosen when H is observed and $\theta = 1/3$ when T is observed. In this procedure, which we shall call T_4, one responds to the evidence in a way which intuitively seems to be the correct one.

By definition, the test T_1 has both size 0 and power 0, because it never leads to the rejection of the null hypothesis, and the test T_2 has both size 1 and power 1, because it always leads to the rejection of the null hypothesis. Among the two more interesting tests T_3 and T_4, the intuitively unreasonable test T_3 has the size 2/3, since when this test is used the probability of rejecting the null hypothesis $\theta = 1/3$ when it is true is 2/3, and the power 1/4, since the probability of choosing the alternative hypothesis $\theta = 3/4$ when it is true is only 1/4 when this test is used. Similarly, the size of T_4 is 1/3 and the power of T_4 is 3/4. The poverty of the test T_3 is reflected in the fact that it has a large size but a small power.

If the four tests T_1, T_2, T_3 and T_4 were all the tests that there are, one would have to view each of them as a best test. Since all these tests have a different size, each of them is in a trivial sense the most powerful test of its size among them. If these tests were all that there are, one would have to conclude that the test T_3, which "goes against the evidence," should also count as a best test. The standard method of avoiding this implausible result is to observe that there are other tests besides T_1, T_2, T_3, and T_4. For example, one might consider a randomized decision procedure **Q** which always chooses $\theta = 3/4$ when H occurs, but utilizes some random procedure for choosing $\theta = 3/4$ with the probability 50% and $\theta = 1/3$ with the probability 50% whenever T occurs. It is easy to verify that the size of this test (i.e. the probability with which it yields the result $\theta = 3/4$ when, as a matter of fact, the correct result would have been $\theta = 1/3$) is 2/3, which is identical with the size of the test T_3. However, the power of this randomized test is larger than the power of T_3 (7/8 > 1/4).

To sum up, we have seen that the standard Neyman-Pearson test procedure contains at least three kinds of arbitrary features: it is not always clear which hypothesis should be taken to be the null hypothesis, the size of the test must often be fixed by convention, and the best test of a given size might well be a *randomized test*, in which case the chosen distribution depends not only on the evidence but also on the result of some random process. On the other hand, when the size of the test has been fixed, and when the alternative of the null hypothesis is simple, the necessity of making a choice between the different tests of the given size does not usually involve any further arbitrary choices. As already stated, one normally chooses the test that has the largest power among the tests of the given size.

A well known theorem, due to Neyman and Pearson ([1933], pp. 298-302), states that *likelihood ratio tests* are always best tests of their size (we present this theorem as a special case of more general theorems in section 10 below). However, when the alternative of the null hypothesis is a composite hypothesis the power of the test is usually not well-defined,[1] and it is well known that the Neyman-Pearson theorem does not generalize in any straightforward way to this situation. In particular, Rubin and Stein (see Lehmann [1950])[2] have independently discovered examples which show that likelihood ratio tests can perform very badly when a composite alternative hypothesis has been chosen suitably. In such cases there is sometimes no obvious answer to the question which of the tests of the given size should be chosen. Hence, also the choice between the tests of a given size can sometimes be rather arbitrary.

3 Estimation and Regression to the Mean

Next we shall have a look at the classical frequentist theory of estimation. We shall consider the problem of estimating an unknown quantity θ^* that is responsible for an experimental outcome x in accordance with the equation

$$x = \theta^* + u ,$$

where u is a 'white noise' term which has a normal (bell-shaped) distribution with the mean 0. More specifically, we shall assume that x denotes the *average* of n results of measurements of the same quantity. If each of these measurement results has a normal distribution whose mean is the correct value θ^* and whose variance is σ^2, and if the distributions of the measurement results are independent of each other, it will be the case that $u \sim N(0,\sigma^2/n)$.

For example, in a case like this x might be the average of n measurements of the weight of an object, in which case θ^* is its true weight, or x might be the average of *e.g.* several cholesterol readings, in which case θ^* is the true cholesterol value. In each case x is what is observed and θ^*

---

[1] In this case one has to rest content with defining for the test a *power function* which specifies, for each of the simple hypotheses h_S which are compatible with h_1, the probability of correctly rejecting h_0 when h_S is true. (See *e.g.*, Hogg and Craig, 1965, pp. 255-256.)

[2] We thank Branden Fitelson for pointing out the existence of these examples.

is unknown, as is *u*. *Estimators* of θ are quantities which are calculated on the basis of the observations without knowing the value of θ^*. Below, we shall denote an estimator of θ by $\tilde{\theta}$. Mathematically estimators can be represented by functions of the evidence which do not depend on θ^* (see e.g. Cramér ([1946], p.477). But which function of the evidence should $\tilde{\theta}$ be?

The *squared error* $\left(\tilde{\theta} - \theta^*\right)^2$ is often used as a measure of success of an estimator $\tilde{\theta}$. However, even if this number *actually* happens to be small, it might still be the case that the estimator $\tilde{\theta}$ is a bad one. The success of an estimator might, of course, also simply be dumb luck. Accordingly, in frequentist statistics estimators are standardly evaluated in terms of their *average* performance. By definition, an *unbiased* estimator is such that its expectation value $E\left(\tilde{\theta}\right)$ is the actual value of the estimated parameter, θ^*, and a *best* estimator is an unbiased estimator for which the expectation value of its squared error, $E\left[\left(\tilde{\theta} - \theta^*\right)^2 \middle| \theta^*\right]$, receives its smallest value within the family of all unbiased estimators (see e.g. Hogg and Craig [1965], p. 205). According to the world-centric, or frequentist, point of view, one should prefer best estimators to other unbiased estimators, since they produce good estimates *on average*, where the average is taken over different possible states of the world, *i.e*. different possible observed data.

There are well-known theorems that show that, when some regularity conditions are satisfied, the *maximum likelihood estimators* are asymptotically (for large *n*) the best estimators even in cases in which the error distribution is not normal. (See e.g. Cramér ([1946], pp. 498-99). A maximum likelihood estimator maps *x* to the value $\tilde{\theta}$ that maximizes the probability density value corresponding to *x*. In the simple case at hand, if the distribution of *x* is $N(\theta^*,\sigma^2/n)$, then the maximum likelihood estimator of θ^* is simply the average of the measurement results, i.e. *x*. In this case the estimator $\tilde{\theta} = x$ has exactly the same standard error for every value of θ^*, since the number $E\left[\left(\tilde{\theta} - \theta^*\right)^2 \middle| \theta^*\right]$ is the variance of the distribution of *x* given θ^*, and this is equal to σ^2/n. Hence, in the context of this example the world-centric approach really appears to work.

No matter what the true value of θ^* is, so long as it is related to x by the equations $x = \theta^* + u$ and $u \sim N(0,\sigma^2/n)$, the maximum likelihood estimator is *the* efficient estimator.[3]

However, what is often called *regression to the mean* causes problems for this frequentist account of estimation. A well known example of this phenomenon is the tendency for a father of a very tall son to be shorter than the son, and for the father of a very short son to be taller than the son. The surprising feature of this phenomenon is that it is largely a population-level effect. For instance, suppose that in every token case the distribution of the heights of children is symmetrically spread above and below the (average) height of the parents. That is, suppose that x represents the height of a child, and θ^* is the (average) height of the parents, and that for every possible value of θ^*, $x = \theta^* + u$, where u is a symmetrical distribution with mean zero. The mere fact that the different values of θ^* within the population are concentrated about a mean value m, according to a normal (bell) distribution is sufficient to imply that if x is well above m in a token case then one should suspect that the value x is too large an estimate of θ^*. However, a maximum likelihood estimator takes the data at face value. To use another example, when a student scores very well on an aptitude test, the classical theory of estimation advises us to suppose that the student has the aptitude that corresponds to that score. Of course, the classical theory recognizes that the high score may be due to good luck—questions that the student happened to know—but it does not take into account the fact that this case is more likely than the case in which the score is too low.

This seems rather unsatisfactory. It is clear that when all the information that we have is the score, it is rational to go by the score. However, if we have additional relevant information, we should be able to use it. In the context of our example such additional information can easily be represented with a prior distribution on the possible values of θ^*. For example, we may suppose

---

[3] This is true for this simple example despite the surprising fact that there exist biased estimators (Stein [1956], James and Stein [1961]) with even smaller standard errors *for every possible value of the parameters* in the case in which there are 3 or more parameters in the problem. Efron [1978] argues that the existence of Stein estimators speaks in favor of the frequentist approach, but his arguments are controversial in light of the more recent discovery that Stein estimators can be obtained by assuming a particular prior distribution (e.g. Lehmann [1983], p. 299). We are grateful to Branden Fitelson for pointing us to these references.

that in the situation which is under consideration the person whose θ^* value is being measured has been selected randomly from a population in which θ^* values are distributed approximately in accordance with a Gaussian (normal) distribution $N(m, s^2)$ with the mean m and the variance s^2. If in this case our test score is x, and if x is much larger than m, it seems reasonable to suggest that the best guess for θ^* is somewhere between m and x.

There is one way in which a frequentist can make use of the prior information contained in the distribution $N(m, s^2)$. Instead of searching for an estimator that is best for a particular value of θ^*, one can ask which estimator is best *on the average* for all values of θ^*, when the average is taken relative to the prior distribution $p(\theta^*)$. In other words, since the success of an estimator $\tilde{\theta}(x)$ is for each fixed value of θ^* measured by

$$E\left[\left(\tilde{\theta}-\theta^*\right)^2\middle|\theta^*\right],$$

its average success can be measured by

$$E_1=\int E\left[\left(\tilde{\theta}-\theta^*\right)^2\middle|\theta^*\right]p(\theta^*)d\theta^* \tag{1}$$

By definition, this quantity equals

$$E_1=\int\left[\int\left(\tilde{\theta}(x)-\theta^*\right)^2 p\left(x|\theta^*\right)dx\right]p(\theta^*)d\theta^* \tag{2}$$

Formula (2) shows that the procedure in which the estimator $\tilde{\theta}$ is chosen by maximizing E_1 can be characterized also as follows. One first calculates, given an arbitrary value of θ^*, the average squared error of each considered estimator $\tilde{\theta}$. The result will be a function of θ^*, but not of x. One then calculates the expected value of this result relative to the prior distribution $p(\theta^*)$. The result of this calculation will depend on neither x nor θ^*. The task is then to find the function $\tilde{\theta}(x)$ which minimizes the latter quantity.

Often one considers the estimators of the form

$$\tilde{\theta}(x)=(1-\eta)x+\eta m, \tag{3}$$

where η is a fixed number between 1 and 0, and *m* is arbitrary fixed number. Also the maximum likelihood estimator is of this form, since the formula (3) yields the maximum likelihood estimator when one puts $\eta = 0$. However, the maximum likelihood estimator will not normally be the optimal one in the sense of having the smallest value of E_1.

The Bayesian solution of the problem that we are considering is seemingly simpler. As we have seen, in this problem the available relevant information consists of the observed value of *x*, the prior distribution of θ^*, and a conditional probability distribution $p(x|\theta^*)$ which is called the likelihood function. These and the Bayes theorem can be used for calculating a posterior distribution $p(\theta^*|x)$ for θ^*. With the help of this distribution one can then subsequently calculate the average loss (in terms of squared distance) of accepting the estimator $\tilde{\theta}$, which is

$$E_2 = \int \left(\tilde{\theta}(x) - \theta^*\right)^2 p(\theta^*|x) d\theta^* . \tag{4}$$

It is well known that in order to minimize E_2 one should choose the *mean of the posterior distribution* as the estimator. This is called the Bayes estimator.[4] In our simple example, the Bayes estimator is for every value of *x* given by the formula (3) with

$$\eta = (\sigma^2/n) \Big/ \left((\sigma^2/n) + s^2\right).$$

Note that the expected squared error of this estimator is smaller than that of the maximum likelihood estimator.

