A Rate of Incoherence Applied to Fixed-Level Testing

Mark J. Schervish, Teddy Seidenfeld,
and Joseph B. Kadane†‡
Carnegie Mellon University

It has long been known that the practice of testing all hypotheses at the same level (such as 0.05), regardless of the distribution of the data, is not consistent with Bayesian expected utility maximization. According to de Finetti's "Dutch Book" argument, procedures that are not consistent with expected utility maximization are incoherent and they lead to gambles that are sure to lose no matter what happens. In this paper, we use a method to measure the rate at which incoherent procedures are sure to lose, so that we can distinguish slightly incoherent procedures from grossly incoherent ones. We present an analysis of testing a simple hypothesis against a simple alternative as a case-study of how the method can work.

1. Introduction. Cox (1958) and Lindley (1972) have shown that the practice of testing all hypotheses at the same level, regardless of the distribution of the data, can lead to inadmissibility and incompatibility with Bayesian decision theory. One of the most compelling arguments for Bayesian decision theory and the use of probability to model uncertainty is the "Dutch Book" argument, which says that if you are willing to accept either side of each bet implied by your statements and finite combinations of these together, either

(a) those statements are "coherent," that is they comport with the axioms of probability, or

†Mark J. Schervish and Joseph B. Kadane are Professor, Department of Statistics and Teddy Seidenfeld is Professor, Departments of Statistics and Philosophy, Carnegie Mellon University, Pittsburgh, PA 15213.

‡This paper is an extract from Schervish, Seidenfeld, and Kadane (1999). This research was supported by NSF grant DMS-98-01401 and NSF grant DMS-96-26181.

Philosophy of Science, 69 (September 2002) pp. S248–S264. 0031-8248/2002/69supp-0022$10.00
Copyright 2002 by the Philosophy of Science Association. All rights reserved.

https://doi.org/10.1086/341850 Published online by Cambridge University Press

(b) a gambler betting against you can choose bets so that you are a sure loser.

Excellent introductions to the concepts of coherence and Dutch Book can be found in Shimony (1955), Freedman and Purves (1969), and de Finetti (1974, Section 3.3).

As a practical matter, it is very difficult to structure one's statements of probabilities (i.e. previsions) in such a way that they both reflect one's beliefs and are coherent (see Kadane and Wolfson 1998). Yet the dichotomy above does not allow for discussion of what sets of previsions may be "very" incoherent or just "slightly" incoherent. This paper explores a remedy for this by studying how quickly an incoherent bookie can be forced to lose money. A faster rate of sure financial decline to the bookie, or a faster rate of guaranteed profit to the gambler, is associated with a greater degree of incoherence.

The problem as stated so far requires some normalization. Suppose that a particular combination of gambles yields a sure loss y for the bookie. Then multiplying each gamble by the same constant $k > 0$ will create a combination of gambles that yields sure loss ky. In this paper we explore how to perform the normalization from the bookie's perspective. We introduce normalizations in Section 3. To fix ideas, however, consider that the bookie cannot be assumed to have infinite resources. A wise gambler would want to be sure that the bookie could cover all the bets. One way to do this would be to require the bookie to escrow the maximum amount that the bookie could lose on each gamble separately. Thus we can ask how much the bookie can be forced to lose for sure, given a specified level of escrow that the bookie can offer. In Section 4, we apply these ideas to assess the incoherence of the practice of fixed level hypothesis testing.

2. Gambles and Incoherence. Think of a random variable X as a function from some space T of possibilities to the real numbers. We assume that, for a bounded random variable X, a bookie might announce some value of x such that he/she finds acceptable all gambles whose net payoff to the bookie is $\alpha(X - y)$ for $\alpha > 0$ and $y < x$. Each such x will be called a *lower prevision for X*. In addition, or alternatively, the bookie may announce some value of x such that the gamble $\alpha(X - y)$ is acceptable when $\alpha < 0$ and $y > x$. These x will be called *upper previsions for X*. We allow that the bookie might announce only upper previsions or only lower previsions or both. For example, if X is the indicator I_A of an event A the bookie might announce that he/she finds acceptable all gambles of the form $\alpha(I_A - y)$ for $y < p$ if $\alpha > 0$ but no other gambles, in particular, not for $y = p$. It will turn out not to matter for any of our results whether or not the bookie finds the gamble $\alpha(I_A - p)$ acceptable. In the special case

https://doi.org/10.1086/341850 Published online by Cambridge University Press

in which x is both an upper prevision and a lower prevision, we call x a *prevision* of X and denote it $P(X)$. Readers interested in a thorough discussion of upper and lower previsions should refer to Walley (1991).

It will be convenient to assume that, whenever both an upper prevision x^+ and a lower prevision x^- have been assessed for the same random variable X, $x^- \leq x^+$, otherwise the bookie is willing to sell X for a certain price and then buy it right back for a higher price. Although such incoherence could be measured, it requires cumbersome bookkeeping that makes general results difficult to understand. (See Examples 4 and 5 of Schervish et al. 1999.) In particular, this assumption implies that there can be at most one prevision of X.

A collection $x_1, \ldots, x_n$ of upper and/or lower previsions for $X_1, \ldots, X_n$ respectively is *incoherent* if there exists $\varepsilon > 0$ and a collection of acceptable gambles $\{\alpha_i(X_i - y_i)\}_{i=1}^n$ such that

$$\sup_{t \in T} \sum_{i=1}^{n} \alpha_i \left(X_i(t) - y_i \right) < -\varepsilon, \tag{1}$$

in which case we say that a *Dutch Book* has been made against the bookie. Of course, we would need $\alpha_i > 0$ and $y_i < x_i$ if x_i is a lower prevision for X_i and we would need $\alpha_i < 0$ and $y_i > x_i$ if x_i is an upper prevision for X_i. When a collection of upper and/or lower previsions is incoherent, we would like to be able to measure how incoherent they are. As we noted earlier, the ε in (1) is not a good measure because we could make ε twice as big by multiplying all of the α_i in (1) by 2, but the previsions would be the same. Instead, we need to determine some measure of the sizes of the gambles and then consider the left-hand side of (1) relative to the total size of the combination of gambles. This is what we do in Section 3.

3. Normalizations. To begin, consider a single acceptable gamble such as $Y = \alpha (X - y)$. There are a number of possible ways to measure the size of Y. For example $\sup_t |Y(t)|$ or $\sup