One can see how the two approaches have entirely different philosophical flavors. The Bayesian approach optimizes the estimator for the particular observed value of *x* and is unconcerned with the frequentist question about how the estimator would perform on the average for different values of *x*. However, it is fairly easy to see that the two approaches yield the same

---

[4] In nice symmetrical cases, the mean value of a distribution coincides with the point at which the posterior probability density is maximum. But it is important not to *define* the Bayes estimator in those terms because the point at which posterior probability density is maximum can be changed by a nonlinear transformation of the coordinates (see Forster [1995] for a detailed discussion of this point). On the other hand, the mean of a distribution is invariant under such transformations.

results: a given function $\tilde{\theta}(x)$ minimizes the quantity E_1, which should be minimized according to the frequentist approach, if for each x it minimizes the quantity E_2, which should be minimized according to the Bayesian approach.[5] One might wonder what the point of averaging over all possible values of the mean x of the data is; however, fortunately for the frequentists, such averaging is harmless in the sense that the same estimators will be optimal in both cases.

The idea of providing a frequentist explanation of the regression-to-the-mean by maximizing a population-level measure of success is not a standard part of the frequentist literature. In fact, we believe that the quick response to our proposal will be to say that it is just a minor re-description of the standard Bayesian solution. The reason given will be, we predict, that it makes essential use of the prior distribution $p(\theta^*)$, and this is the defining characteristic of Bayesianism. In our view, we are providing a *generalized* frequentist approach in which $p(\theta^*)$ is interpreted as representing the *frequencies* within a real or hypothetical *ensemble* of token cases, and this provides a subtly different foundation.[6] To argue our case, we need to spell out the idea with some care. In doing so, we shall also show how to remove the *ad hoc* features of hypothesis testing within a frequentist paradigm.

---

[5] This result is valid because

$$E_1 = \int \left[\int \left(\tilde{\theta}(x) - \theta^*\right)^2 p\left(x|\theta^*\right) dx \right] p(\theta^*) d\theta^* =$$
$$\int \left[\int \left(\tilde{\theta}(x) - \theta^*\right)^2 p\left(x|\theta^*\right) p(\theta^*) dx \right] d\theta^* =$$
$$\int \left[\int \left(\tilde{\theta}(x) - \theta^*\right)^2 p\left(\theta^*|x\right) p(x) dx \right] d\theta^* =$$
$$\int \left[\int \left(\tilde{\theta}(x) - \theta^*\right)^2 p\left(\theta^*|x\right) d\theta^* \right] p(x) dx = \int E_2 p(x) dx$$

Hence, the value of the integral E_1 receives its smallest value if and only if the estimator $\tilde{\theta}(x)$ has been chosen so that E_2 receives its smallest value for each x (or to be quite precise, for almost all values of x).

[6] Apparently, Wald originally conceived of his classic work in decision theory (Wald [1950]) as a frequentist decision theory, but was later persuaded that it should be given a Bayesian interpretation. This conclusion was, of course, correct given that Wald did not introduce the notion of an *ensemble* of token experiments which we introduce below.

4 Experiment types

The version of our framework, as we present it here, represents the observed data as being 'generated' by a probability distribution belonging to a finite set of possible probability distributions. What we call an *experiment type* is, essentially, an ordered collection of such possible distributions. The following definition is a slightly modified version of a definition in Blackwell ([1953], p. 265).

Definition 1: An *experiment type* is an ordered pair $\varepsilon = (\Lambda, \Omega)$ where $\Lambda = (\lambda_1, \lambda_2, ..., \lambda_n)$ is for some $n \in \mathbb{N}$ an *n*-tuple of probability measures on the same collection of subsets of the set Ω.

The space Ω is a set of possible outcomes of the experiment, where each outcome may describe a sequence of events. For example, the outcome in a coin flipping experiment might be a sequence of heads and tails. Each λ_i in Λ specifies a different probability distribution for these outcomes. Any particular (token) experiment is of type ε if and only if the class of possible outcomes is Ω and the physical process that produces the outcome is such that one of the probability measures in Λ is the true probability distribution, which generates the outcomes. A token experiment is simultaneously of many types. The reason why Λ has other components besides the actual one is that we intend to classify different token experiments together as tokens of the same type even when the generating distributions are different.

Just as in Blackwell's original definition, λ_i is a *probability measure*. This means that each λ_i assigns a probability value to each member of a collection of subsets of Ω. In the context of the two examples that we have already discussed, it is more customary to specify a probability distribution by a *probability density* directly over the elements of Ω, rather than with a probability *measure* over a collection of subsets. However, we conform to the practice of using measures to specify the probability distributions for reasons of generality: when the set of the outcomes of the experiment is taken to be an arbitrary set, the use of measures is more fundamental. Our assumption that Ω is finite in the coin-flipping example and the assumption that it is the set of the real numbers in our estimation example are special cases. We shall

explain below (in section 6) how one arrives at the more usual representation of probability distributions in terms of probability densities and likelihood functions from the measure-theoretic description.

As an example, the coin-flipping example of section 2 is an experiment of type $\varepsilon_1 = (\Lambda_1, \Omega_1)$, where $\Lambda_1 = (m_{\theta=1/3}, m_{\theta=3/4})$ are probability measures over the event space $\Omega_1 = \{\mathrm{H}, \mathrm{T}\}$ (the subscript '1' refers to the fact that the outcome of only a single coin toss is under consideration). The probability measures $m_{\theta=1/3}$ and $m_{\theta=3/4}$ satisfy the conditions $m_{\theta=1/3}(\{\mathrm{H}\}) = \frac{1}{3}$, $m_{\theta=1/3}(\{\mathrm{T}\}) = \frac{2}{3}$, $m_{\theta=1/3}(\{\mathrm{H},\mathrm{T}\}) = 1$, $m_{\theta=3/4}(\{\mathrm{H}\}) = \frac{3}{4}$, $m_{\theta=3/4}(\{\mathrm{T}\}) = \frac{1}{4}$, and $m_{\theta=3/4}(\{\mathrm{H},\mathrm{T}\}) = 1$. The probability of the set $\{\mathrm{H},\mathrm{T}\}$ is 1 because the set $\{\mathrm{H},\mathrm{T}\}$ represents the event of the coin landing either 'heads up' or 'tails up'.

In section 2 we considered a case in which a statistician observed an element of Ω_1 — that is, either the result H or the result T — and knew that this result was generated by one of the probability distributions in $\Lambda_1 = (m_{\theta=1/3}, m_{\theta=3/4})$, but did not know which one. The tests considered in section 2 were, essentially, procedures by which a statistician could choose an element of Λ_1 — i.e. either $m_{\theta=1/3}$ or $m_{\theta=3/4}$ — on the basis of an observed element of Ω_1.

The problem of estimating θ^*, as we described it in section 3, does not correspond to an experiment type in the sense of Definition 1 because the set of probability distributions is *infinite*. However, if we modify the problem by assuming that the actual value of θ^* has to be one of the finite set of numbers $\theta_1, \theta_2, ..., \theta_k$, then the modified version of the example corresponds to an experiment type $(\Lambda_\theta, \Omega_\theta)$, in which

$$\Lambda_\theta = \left(N(\theta_1, \sigma^2/n), N(\theta_2, \sigma^2/n), ..., N(\theta_k, \sigma^2/n)\right)$$

and $\Omega_\theta = \mathbb{R}$. Recall that in this example a statistician observed an element of $\mathbb{R}$, the average x, which is the value of a random variable whose probability distribution is of the form $N(\theta, \sigma^2/n)$ for some value of θ. The problem is to choose one of these probability distributions of the basis of an observation.

5 Payoffs and Decision Problems

We intend our frequentist framework to be more general than the standard framework in several respects. One of these has to do with the hypotheses that a statistician might consider or choose. It is often (as in *e. g.*, Blackwell [1953]) assumed that a statistician who is presented with an experiment of type (Λ, Ω) must select a probability distribution from Λ. E.g., in the context of the problem of estimating θ^* this means that the set $\{\theta_1, \theta_2, ..., \theta_k\}$ at the same time defines *both* the experiment type *and* the space of the hypotheses that the statistician might choose to consider. However, we wish to allow for the case in which the restriction in the possible values of θ^* is not known by the statistician. An obvious reason for considering such cases is that scientists often choose *idealized* hypotheses which they know to be strictly speaking false. Accordingly, we shall introduce in our framework a set *M* of *hypothetical probability distributions* on the set Ω which contains the probability measures that the statistician is willing to consider. These do not have to include the distributions in Λ.

We also want to take account of the fact that some false probability distributions can be better than others according to the cognitive aims of scientific theorizing. For this reason we shall introduce the notion of the *payoff* of accepting a probability distribution *m* when the correct probability distribution is λ. As the following definition states, payoffs are specified by a *payoff function* in our framework.

Definition 2: (Payoffs) Suppose that $\varepsilon = ((\lambda_1, \lambda_2, ..., \lambda_n), \Omega)$ is an experiment type and that *M* is a set of probability measures which have the same domain as the measures $\lambda_1, \lambda_2, ..., \lambda_n$. If *pay* is a real-valued function on $M \times \{\lambda_1, \lambda_2, ..., \lambda_n\}$, it is called a *payoff function*, and its value at the point (m, λ_i) (where $m \in M$ and $i \in \{1, ...n\}$) is denoted by $pay(m|\lambda_i)$ and called the payoff of *m* given λ_i. In this case the vector

$$pay(m) = (pay(m|\lambda_1), pay(m|\lambda_2), \ldots, pay(m|\lambda_n))$$

is called the *payoff vector* of the distribution *m*.

We emphasize that our notion 'payoff' is not meant to be restricted in any way. In many of the applications of decision theory to science, the payoff of a hypothesis refers to its epistemic or cognitive *value*. In such cases, we view the process of accepting a hypothesis on the basis of an observation as a kind of ampliative inference. A decision theory is, in part, a theory about the relative merits of such inferences even when they are considered in abstraction, without any reference to the aims of actual scientists or the practical consequences of accepting hypotheses.

Our next definition is a modified version of a definition in Blackwell ([1953], p. 265). It specifies how we represent the case in which the payoff of accepting a hypothesis depends only on its truth value. Within our framework this situation is represented with a payoff function which gives the value 1 to choosing the true distribution, and the value 0 to choosing any other distribution. We shall call such payoffs *simple* payoffs.[7]

Definition 3. (Simple payoffs) Suppose that $\varepsilon = \left(\left(\lambda_1, \lambda_2, ..., \lambda_n\right), \Omega\right)$ is an experiment type, and that *M* is a set of probability measures which have the same domain as the measures $\lambda_1, \lambda_2, ..., \lambda_n$. If a payoff function *pay* on $M \times \{\lambda_1, \lambda_2, ..., \lambda_n\}$ is such that, whenever $m \in M$ and $i \in \{1, ...n\}$, $pay\left(m \middle| \lambda_i\right) = 1$ when $m = \lambda_i$ and $pay\left(m \middle| \lambda_i\right) = 0$ when $m \neq \lambda_i$, we say that *pay* is a *simple* payoff function and that it defines *simple payoffs*.

Simple payoffs may appear to be *so* simple that they are of little interest. However, we shall see below that the Neyman-Pearson theorem, which says that best tests are likelihood ratio tests, is deduced from the assumption that the payoffs are simple. On the other hand, as we saw above, the standard frequentist theory of *estimation* does not utilize simple payoffs; rather, in it the payoff is defined in terms of the expectation value of the square of the difference between the estimated value and the true value of the parameter θ.

In addition to allowing for the case in which the distributions in *M* and the distributions in Λ are not identical, our framework is more general than its more traditional frequentist alternatives also in that we do not assume that the experimentally observed outcome was always

[7] The term is borrowed from Wald ([1950]).

simply an element of Ω. Rather, we wish to allow also for cases in which the probabilities are not only for the observed outcomes but also for some *potential* observations that have not yet occurred. For example, we want to be able to consider the result of a single coin toss, and make a choice between hypotheses that assign probabilities to future outcomes. For example, such hypotheses might be concerned with the probability distribution of a sequence of *three* coin tosses, of which only the first one has already taken place. In this case the considered probability distributions will be measures on the set

$$\Omega_3 = \{\mathrm{HHH}, \mathrm{HHT}, \mathrm{HTH}, \mathrm{HTT}, \mathrm{THH}, \mathrm{THT}, \mathrm{TTH}, \mathrm{TTT}\}.$$

However, the available information cannot be represented by an element of Ω_3 in this case. Rather, if the result of the first coin toss has been H, the statistician knows only that the element of Ω_3 will be one of the sequences in $\{\mathrm{HHH}, \mathrm{HHT}, \mathrm{HTH}, \mathrm{HTT}\}$. Similarly, if the result had been T, she will have to make a choice knowing only that one of the sequences in $\{\mathrm{THH}, \mathrm{THT}, \mathrm{TTH}, \mathrm{TTT}\}$ will be the actual one. In a case like this, the appropriate mathematical representation of the knowledge of the statistician is not an *element* of the set Ω, but *a non-empty subset* of that set. Clearly, the subsets that can represent such information will form a *partitioning* of Ω, since the intersection of any two such sets must be empty, and their union is Ω.

Accordingly, we shall represent the information that a statistician uses to choose a probability distribution as *an element of a partitioning X of* Ω. The case in which the statistician knows which element of Ω is the actual one is then represented by the situation in which the partitioning X is trivial in the sense that

$$X = \{\{\omega\} | \omega \in \Omega\}.$$

When this condition is valid, the sets X and Ω are essentially identical. We nevertheless introduce the distinction into the following definitions, because the distinction between what is observed and what is predicted is philosophically important (Forster and Sober [1994]), and because a major goal of this paper is to present a frequentist theory of decision-making in a *general* framework.

Definition 4. (Decision problem) The four-tuple $\mathbf{D} = (\varepsilon, M, X, pay)$ is called a *decision problem* if the following conditions (i)-(iv) are valid:

(i) $\varepsilon = ((\lambda_1, \lambda_2, ..., \lambda_n), \Omega)$ is an experiment type.

(ii) M is a set of probability measures which have the same domain as the measures $\lambda_1, \lambda_2, ..., \lambda_n$.

(iii) X is a partitioning of the set of **Ω.**

(iv) *pay* is a real-valued function on $M \times \{\lambda_1, \lambda_2, ..., \lambda_n\}$.

If $\mathbf{D} = (\varepsilon, M, X, pay)$ is a decision problem, the elements of M are called *its hypothetical probability distributions*.

In this definition the set M does not have to be identical with the set $\{\lambda_1, \lambda_2, ..., \lambda_n\}$ of the probability distributions of the experiment type ε. We shall refer to the special case in which the statistician knows that one of the probability distributions $\lambda_1, \lambda_2, ..., \lambda_n$ has generated the data, and restricts her attention to these distributions, as an *ideal* decision problem.

Definition 5. (Ideal decision problem) Suppose that $\varepsilon = ((\lambda_1, \lambda_2, ..., \lambda_n), \Omega)$ is an experiment type, and that $\mathbf{D} = (\varepsilon, M, X, pay)$ is a decision problem. We say that **D** is an *ideal decision problem* if $M = \{\lambda_1, \lambda_2, ..., \lambda_n\}$.

For example, the coin tossing example of section 2 corresponds to an ideal decision problem in the sense of this definition. More specifically, the experiment type $\varepsilon_1 = (\Lambda_1, \Omega_1)$ in which $\Lambda_1 = (m_{\theta=1/3}, m_{\theta=3/4})$ and $\Omega_1 = \{\mathrm{H}, \mathrm{T}\}$ can be embedded in the ideal decision problem $\mathbf{D}_{coin} = (\varepsilon_1, M_1, X_1, pay_{simple})$ in which $M_1 = \{m_{\theta=1/3}, m_{\theta=3/4}\}$, $X_1 = \{\{H\}, \{T\}\}$, and the payoff function pay_{simple} is the simple payoff function for which

$$pay_{simple}(m_{\theta=1/3} | m_{\theta=1/3}) = pay_{simple}(m_{\theta=3/4} | m_{\theta=3/4}) = 1 \text{ and}$$

$$pay_{simple}(m_{\theta=3/4} | m_{\theta=1/3}) = pay_{simple}(m_{\theta=1/3} | m_{\theta=3/4}) = 0.$$

Since what the statistician observes is, in this case, simply an element of Ω_1, the elements of the set Ω_1, H and T, and the elements of X_1, $\{\mathrm{H}\}$ and $\{\mathrm{T}\}$, correspond to each other in a one-to-one fashion.

Similarly, our second example corresponds to the experiment type $(\Lambda_\theta, \Omega_\theta)$ in which $\Lambda_\theta = \left(N(\theta_1, \sigma^2/n), N(\theta_2, \sigma^2/n), ..., N(\theta_k, \sigma^2/n)\right)$ and $\Omega_\theta = \mathbb{R}$, and it can be embedded in a decision problem $\mathbf{D}_\theta = (\varepsilon_\theta, M_\theta, X_\theta, pay_\theta)$ in which $M_\theta = \left\{N(\theta, \sigma^2/n) \middle| \theta \in \mathbb{R}\right\}$ is the set of all normal distributions[8] with variance σ^2/n and in which $X_\theta = \left\{\{x\} \middle| x \in \mathbb{R}\right\}$, is essentially identical with $\mathbb{R}$. If we stick to the idea that the success of an estimate of θ should be evaluated with its squared distance from the correct value θ^*, it is natural to define the payoff function of the decision problem $\mathbf{D}_\theta$ by the formula

$$pay_\theta\left(N(\theta, \sigma^2/n) \middle| N(\theta_j, \sigma^2/n)\right) = -(\theta - \theta_j)^2$$

where $\theta_j \in \{\theta_1, \theta_2, ..., \theta_k\}$ and $\theta \in \mathbb{R}$ are arbitrary.

As we stated above, Definition 4 is meant to apply to a case in which a statistician chooses an element of *M* on the basis of an observation represented by an element of *X*. *Decision functions* represent the rules by which such choices are made. However, before turning to a discussion of decision functions we shall still connect our measure theoretical representation of probability distributions with likelihood functions.

6 Likelihoods

It is easy to see how the likelihood function should be defined in the contexts of the two decision problems, $\mathbf{D}_{coin} = (\varepsilon_1, M_1, X_1, pay_{simple})$ and $\mathbf{D}_\theta = (\varepsilon_\theta, M_\theta, X_\theta, pay_\theta)$, which we used as

---

[8] To be quite precise, the notation $N(\theta, \sigma^2/n)$ refers to the representation of a normal distribution as a measure on $\mathbb{R}$, rather than as a density function on $\mathbb{R}$.

our examples. In the first example, and more generally in all cases in which the set X which represents the different possible observations is finite, the likelihood of a probability distribution λ relative to an element x of X is simply $\lambda(x)$, i.e. the probability that what is observed is x according to λ. In the second example the set of the different logically possible observations is isomorphic to the set of the real numbers and, hence, uncountably infinite. In this case the probability of each element of the set X is zero relative to the measures of the considered experiment type, and the above definition is useless. However, in such cases likelihoods can be represented in terms of *probability densities*. When X is isomorphic with $\mathbb{R}$, a probability density on X is a real-valued function p on X which specifies a probability density for the observed outcome. This probability density is such that the probability of the observed outcome belonging to the set C is

$$\int_C p(x)dx,$$

whenever C is an arbitrary measurable subset of X. In this case the value $p(x)$ is called the likelihood of the probability distribution p relative to x.

If our framework was supposed to be applicable to these two special cases only, we could rest content with these two standard definitions of 'likelihood'. However, because our framework is meant to be more general (*e.g.*, in section 11 we consider a case in which Ω is countably infinite), and since we allow the set Ω whose elements represent potential and actual observations to be an arbitrary set, which in general need not be finite or isomorphic with $\mathbb{R}$, we need to provide a more general definition of the notion of likelihood. This definition is presented as Definition 6:

Definition 6. Suppose that $\varepsilon = ((\lambda_1, \lambda_2, ..., \lambda_n), \Omega)$ is an experiment type, that $\mathbf{D} = (\varepsilon, M, X, pay)$ is a decision problem, and that μ is a measure on $\sigma_\varepsilon(X)$, where $\sigma_\varepsilon(X)$ is a collection of subsets of X called the σ-algebra of X. If $\lambda_i \in \{\lambda_1, ..., \lambda_n\}$ and if the non-negative function L_i on X is such that, for all sets C that belong to $\sigma_\varepsilon(X)$

$$\lambda_i^X(C) = \int_C L_i(x)d\mu(x),$$

we say that L_i is the *likelihood of* λ_i *relative to the measure* μ. Further, if the measures $\lambda_1, \lambda_2, ..., \lambda_n$ have the likelihood functions $L_1, L_2, ..., L_n$ relative to the same measure μ, we call the (n+1)-tuple $(\mu, L_1, L_2, ..., L_n)$ a *likelihood vector* of the decision problem **D**.

This definition contains two symbols which we have not defined, "$\sigma_\varepsilon(X)$" and "λ_i^X". Their definitions involve technical complications which are irrelevant for our current purposes and accordingly, we present their definition as Definitions A1 and A2 of a separate mathematical appendix. The appendix also contains the proofs of the theorems in this paper.

For our current purposes the essential feature of Definition 6 is that when the measure μ which specifies measures for the subsets of X is chosen suitably, the value $L_i(x)$ turns out to be what one normally means by the likelihood of the hypothesis λ_i relative to the observed outcome x. In our coin-tossing examples, as well as in all other cases in which X is finite, this will be the case if μ is chosen to be the *counting measure* on X. With this choice of μ the likelihood vector of *e.g.* our coin-tossing decision problem $\mathbf{D}_{coin} = (\varepsilon_1, M_1, X_1, pay_{simple})$, in which $\varepsilon_1 = (\Lambda_1, \Omega_1)$ and $\Lambda_1 = (m_{\theta=1/3}, m_{\theta=3/4})$, turns out to be $(\mu, L_{\theta=1/3}, L_{\theta=3/4})$, where the function $L_{\theta=1/3}$ gives the likelihoods of $\theta = 1/3$ when the outcomes are H and T, and the function $L_{\theta=3/4}$ gives the likelihoods of $\theta = 3/4$ when the outcomes are H and T. In other words, the functions $L_{\theta=1/3}$ and $L_{\theta=3/4}$ are such that $L_{\theta=1/3}(\{\mathrm{H}\}) = \frac{1}{3}$ and $L_{\theta=1/3}(\{T\}) = \frac{2}{3}$, and that $L_{\theta=3/4}(\{\mathrm{H}\}) = \frac{3}{4}$ and $L_{\theta=3/4}(\{T\}) = \frac{1}{4}$. Below we shall always take the measure μ to be the counting measure when X is finite.

Similarly, when X is isomorphic with the set of the real numbers, we shall always implicitly assume that the measure μ has been chosen to be the counterpart of the Lebesgue measure on X. In this case the likelihood functions will turn out to specify what one normally means by likelihood in these cases. For example, the likelihood functions of the distributions $N(\theta_i, \sigma^2/n)$ which Λ_θ contains will turn out to functions whose graph is the familiar bell-shaped curve.

It should be observed that we have not yet shown that all decision problems actually *have a likelihood vector* in the sense we use the term, and we also need to define the sense in which likelihood vectors are unique. These results are stated in theorems 1 and 2 below.

Theorem 1. All decision problems have likelihood vectors.

It is well known that when the evidence consists of the value of a continuous quantity, the numerical values of the likelihoods of the considered hypotheses can depend on the way the hypotheses are represented. For example, when the available data consists of the result of the measurement of the length of an object, and one considers a simple statistical hypothesis that specifies a probability distribution for its length, the numerical value of the likelihood of the hypothesis when lengths are measured in inches will be different from the likelihood that the hypothesis has when lengths are measured in centimeters. However, it is also well-known that the *likelihood ratio* of two simple statistical hypotheses will be independent of the choice of units, and of other similar choices of representation. Our Theorem 2 is the counterpart of this well-known result within our framework.

Theorem 2. Suppose that $\varepsilon = \left((\lambda_1, \lambda_2, ..., \lambda_n), \Omega \right)$ is an experiment type, and that $\mathbf{D} = (\varepsilon, M, X, pay)$ is a decision problem. If the measures λ_i and λ_j (where $i, j \in \{1, ..., n\}$) have the likelihood functions L_i and L_j relative to the measure μ, and the likelihood functions L_i' and L_j' relative to the measure μ', it must be the case that

$$L_i(x) / L_j(x) = L_i'(x) / L_j'(x)$$

with the possible exception of a subset of X which has zero measure with respect to both λ_i^X and λ_j^X. (Here the value of a ratio a/b counts as $+\infty$ if $a \neq 0$ and $b = 0$, and as undefined if $a = b = 0$.)

7 Decision Functions

Clearly, each rule for choosing a hypothetical probability distribution on the basis of the available observations can be represented by a mapping of elements of the set X of the possible

observed outcomes into the set M. We shall call such mappings *decision functions*. We could define a decision function to be an *arbitrary* function from X to M if we were concerned only with cases in which either the set M or the set X is finite. However, when neither of the sets is finite, it is convenient to restrict the set of acceptable decision functions by introducing two rather weak regularity conditions, which we shall call the DF-conditions. The precise statement of these conditions is a part of Definition A3 in the mathematical appendix. Those decision functions that conform to the following definition are called *ordinary*, in order to distinguish them from *randomized* decision functions, which we shall discuss at the end of this section.

Definition 7. (Ordinary decision function) Suppose that $\varepsilon = \left(\left(\lambda_1, \lambda_2, ..., \lambda_n\right), \Omega\right)$ is an experiment type, and that $\mathbf{D} = \left(\varepsilon, M, X, pay\right)$ is a decision problem. A function f is called a *decision function*, or an *ordinary decision function*, associated with **D** if f is a function which maps X into M, and if f satisfies the DF-conditions. The set of all decision functions associated with **D** will be denoted by $F\left(X, M; \varepsilon, pay\right)$ or, when it is clear from the context what the considered experiment and payoff function are, simply by $F\left(X, M\right)$.

For example, the decision functions which are associated with $\mathbf{D} = \left(\varepsilon_1, M_1, X_1, pay_{simple}\right)$ are functions from the set $X_1 = \left\{\left\{\mathrm{H}\right\}, \left\{\mathrm{T}\right\}\right\}$ into the set $M_1 = \left\{m_{\theta=1/3}, m_{\theta=3/4}\right\}$. There are four ordinary decision functions in this case , and the DF-conditions are trivially valid for all of them. These functions correspond in an obvious way to the four tests T_1, T_2, T_3 and T_4 that we considered in section 2 and accordingly, we shall denote them by f_1, f_2, f_3 and f_4, respectively. Hence, in this case $F\left(X_1, M_1; \varepsilon_1, pay_{simple}\right) = \left\{f_1, f_2, f_3, f_4\right\}$. The function f_1 maps both elements of X_1 to the distribution determined by $\theta = 1/3$, and the function f_2 maps them both to the distribution corresponding to $\theta = 3/4$. The function f_3 is the counterpart of the "unreasonable" test T_3: it satisfies the conditions $f_3\left(\left\{\mathrm{H}\right\}\right) = m_{\theta=1/3}$ and $f_3\left(\left\{\mathrm{T}\right\}\right) = m_{\theta=3/4}$. The remaining function f_4 is the counterpart of the "reasonable" test T_4, and it satisfies the conditions $f_4\left(\left\{\mathrm{H}\right\}\right) = m_{\theta=3/4}$ and $f_4\left(\left\{\mathrm{T}\right\}\right) = m_{\theta=1/3}$.

The following definition defines the *average success* of a decision function *f* relative to a probability distribution λ — or, more rigorously, the expected payoff of a function *f*, given a distribution λ — for all legitimate decision functions.

Definition 8. (Expected payoff) Suppose that $\varepsilon = \left(\left(\lambda_1, \lambda_2, ..., \lambda_n\right), \Omega\right)$ is an experiment type, that $\mathbf{D} = \left(\varepsilon, M, X, pay\right)$ is a decision problem, and that $\left(\mu, L_1, L_2, ..., L_n\right)$ is a likelihood vector of **D**. Whenever $\lambda_i \in \left\{\lambda_1, \lambda_2, ..., \lambda_n\right\}$ and f is a decision function which belongs to $F\left(X, M; \varepsilon, pay\right)$, the quantity $pay\left(f \middle| \lambda\right)$ is defined with the formula

$$pay\left(f \middle| \lambda_i\right) = \int_X pay\left(f\left(x\right) \middle| \lambda_i\right) L_i\left(x\right) d\mu\left(x\right)$$

and called which we shall call *the expected payoff of f given* λ. Further, in this case the vector $\left(pay\left(f \middle| \lambda_1\right), pay\left(f \middle| \lambda_2\right), ..., pay\left(f \middle| \lambda_n\right)\right)$ is called the *payoff vector of f*.

It is straightforward to show that the integral in this definition exists whenever *f* satisfies the DF-conditions, and that its value is independent of the choice of the likelihood vector $\left(\mu, L_1, L_2, ..., L_n\right)$.

Note that the likelihood values in the definition of the expected payoff of a decision rule refer to the generating distributions in Λ, as specified by the experiment type under consideration. This means that the payoff of a decision rule is an *objective* quantity from a world-centric point of view. In the person-centric view, the expected payoff would be defined in terms of the likelihoods of the hypothetical distributions. This is why our framework is frequentist rather than Bayesian.

When the set *X* is isomorphic with $\mathbb{R}$ and the measure μ is chosen to be the counterpart of the Lebesgue measure, the expression of $pay\left(f \middle| \lambda_i\right)$ turns into

$$pay\left(f \middle| \lambda_i\right) = \int_X pay\left(f\left(x\right) \middle| \lambda_i\right) L_i\left(x\right) dx$$

Similarly, when the set *X* is finite and the measure μ is chosen to be the counting measure, Definition 8 implies that

(5) $$pay(f|\lambda_i) = \sum_{x \in X} pay(f(x)|\lambda_i)\lambda_i(x)$$

for each $\lambda_i \in \{\lambda_1, \lambda_2, ..., \lambda_n\}$. It is easy to see the intuitive significance of each of these formulas: in each case $pay(f|\lambda_i)$ is the average payoff of the function *f* when the actual probability distribution in Ω is given by λ_i.

In our coin-tossing example the function f_3 was a mathematical representation of the procedure of choosing $m_{\theta=1/3}$ when H is observed and $m_{\theta=3/4}$ when T is observed. If we follow this procedure, and if the actual distribution happens to be $m_{\theta=1/3}$, we shall have a chance of 1/3 of picking up a distribution with the payoff 1, and a chance of 2/3 of picking up a distribution with the payoff 0. Hence, in this case the expectation value of the payoff of this procedure, $pay_{simple}(f_3 | m_{\theta=1/3})$, is $\frac{1}{3} \times 1 + \frac{2}{3} \times 0 = \frac{1}{3}$.

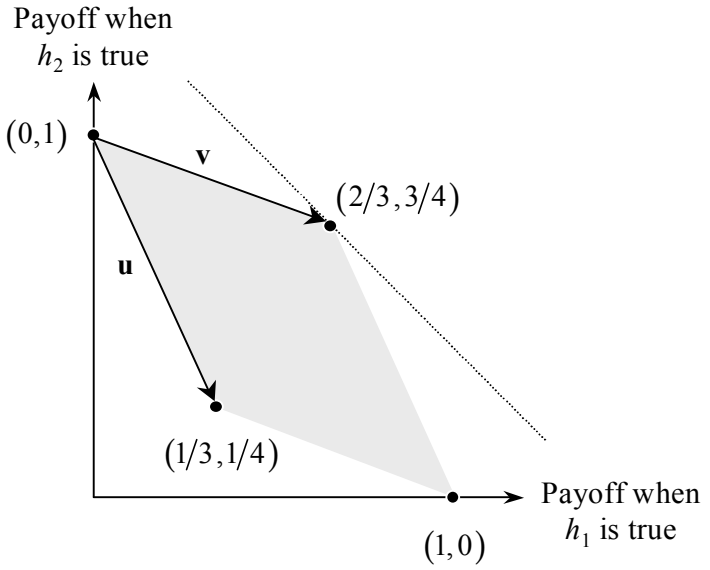

Figure 1: The payoff vectors of the decision problem $\mathbf{D} = (\varepsilon_1, Q_1, X_1, pay_{simple})$. Here h_1 is the hypothesis that the actual distribution is $q_{\theta=1/3}$, and h_2 is the hypothesis that it is $q_{\theta=3/4}$. The four points (1,0), (0,1), (1/3,1/4), and (2/3,3/4) represent the payoff vectors of the four decision functions f_1, f_2, f_3 and f_4, respectively.

Similarly, it is easy to see that expectation value of the payoff of using f_3 when the correct distribution is $m_{\theta=3/4}$ is $\frac{3}{4} \times 0 + \frac{1}{4} \times 1 = \frac{1}{4}$ and that, accordingly, $pay_{simple}(f_3 | m_{\theta=3/4}) = \frac{1}{4}$. Hence, the payoff vector of the function f_3 is (1/3, 1/4).

It is also easy to see that the payoff vector of the function f_4 is (2/3, 3/4), and that the two "*a priori* functions" f_1 and f_2 have the payoff vectors (1,0) and (0,1), respectively. These payoff vectors are shown in Figure 1. The poverty of the function f_3 is reflected in the position of its

payoff vector (1/3, 1/4) in this figure; as the figure shows, it performs worse than the function f_4 *both* when $\theta = 1/3$ *and* when $\theta = 3/4$.

We have not introduced any rigorous mathematical representation of *randomized decision procedures*. In the context of our current example, a randomized decision procedure can be characterized by specifying the values of two quantities: the probability ρ_{H} with which $\theta = 1/3$ gets chosen when the result of the coin toss is H, and the probability ρ_{T} with which this distribution gets chosen when the result of the coin toss is T. The probabilities with which $\theta = 3/4$ gets chosen in the two cases are then $1-\rho_{\mathrm{H}}$ and $1-\rho_{\mathrm{T}}$, respectively.

What are the payoff vectors of such randomized decision procedures? When the value of θ is actually 1/3, the probability of choosing the actual distribution, which equals the expected payoff when payoffs are simple, is

$$\begin{gathered}(\text{probability of H})(\text{probability of choosing } \theta = 1/3 \text{ when H has been observed}) + \\ (\text{probability of T})(\text{probability of choosing } \theta = 1/3 \text{ when T has been observed}) = \\ (1/3)\rho_{\mathrm{H}} + (2/3)\rho_{\mathrm{T}}\end{gathered}$$

On the other hand, when the truth is that $\theta = 3/4$, the probability of choosing the correct hypothesis is $(3/4)(1-\rho_{\mathrm{H}}) + (1/4)(1-\rho_{\mathrm{T}})$. Hence, the randomized decision procedure which we are considering corresponds to the payoff vector

$$\left((1/3)\rho_{\mathrm{H}} + (2/3)\rho_{\mathrm{T}}, (3/4)(1-\rho_{\mathrm{H}}) + (1/4)(1-\rho_{\mathrm{T}})\right) = (0,1) + \rho_{\mathrm{H}}\mathbf{u} + \rho_{\mathrm{T}}\mathbf{v}$$

where the vector $\mathbf{u} = (1/3, -3/4)$ and the vector $\mathbf{v} = (2/3, -1/4)$. These vectors have also been shown in Figure 1. Since all payoff vectors are of the form $(0,1) + \rho_{\mathrm{H}}\mathbf{u} + \rho_{\mathrm{T}}\mathbf{v}$, and since the possible values of ρ_{H} and ρ_{T} range from 0 to 1, the range of the payoff vectors of randomized decision procedures is represented by the shaded area in this figure.

8 Ensembles of Token Experiments and Optimality

The Neyman-Pearson paradigm, which states that one should make use of a best test, determines a unique best test only after the size of the test is fixed by convention. In Bayesian

statistics one can avoid such conventional choices, but they can be avoided only by introducing subjective prior probabilities for the considered hypotheses. Our aim is to develop a new version of the frequentist paradigm, in which the role of such conventional and subjective features is smaller than it is in its two traditional alternatives. A novel feature of our approach is to consider each experiment as an element of an *ensemble of token experiments*.

Each ensemble of token experiments contains experiments that are of the same type in the sense of Definition 1: in these experiments the outcome space Ω and the set of probability distributions $\{\lambda_1, \lambda_2, ..., \lambda_n\}$ that generate an outcome are the same. Hence, each of these experiments is represented by the same mathematical entity, $((\lambda_1, \lambda_2, ..., \lambda_n), \Omega)$, and the experiments are similar also in so far that in each of them a statistician observes either an element of Ω or a part of it, and then chooses a probability distribution on the space Ω on the basis of this observation (although the chosen distribution on Ω need not be one of the λ_i).

A single token experiment may, of course, be a member of many different ensembles. The *optimality* of a decision function within a particular ensemble of token experiments is a useful theoretical device which helps us to understand and discuss the relative merits of decision functions. Such optimality turns out to have an obvious definition if it is assumed that each of the considered distributions λ_i occurs within each ensemble of token experiments with some well defined frequency p_i. Accordingly, our mathematical representation of ensembles of token experiments specifies in addition to the relevant experiment type also the values of such frequencies.

Definition 9. (Ensemble) The pair $S = (\varepsilon, \mathbf{p})$ is called *an ensemble of token experiments* if (i) $\varepsilon = ((\lambda_1, \lambda_2, \ldots, \lambda_n), \Omega)$ is an experiment type and (ii) $\mathbf{p} = (p_1, p_2, \ldots, p_n)$ is an n-tuple of non-negative real numbers which satisfies the condition $p_1 + p_2 + \ldots + p_n = 1$.

It is our intention that, depending on the application, ensembles may be real or imaginary sets of token experiments. Although the epistemological import of each case is different, we shall not attend to this important difference at the present time. Our aim is only to define optimality in an

objective way. So, in either case, if $S = (\varepsilon, \mathbf{p})$ is an ensemble containing N token experiments, then λ_i is the true generating probability distribution for $p_i \times N$ experiments in the ensemble.

Decision functions map each element of a partitioning X of Ω to an element of the set M of hypothetical distributions on Ω. The payoff of a decision function f relative to a distribution λ on Ω, $pay(f \mid \lambda)$, was above defined to be the expected payoff that an application of the function yields when the distribution λ is the actual one. When an experiment is viewed as an element of an ensemble of token experiments, an obvious measure for the success of a decision function within the ensemble is the average value of $pay(f \mid \lambda)$ within it. This is given by

$$p_1 pay(f \mid \lambda_1) + p_2 pay(f \mid \lambda_2) + \cdots + p_n pay(f \mid \lambda_n)$$

Accordingly, we shall define a decision function to be *optimal* if it maximizes the value of this quantity.

Definition 10. (Optimality) Suppose that $\varepsilon = ((\lambda_1, \lambda_2, ..., \lambda_n), \Omega)$ is an experiment type, that $\mathbf{D} = (\varepsilon, M, X, pay)$ is a decision problem, and that $S = (\varepsilon, (p_1, p_2, \ldots, p_n))$ is an ensemble of token experiments. If f belongs to $F(X, M; \varepsilon, pay)$, the quantity $pay_S(f)$ is defined by the formula

$$pay_S(f) = \sum_{i=1}^{n} p_i pay(f \mid \lambda_i),$$

and it is called *the expected payoff of f within S*. If f^* belongs to $F(X, M; \varepsilon, pay)$, and if the quantity $pay_S(f)$ receives its maximum value within $F(X, M; \varepsilon, pay)$ when $f = f^*$, we say that the decision function f^* is *optimal* for S.

This concept of optimality defines an *objective* sense in which one decision function is better than another relative to a given ensemble of token experiments. In section 3 we saw how one could solve the problem of regression to the mean within the frequentist framework by introducing a prior distribution $p(\theta^*)$ for the value of θ^*, and by requiring that the estimator $\tilde{\theta}$ that one uses minimizes the integral E_1, which depends on the distribution $p(\theta^*)$ (see

formula (2)). Clearly, in the context of this example the recommendation that one should use an optimal decision function is simply a discrete version of the idea the integral E_1 should be minimized: if one replaces in formula (2) the continuous probability distribution $p(\theta^*)$ with a discrete probability distribution that gives probabilities $p_1, p_2, \ldots, p_n$ to the parameter values $\theta_1, \theta_2, \ldots, \theta_n$, respectively, then the integral E_1 turns into the sum

$$S_1 = \sum_{i=1}^{n} p_i \left[\int \left(\tilde{\theta}(x) - \theta_i\right)^2 p\left(x|\theta_i\right) dx \right] = -\sum_{i=1}^{n} p_i \left[\int -\left(\tilde{\theta}(x) - \theta_i\right)^2 p\left(x|\theta_i\right) dx \right]$$

However, earlier we defined the payoff function of the decision problem $\mathbf{D}_\theta = (\varepsilon_\theta, M_\theta, X_\theta, pay_\theta)$ to be

$$pay_\theta\left(N\left(\theta, \sigma^2/n\right) \middle| N\left(\theta_j, \sigma^2/n\right)\right) = -\left(\theta - \theta_j\right)^2,$$

and when payoffs are defined in this way, the expression in square brackets in the formula of S_1 equals the expected payoff, given the parameter value θ_i, of the decision function that chooses the hypothesis $\tilde{\theta}(x)$ for an observed x value. Hence, S_1 is the negative of the quantity maximized by optimal decision functions, and the practice of choosing the estimator that minimizes the sum S_1 is essentially identical to choosing an optimal decision function.

On the other hand, our earlier example of the decision problem $\mathbf{D}_{coin} = (\varepsilon_1, M_1, X_1, pay_{simple})$ corresponded to a situation in which a coin was chosen at random from two coins described by the Bernoulli parameter values $\theta = 1/3$ and $\theta = 3/4$. It is natural to embed an experiment of type ε_1 into an ensemble of token experiments $(\varepsilon_1, \mathbf{p})$ in which $\mathbf{p} = (p_1, p_2) = (0.5, 0.5)$. This is because each of the distributions $m_{\theta=1/3}$ and $m_{\theta=3/4}$ would turn out to be the correct one in an approximately half of the cases when the experiment is repeated.

As we explained in section 7, the points of the shaded area in Figure 1 correspond to the payoff vectors that decision functions and randomized decision procedures can have. On the other hand, for each fixed value of a number *C*, the set of payoff vectors of the set *f* such that $pay_S(f) = C$ is represented by a straight line. The slopes of such straight lines will depend on the numerical values that are given to p_1 and p_2, but it will always be the case that, the higher

the value of $pay_S(f)$ on such a straight line, the higher and more to the right that straight line will be located. In particular, the decision procedure which is optimal for the given values of p_1 and p_2 is located at the point at which the highest straight line of the corresponding slope which touches the shaded area touches it. The dotted line drawn in Figure 1 corresponds to the values $p_1 = p_2 = 0.5$, and it touches the shaded area at the point (2/3, 3/4). Hence, the unique optimal decision function relative to an ensemble for which $p_1 = p_2 = 0.5$ is f_4, because it is the only rule corresponding to the payoff vector (2/3, 3/4). In fact, it will be the optimal rule for a variety of ensembles with different values of p_1 and p_2. It is only when p_1 is substantially greater than p_2, or vice versa, that one of the 'a priori' rules will be optimal. The rule f_3 is never optimal.

In a similar manner, one can also see that *a randomized decision procedure cannot be the only optimal one in the situation of Figure 1:* since the top-most straight line of a given slope which meets the shaded area must meet it at one of the points (0, 1), (2/3, 3/4), and (1, 0), one of the ordinary decision functions f_1, f_2, and f_4 has to be an optimal one. This geometric argument applies only to the case in which the set *M* contains only two probability distributions. However, in the next section we shall see that the result is valid also more generally: when a decision problem **D** has been fixed, and when its experiment has been embedded in an ensemble of token experiments *S*, there will always be ordinary (and not just randomized) decision functions which are optimal for **D** and *S*. In this respect our notion of optimality differs from the notion of being a best test since, as we saw above, it might turn out that all best tests of the given size are randomized tests (or even that all tests of the given size are randomized tests).

9 Sufficient and Necessary Conditions for Optimality

A combination of the definition of the expected payoff $pay_S(f)$ of a decision function *f* within an ensemble *S* and the definition of the expected payoff of *f* relative to a distribution λ_i yields an explicit formula for $pay_S(f)$. We shall present this formula as our next theorem.

Theorem 3. If $\varepsilon = \left(\left(\lambda_1, \lambda_2, ..., \lambda_n\right), \Omega\right)$ is an experiment type, $\mathbf{D} = \left(\varepsilon, M, X, pay\right)$ is a decision problem, and $\left(\mu, L_1, L_2, ..., L_n\right)$ is a likelihood vector of **D**, the quantity $pay_S\left(f\right)$ is given by the formula

$$pay_S\left(f\right) = \int_X pay_{f,S}\left(x\right) dx\,,$$

where

$$pay_{f,S}\left(x\right) = \sum_{i=1}^{n} p_i pay\left(f\left(x\right)\middle|\lambda_i\right) L_i\left(x\right)$$

By definition, an optimal decision function is a decision function for which $pay_S\left(f\right)$ receives its largest possible value. Since according to Theorem 3 the value of $pay_S\left(f\right)$ is the integral of the function $pay_{f,S}\left(x\right)$ over the space *X*, the theorem implies that the value of $pay_S\left(f\right)$ will be maximized by the decision function *f* for which $pay_{f,S}\left(x\right)$ receives its largest possible value for each *x*. This observation constitutes our main theorem.

Theorem 4 (Main Theorem). Suppose that $\varepsilon = \left(\left(\lambda_1, \lambda_2, ..., \lambda_n\right), \Omega\right)$ is an experiment type, that $\mathbf{D} = \left(\varepsilon, M, X, pay\right)$ is a decision problem, and that $\left(\mu, L_1, L_2, ..., L_n\right)$ a likelihood vector of the decision problem **D**. If the decision function $f \in F\left(X, M; \varepsilon, pay\right)$ is such that it chooses for each *x* a measure $f\left(x\right) \in M$ for which the quantity

$$Q = \sum_{i=1}^{n} p_i pay\left(f\left(x\right)\middle|\lambda_i\right) L_i\left(x\right)$$

is largest among the measures of *M*, the decision function *f* is optimal for *S*.

In other words, the problem of choosing an appropriate distribution in response to the empirical information *x* can be solved separately for each *x*, so that the "global" problem of maximizing $pay_S\left(f\right)$ can be achieved by the separate "local" maximizations of $pay_{f,S}\left(x\right)$.

We have not introduced a rigorous definition for randomized decision procedures into our framework. A straightforward generalization of our Main Theorem justifies this omission by

showing that if we did introduce such a rigorous definition, and if we defined the function pay_S also for randomized decision functions, it could not happen that $pay_S(f)$ received its largest value only for randomized decision functions. In other words, it is impossible that some randomized decision functions are optimal while there are no ordinary decision functions that are optimal. In order to see why this cannot be the case, consider an arbitrary randomized decision function *g* which for each *x* chooses one of the distributions in $f_1(x), f_2(x), ..., f_k(x)$ at random. Since the probabilities by which the measures $f_1(x), f_2(x), ..., f_k(x)$ get chosen in a randomized decision procedure may depend on the observed *x*, these probabilities must be represented as functions of *x*. We shall denote these functions by $\rho_1(x), \rho_2(x), ..., \rho_k(x)$, respectively. These functions must, of course, satisfy the condition $\rho_1(x) + \rho_2(x) + ... + \rho_k(x) = 1$ for each *x*.

It is clear that the expected payoff of the randomized decision function *g* when the actual distribution is λ_i can be defined with the formula

$$pay(g|\lambda_i) = \sum_{j=1}^{k} \int_X \rho_j(x)\, pay(f_j(x)|\lambda_i) L_i(x)\, dx$$

and its expected payoff within an ensemble *S* can be defined just like we have defined the expected payoff of an ordinary decision function, with the formula

$$pay_S(g) = \sum_{i=1}^{n} p_i\, pay(g|\lambda_i)$$

Now a straightforward generalization of Theorem 3 yields the result that $pay_S(g)$ can be put into the form

$$pay_S(g) = \int_X pay_{g,S}(x)\, d\mu(x),$$

where $$pay_{g,S}(x) = \sum_{i=1}^{n} \sum_{j=1}^{k} \rho_j(x)\, p_i\, pay(f_j(x)|\lambda_i) L_i(x),$$

and a straightforward generalization of Theorem 4 states that a randomized decision function *g* is optimal if it is such that the quantity $pay_{g,S}(x)$ receives for each *x* its largest possible value within the class of all randomized and ordinary decision functions.

However, this quantity equals

$$\sum_{j=1}^{k} \rho_j(x) \sum_{i=1}^{n} p_i pay\left(f_j(x) \middle| \lambda_i\right) L_i(x)$$

and this means that it can be maximized by choosing for each *x* the value *j** of *j* for which

$$\sum_{i=1}^{n} p_i pay\left(f_j(x) \middle| \lambda_i\right) L_i(x)$$

receives its largest value, and by setting $\rho_{j*}(x) = 1$ and $\rho_j(x) = 0$ for all $j \neq j*$. However, this choice of the functions $\rho_1(x), \rho_2(x), ..., \rho_k(x)$ represents the *ordinary* decision function which corresponds to choosing $f_{j*}(x)$ for each *x*. Hence, the value of the quantity that optimal decision functions maximize gets maximized by an ordinary decision function, and it cannot be the case that the set of all optimal decision functions contains only randomized decision functions.

As the following corollary shows, the methodological recommendation of our main theorem is simple when the considered decision problem is an *ideal* decision problem with *simple* payoffs.

Corollary. Suppose that $\varepsilon = \left((\lambda_1, \lambda_2, ..., \lambda_n), \Omega\right)$ is an experiment type, that $\mathbf{D} = \left(\varepsilon, M, X, pay_{simple}\right)$ is an ideal decision problem with simple payoffs, and that $(\mu, L_1, L_2, ..., L_n)$ a likelihood vector of **D**. If a decision function $f \in F\left(X, M; \varepsilon, pay_{simple}\right)$ is such that, for each *x* in *X*, $f(x)$ is a measure $\lambda_i \in \{\lambda_1, \lambda_2, ..., \lambda_n\}$ for which $p_i L_i(x)$ receives its largest value, then *f* is optimal for *S*.

Our main theorem and its corollary will look familiar when they are given a Bayesian interpretation. Under this interpretation each p_i, $i = 1, \ldots n$, is the prior probability $\Pr(h_i)$, where h_i is denotes the hypothesis that the actual distribution of the considered random variable is given by λ_i, and $\Pr(\cdot)$ denotes the prior probability of a hypothesis. Since the quantity $L_i(x)$ is

the likelihood of h_i relative to x, according to Bayes's theorem the product $p_i L_i(x)$ is proportional to the posterior probability of h_i given x, $\Pr(h_i|x)$. Under this Bayesian interpretation the quantity $pay_{f,S}(x)$ of Theorem 3 is proportional to

$$\sum_{i=1}^{n} \Pr(h_i|x)\, pay(f(x)|\lambda_i). \tag{6}$$

This is the expected utility of accepting the hypothesis recommended by the decision procedure f, where the expectation is calculated according to the posterior distribution of the hypotheses $h_1, h_2, ..., h_n$, given the observed outcome x. In particular, in the special case in which the payoffs are simple the optimal decision function is the one which chooses the hypothesis with the highest posterior probability.

This Bayesian interpretation is, of course, not the one we are giving to the quantities that occur in the Main Theorem and its corollary. Rather, as we have seen, Theorem 4 is concerned with the optimality of decision functions within an ensemble of token experiments, and the values p_i are the relative frequencies with which the members of Λ occur as the generating distributions within it. A particular token experiment may be viewed as belonging to many different ensembles of token experiments and, unlike in Bayesian statistics, the values of p_i (i = 1,…, n) are functions of the considered ensemble.

Moreover, the 'Bayesian' interpretation, as we describe it, is a world-centric formula in which the hypotheses h_i refer to the distributions of the set of generating distributions Λ, rather than the hypothetical distributions in M. In contrast, the usual Bayesian formula for maximizing expected utility is a person-centric formula (which appeals to a subjective notion of optimality). When the decision problem is ideal (see Definition 5), the two points of view are equivalent. But when a decision problem is not ideal, the Main Theorem is incorrectly interpreted by the person-centric Bayesian formula.

10 Optimality, Best Tests, and Likelihood Ratio Tests

We have already pointed out that one can use our definition of an optimal decision function in an ensemble for solving the problem of the regression to the mean within the frequentist framework. In this section we take a closer look at the relationship between our framework and Neyman-Pearson hypothesis testing, which we introduced in section 2.

In section 2 we focused most of our attention on the case in which a choice was made between just two probability distributions, and most of this section will be concerned with the same example. We begin by reformulating the relevant notions of the Neyman-Pearson theory of tests within our framework.

Definition 11. Suppose that $\varepsilon = ((\lambda_1, \lambda_2), \Omega)$ is an experiment type with two probability distributions, that $\mathbf{D} = (\varepsilon, M, X, pay)$ is an ideal decision problem, and that (μ, L_1, L_2) is a likelihood vector of **D**. In this case the elements of $F(X, M; \varepsilon, pay)$ are called *tests of* λ_1 *against* λ_2, and if *f* belongs to $F(X, M; \varepsilon, pay)$, the set $C = \{x \in X \mid f(x) = \lambda_2\}$ is called the *critical region* of the test *f*. Further, in this case the quantity $\int_C L_1(x)\, d\mu(x)$ is called the *size of the test f*, and the quantity $\int_C L_2(x)\, d\mu(x)$ is called the *power of the test f*. If the power of *f* is maximal among all those tests of λ_1 against λ_2 that have the same size, then *f* is called a *best test of* λ_1 *against* λ_2 *of that size*, or simply a *best test of* λ_1 *against* λ_2.

The function *f* mentioned in this definition has, of course, been meant to correspond in our framework to a test *whose null hypothesis states that the correct distribution is* λ_1. It is easy to see that the size and the power which get defined in this definition are, indeed, what one normally means by the size and the power of such a test.

The practice of calling the test which has the maximal power among all the tests of its size a "best test" is motivated by the ideas that *1)* one of the considered probability distributions is, as a matter of fact, generating the data, and that *2)* the aim of the statistician who performs the test is

to find out which distribution is the actual one. The first of these assumptions corresponds within our framework to the assumption that the considered decision problem is ideal, and Definition 11 is applicable only to cases in which this assumption is valid. However, this definition does not mention any counterpart of assumption *2)*. We shall next produce such a counterpart.

We have already seen that decision problems with *simple payoffs* can be used for representing situations in which the statistician has the aim of finding the actual distribution, and no other aims beside it. However, a simple payoff function represents a situation in which a statistician views all false probability distributions as equally "bad". Since statisticians actually often find some false probability distributions preferable to others, the claim that payoffs are simple is not an appropriate rigorous counterpart of the assumption *2)*. Rather, such a counterpart should state that, whatever the actual probability distribution should happen to be, choosing it is preferable to choosing any other probability distribution. The situations of this kind are within our framework represented by decision problems which have *strictly epistemic* payoffs in the sense of the following definition.

Definition 12. (Epistemic payoffs) Suppose $\varepsilon = \left(\left(\lambda_1, \lambda_2, ..., \lambda_n\right), \Omega\right)$ is an experiment type, that *M* is a set of probability distributions on Ω, and that $\left\{\lambda_1, \lambda_2, ..., \lambda_n\right\} \subseteq M$. The payoff function *pay* on $M \times \left\{\lambda_1, \lambda_2, ..., \lambda_n\right\}$ is called *epistemic* if for all $m \in M$ and all $\lambda \in \left\{\lambda_1, \lambda_2, ..., \lambda_n\right\}$ $pay\left(\lambda \middle| \lambda\right) \geq pay\left(m \middle| \lambda\right)$, and *strictly epistemic* if $pay\left(\lambda \middle| \lambda\right) > pay\left(m \middle| \lambda\right)$ whenever $m \neq \lambda$.

The decision problems that are ideal in the sense of Definition 5, and have strictly epistemic payoffs in the sense of Definition 12, constitute an appropriate rigorous counterpart for the class of cases to which Neyman-Pearson theory of tests was originally supposed to cover. It is easy to see that in the context of such decision problems all optimal decision functions are best tests (although the converse does not hold).

Theorem 5. Suppose that $\varepsilon = \left(\left(\lambda_1, \lambda_2\right), \Omega\right)$ is an experiment type with two probability distributions, that $\mathbf{D} = \left(\varepsilon, M, X, pay\right)$ is an ideal decision problem, and that *pay* is a strictly epistemic payoff function. If $S = \left(\varepsilon, \mathbf{p}\right)$ is an ensemble of token experiments in which

$\mathbf{p} \neq (1,0)$, and if the decision function $f^* \in F(X, M; \varepsilon, pay)$ is optimal within *S*, then *f** is a best test of λ_1 against λ_2.

According to the Neyman-Pearson theorem *likelihood ratio tests* between two simple hypotheses are always best tests. A likelihood ratio test is a test in which one sets a critical value *K* for the ratio of the likelihoods of the null hypothesis and its alternative, and chooses the null hypothesis whenever the ratio is larger than *K*, and its alternative whenever the ratio is smaller than *K*. On the other hand, in section 2 above we saw that if one is not willing to consider tests that involve randomized decision making, the best available tests of a given size might fail to be likelihood ratio tests. There we considered an example in which a choice was made between the hypotheses $\theta = 1/3$ or $\theta = 3/4$, and there were only four non-randomized tests. One of these, T_3, was quite unintuitive in the sense that in this test one chose the hypothesis that "went against the evidence": if the observed outcome of the coin toss was H, one chose $\theta = 1/3$, and if the result was T, one chose $\theta = 3/4$. In the case of this procedure the likelihood ratio, i.e.

$$\frac{\text{the likelihood of } \theta = 1/3}{\text{the likelihood of } \theta = 3/4},$$

was smaller when the null hypothesis $\theta = 1/3$ was chosen than when its alternative $\theta = 3/4$ was chosen. (In the former case it is 4/9, and in the latter case it is 8/3.) Hence, the test T_3 is not a likelihood ratio test. Yet the test T_3 is the only test of its size, and therefore the best test of its size, if randomized tests are not taken into consideration.

This makes it natural to ask how our notion of optimality is related to the notion of a likelihood ratio test. This question will be answered by Theorem 6 below. The theorem makes use of a rigorous definition of a likelihood ratio test, which we present as our Definition 13.

Definition 13. Suppose that $\varepsilon = ((\lambda_1, \lambda_2), \Omega)$ is an experiment type with two probability distributions, that $\mathbf{D} = (\varepsilon, M, X, pay)$ is an ideal decision problem, and that (μ, L_1, L_2) is a likelihood vector of **D**. In this case the test *f* of λ_1 against λ_2 is called a *likelihood ratio test*, and the number *K* is called the *critical ratio* of the test *f*, if the following condition is valid for all values of $x \in X$:

If $L_1(x)/L_2(x) > K$, then $f(x) = \lambda_1$, and if $L_1(x)/L_2(x) < K$, then $f(x) = \lambda_2$.

(In this condition the value of $L_1(x)/L_2(x)$ is taken to be $+\infty$ if $L_2(x) = 0$ and $L_1(x) \neq 0$, and it is taken to be undefined if $L_1(x) = L_2(x) = 0$.)

Theorem 6 will show that in our framework one can avoid the unsavory result that, if one does not consider randomized tests, a best test can make a choice that "goes against the evidence." According to this theorem all optimal tests must be a likelihood ratio tests when the payoffs are strictly epistemic, *i.e.* when the statistician prefers choosing the actual probability distribution to choosing one of the other distributions.

Theorem 6. Suppose that $\varepsilon = ((\lambda_1, \lambda_2), \Omega)$ is an experiment type with two probability distributions, that $\mathbf{D} = (\varepsilon, M, X, pay)$ is an ideal decision problem, and that *pay* is a strictly epistemic payoff function. Suppose further that $S = (\varepsilon, \mathbf{p})$ is an ensemble of token experiments in which $\mathbf{p} = (p, 1-p)$, $p \neq 1$, and define the number *K* by

$$K = \frac{1-p}{p} \ \frac{\Delta_2}{\Delta_1},$$

where $\Delta_2 = pay(\lambda_2|\lambda_2) - pay(\lambda_1|\lambda_2)$, and $\Delta_1 = pay(\lambda_1|\lambda_1) - pay(\lambda_2|\lambda_1)$. Now a decision function $f^* \in F(X, M; \varepsilon, pay)$ is optimal within *S* if it is a likelihood ratio test with the critical ratio *K*.

In the special case of simple payoffs, $\Delta_1 = \Delta_2 = 1$. If we were to further assume that $p = ½$, then we would end up with a simple likelihood test in which one chose the hypothesis with the greater likelihood whenever the likelihoods of the two hypotheses are different.

When Theorem 5 is combined with Theorem 6 it yields an easy proof for a version of the Neyman-Pearson theorem which has been adapted to our framework.

The Neyman-Pearson Theorem. Suppose that $\varepsilon = ((\lambda_1, \lambda_2), \Omega)$ is an experiment type with two probability distributions, that $\mathbf{D} = (\varepsilon, M, X, pay)$ is an ideal decision problem, and that

pay is a strictly epistemic payoff function.. If $f \in F(X, M; \varepsilon, pay)$ is a likelihood ratio test, it is a best test of λ_1 against λ_2.

This theorem provides a partial answer to the question of optimality by showing that likelihood ratio tests should be preferred to others. However, in section 2 we described three well-known limitations of this partial answer. These were the facts that the Neyman-Pearson theorem does not answer the questions which hypothesis should count as the null hypothesis and what the size of the chosen test should be, and that the best test of a given size might be a randomized test. A less obvious limitation of the Neyman-Pearson theorem that our framework has made apparent is that the theorem applies only to ideal decision problems. If neither of the distributions in *M* has actually generated the data, so that the decision problem is not ideal, and if the payoffs are simple, then the optimal payoff for choosing either of the members of *M* is the same; it is zero! Although the Neyman-Pearson theorem is valid also in this case, in this case it does not by itself provide a reason for preferring likelihood ratio tests to other tests.

Our Theorem 6 provides us with something further to say about the problems of arbitrariness. When the size of a likelihood ratio test is chosen to be e.g. 5%, this choice determines one way of setting the value of critical ratio *K* of the likelihood ratio test. Theorem 6 describes an alternative way of setting the value of *K*: its value can be calculated on the basis of the value of *p* and the properties of the relevant payoffs. When *p* is interpreted to be a measure of a person's subjective prior belief about the token experiment at hand, the choice between the two ways of choosing the value of *K* is a choice between conventionalism and subjectivism. As a third alternative, we have proposed a different way of choosing *K*: the probability *p* represents the frequency with which a probability distribution occurs in an ensemble of token experiments. There will still be many *K* values because there are many ensembles, but for each ensemble, the choice of *K* is fixed.

11 Why Stopping Rules are Irrelevant

There is a long-standing debate between N-P theorists and those who accept the Likelihood Principle concerning *optional stopping rules*. The issue which is at stake in this debate can be

illustrated with the following simple coin flipping example (Pratt *et al*. [1995], p. 542).[9] Let Z be the number of heads up, and R be the number of tails up in a sequence of flips of a coin, and suppose that in one particular sequence Z=9 and R=3. Now there are at least two experimental designs that may have led to this result. One is the standard one in which the coin is tossed 12 times. The other involves a stopping rule that says "stop the experiment when $R = 3$." If one works out the probability that $Z \geq 9$ given the null hypothesis $\theta = 0.5$, one finds that the probabilities depend on the experimental design, as follows:

$$\Pr_1\left(Z \geq 9 \middle| \theta = 0.5\right) = \Pr_1\left(Z = 9, R = 3 \middle| \theta = 0.5\right) + \cdots + \Pr_1\left(Z = 12, R = 0 \middle| \theta = 0.5\right),$$

$$\Pr_2\left(Z \geq 9 \middle| \theta = 0.5\right) = \Pr_2\left(Z = 9, R = 3 \middle| \theta = 0.5\right) + \Pr_2\left(Z = 10, R = 3 \middle| \theta = 0.5\right) + \cdots,$$

Here $\Pr_1$ denotes the probability given the standard version of the experiment, and $\Pr_2$ denotes the probability under the rule "stop the experiment when $r = 3$."

It is uncontroversial to say that these probabilities are different. It is even uncontroversial to say that the probability of the actual occurrence, $Z = 9$ and $R = 3$, is different in both case, since in the second case we know that the last toss is 'tails up' so that $\Pr_2\left(Z = 9, R = 3 \middle| \theta = 0.5\right)$ depends on the number of ways in which 9 heads can be distributed amongst the first 11 tosses. On the other hand, $\Pr_1\left(Z = 9, R = 3 \middle| \theta = 0.5\right)$ is calculated in terms of how many ways there are for 9 heads to be distributed amongst 12 tosses.

These probabilities are different because the second stopping rules are different. The first set of token experiments are stopped when $z + r = 12$, whereas the second ensemble is a set of token experiments that are stopped when $r = 3$. While the set of observable outcomes is different, some outcomes are possible under both rules. An experiment in which $Z = 9$ and $R = 3$ is such a case. The question is whether the difference in experimental design makes a difference to 'what the evidence says' according to an optimal rule.[10]

---

[9] The example was pointed out to us in a draft of an article by Deborah Mayo and Mike Kruse.

[10] While the differences are clear, the similarities between the two experiments are not. For example, can the token experiments in S_1 be described as the same experiment type as those in S_2? Although our argument for the irrelevance of stopping rules does not depend on it, it is worth pointing out that there is an experiment type that

According to a classical Neyman-Pearson theorist, it does. For in general we have

$$\Pr_1\left(Z=z,R=r\middle|\theta\right)=\binom{z+r}{z}\theta^z\left(1-\theta\right)^r,$$

and

$$\Pr_2\left(Z=z,R=r\middle|\theta\right)=\binom{z+r-1}{z}\theta^z\left(1-\theta\right)^r,$$

where

$$\binom{z+r}{z}=\frac{(z+r)!}{z!r!}.$$

If we are comparing these probabilities for the same values of z and r, they clearly differ only by the factor $(z+r)/r$. Further, it is easy to see that $\Pr_1\left(Z\geq 9\middle|\theta=0.5\right)\cong 0.073$ and that $\Pr_2\left(Z\geq 9\middle|\theta=0.5\right)\cong 0.0327$. For a one-sided test of size 5%, the null hypothesis is not rejected in the first case, but it is rejected in the second case. Therefore, the experimental design makes a difference to the evidential important of the outcome x, even though the outcome is the same in both cases. This conclusion contradicts the idea that only the likelihood of the *actual evidence*—rather than the features of some other, logically possible evidence which is not actually there—should determine which hypothesis gets chosen.

When our framework is applied to this example, Theorem 6 yields the result that, contrary to what N-P theorists say, it is, indeed, only the available evidence that counts. For, when one considers a case in which the value of θ is 0.5 with some fixed frequency p, and in which it has some other fixed value with the frequency $1-p$, the value of p will, of course, be identical in the two ensembles. Therefore, if also the payoffs are the same in the two cases, the value of K which

---

applies to both cases. We need to set Ω to be the countably infinite set of all possible infinite sequences of heads and tails because it is possible for the observed sequences of tosses to be arbitrarily long in S_2 (if per chance we got a very long string of heads). Then the two rival hypotheses are distributions defined over the same Ω in each case, so the experiment types are the same according our definition. Clearly, the frequency of the generating distributions in the experiment are also the same in both cases, so *the ensembles are the same*! The only difference is the stopping rules define different partitions X on Ω, and this means that the decision problems are different. However, the considered outcome, $Z = 9$ and $R = 3$, is a member of both partitions, so the optimal inference is the same in both cases.

is associated with the optimal decision function is the same in both cases. This means that the fact that the likelihood ratios are the same in both cases tells us that the optimal hypothesis is the same in both cases.

The difference between our framework and the Neyman-Pearson paradigm arises from the fact that different features of the ensemble count as relevant in the two approaches. The features of the ensemble which affect optimality as we have defined it include only the likelihoods, and the relative frequencies of the different distributions within the ensemble. Hence, if the existence of a stopping rule makes no difference to either of these things, it does not make any difference to the optimal choice of a hypothesis either. Moreover, our Main Theorem shows that the standard N-P practice of setting the size of a test at 5% is not only arbitrary, but it is sometimes *inconsistent* with optimality as we define it!

A common diagnosis of what is wrong with standard N-P practice is that is violates the *actualist* idea that the bearing of evidence should depend only on the actual evidence (Sober [1993]). The Neyman-Pearson Theorem appears to provide a response to this charge by showing that there is always a best test of each fixed size, and that this test is characterized as a likelihood ratio test with a fixed ratio K. The results of this test seem to depend on only the likelihood of the actually observed value of x, rather than on the likelihoods of some non-actual values of x. More specifically, if the value of K is the same for two similar experimental designs, and the likelihood *ratios* of the considered hypotheses are the same in the two cases, it is impossible that the null hypothesis is rejected in one case but not the other.

The reason why this nevertheless happens in the above example is that fixing the size of test does not fix the value of K, and its value is different in the two cases that we considered above. Hence, the stopping rules become relevant in the Neyman-Pearson paradigm because the value of K is different in the two cases. On the other hand, when our notion of optimality is applied, Theorem 6 shows that there is a *unique* optimal value of K. In this case there is no conventional choice of K to be made.

Previous arguments for the irrelevance of stopping rules have appealed to what is known as the Likelihood Principle. Here is one formulation of the principle:

The Likelihood Principle. In making inferences or decisions about θ after x is observed, all relevant experimental information is contained in the likelihood function for the observed x. Furthermore, two likelihood functions contain the same information about θ if they are proportional to each other (as functions of θ). (Berger [1985], p. 28.)

Unfortunately, this principle does not follow from our main theorem in its general form.[11] In particular, when a decision problem is not ideal, the likelihood vector mentioned in our main theorem is associated with the set of *generating* distributions in Λ, whereas the likelihoods in the Likelihood Principle are the likelihoods of the *hypothetical* distributions in M. As we have already explained, these are often conflated by frequentists and Bayesians alike, but this conflation is clearly a mistake whenever the hypothetical distributions are idealized and simplified to the point where they are plainly false. It is one of the strengths of our approach is that it proves the irrelevance of stopping rules from weaker assumptions.

12 Concluding Remarks

In section 2, we recounted three well known problems with the classical methods of hypothesis testing—the arbitrary choice of a null hypothesis, the conventional setting of the size of a test, and the unintuitive fact that sometimes the only tests that are best in the sense of Neyman and Pearson turn out to be randomized tests. In section 3, we turned to the classical frequentist method of parameter estimation, the practice of using *maximum likelihood* estimates. There we considered the problem of "regression to the mean", which is also known as "the base rate phenomenon". Each of these problems has contributed to the rise of Bayesian statistics and a widening acceptance of a subjectivist philosophy of statistics and science.

While Bayesianism has gained much ground, there still exists a strong core of classical statisticians who stand their ground despite of the seeming strength of the arguments against them (Mayo [1996]). Moreover, there are the neo-Fisherian likelihoodists (*e.g.*, Edwards [1987], Royall [1997]) who have maintained some kind of middle ground between the two factions.

---

[11] See Forster and Sober [forthcoming] for independent reasons for limiting the validity of the Likelihood Principle.

More recently, there have been followers of the "predictive paradigm" invented by Akaike ([1973]), such as Sakamoto et al ([1986]), Forster and Sober ([1994]) and Burnham and Anderson ([1998]). It appears to us that these splinter groups have philosophical tendencies that lie more towards a frequentist view of statistics.

Philosophically, the most important issue at stake is the choice between objectivism and subjectivism. However, it should be noted that even within the Bayesian camp there has always been a strong objectivist contingent (*e.g.*, Jaynes [1979]), which wants to restrict the choice of prior distributions (or at least the choice of the "first" prior) to priors that can be justified as representing *pure* objective states of ignorance.[12] When faced with persistent opposition and successes of opposing camps, subjective Bayesians are quick to claim that subjective Bayesianism provides the best theory that we have of decision-making. The present paper is an attempt to develop an alternative to the Bayesian theory that is clearly, and distinctly, grounded within the frequentist paradigm, without being vulnerable to the (strong) objections that have been made against classical frequentist methods.

In our view, a decision theory is not restricted to a theory of actions and practical decision making. In the case in which the payoffs represent the truth-related cognitive, or epistemic values of competing theories, then the objectivity of optimality leads to a theory of evidential support. In this theory a decision function is viewed as a pattern of inference that takes us from the observed outcome of an experiment to the 'best' hypothesis describing the underlying mechanism. If 'best' is described in purely subjective terms, then we have a description in need of a theory. All this changes when 'best' is defined in objective truth-related terms, for then optimality is the yardstick of statistical inference in the same way that 'truth preservation' is the cornerstone of deductive logic.

The key ingredient that allows frequentism to compete with the power of Bayesian decision theory is the notion of an ensemble, and the relativization of utilities, or payoffs, to ensembles. Inference *is* founded on ignorance (for if we were all-knowing, there would be no need for inference), but within our framework there is no such thing as *pure* ignorance. Rather, we define what it means to reduce our ignorance of a token case by considering the token case as a member

of an ensemble of cases. A token experiment is simultaneously a member of many ensembles and hence, there are simultaneous many ways in which one can reduce one's ignorance of it. Because the optimal form of inference may turn out to be different for different ensembles, the logical incompatibility of the inferred hypotheses need not impugn the rationality of either inference. At the same time, when all the *relevant* features of the contexts are fixed, the optimality of statistical inference is unambiguously defined in objective terms. In that way, the introduction of ensembles puts the world back into frequentism, and in doing so it provides a conceptually clear alternative to all Bayesian theories of inductive inference.

Acknowledgements

We thank Peter Andreae, Marty Barrett, Dan Hausman, Branden Fitelson, Stephen Leeds, and Elliott Sober for reading or listening to previous versions of this paper.

I. A. Kieseppä
Department of Philosophy
P. O. Box 24 (Unioninkatu 40)
00014 University of Helsinki
Finland

Malcolm Forster
Department of Philosophy
5185 Helen C. White Hall
University of Wisconsin
Madison, WI 53706
U. S. A.

[12] See Forster [1999], section 6, for a summary of the problems with this program.

References

Akaike, H. [1973]: 'Information Theory and an Extension of the Maximum Likelihood Principle', in B. N. Petrov and F. Csaki (eds.), *2nd International Symposium on Information Theory*, Budapest: Akademiai Kiado, pp. 267-81.

Blackwell, D. [1953]: 'Equivalent Comparisons of Experiments', *Annuls of Mathematical Statistics*, **24**, pp. 265-272.

Burnham, Kenneth P and Anderson, David R. [1998]: *Model Selection and Inference: A Practical Information-Theoretic Approach*, New York: Springer.

Cramér H. [1946]: *Mathematical Methods of Statistics*, Princeton, NJ: Princeton University Press.

Edwards, A. W. F. [1987]: *Likelihood*, Expanded Edition, Baltimore and London: The John Hopkins University Press.

Efron, Bradley [1978]: 'Controversies in the Foundations of Statistics', *American Mathematical Monthly* **85**: pp. 231-246.

Forster, M. R. [1995]: 'Bayes and Bust: The Problem of Simplicity for a Probabilist's Approach

to Confirmation', *British Journal for the Philosophy of Science* **46**, pp. 399-424.

Forster, M. R. [1999]: 'How Do Simple Rules "Fit to Reality" in a Complex World?', *Minds and Machines* **9**, pp. 543-564.

Forster, M. R. [2000]: 'Key Concepts in Model Selection: Performance and Generalizability', *Journal of Mathematical Psychology*, **44**, pp. 205-231.

Forster, M. R. and Elliott Sober [1994]: 'How to Tell when Simpler, More Unified, or Less *Ad Hoc* Theories will Provide More Accurate Predictions', *British Journal for the Philosophy of Science*, **45:** pp. 1 - 35.

Forster, M. R. and Elliott Sober [forthcoming]: 'Why Likelihood,' in Mark Taper and Subhash Lele (*eds*), *Likelihood and Evidence*, Chicago and London: University of Chicago Press.

Hogg, R. V. and A. T. Craig [1965]: *Introduction to Mathematical Statistics*. Second Edition. New York: Macmillan Publishing Co.

James, W., and Stein, C. M. [1961]: 'Estimation with Quadratic Loss', *Proceedings of the Fourth Berkley Symposium on Mathematical Statistics and Probability*, Vol. 1, Berkley: University of California Press, pp. 361-380.

Jaynes, E. T. [1979]: 'Where Do We Stand on Maximum Entropy?', in Ralph D. Levine and Myron Tribus (eds.) *The Maximum Entropy Formalism*, Cambridge, Mass.: The MIT Press, pp. 15-118.

Lehmann, E. L. [1950]: 'Some Principles of the Theory of Testing Hypotheses', *Annals of Mathematical Statistics*, **21**, pp. 1-26.

Lehmann, E. L., [1983]: *Theory of Point Estimation*, New York: John Wiley & Sons.

Mayo, Deborah G. [1996]: *Error and the Growth of Experimental Knowledge*, Chicago and London: The University of Chicago Press.

Neyman, J. and E. S. Pearson [1933]: 'On the Problem of the Most Efficient Tests of Statistical

Hypotheses', *Philosophical Transactions of the Royal Society of London*, Series A, **231**, pp. 289-337.

Royall, Richard M. [1997]: *Statistical Evidence: A likelihood paradigm*, Boca Raton: Chapman & Hall/CRC.

Sakamoto, Y., M. Ishiguro, and G. Kitagawa [1986]: *Akaike Information Criterion Statistics*, Dordrecht: Kluwer Academic Publishers.

Stein, C. M. [1956]: 'Inadmissibility of the Usual Estimator for the Mean of a Multivariate Normal Distribution', *Proceedings of the Third Berkley Symposium on Mathematical Statistics and Probability*, Vol. 1, Berkley: University of California Press, pp. 197-206.

Wald, Abraham [1950]: *Statistical Decision Functions*, New York: John Wiley & Sons.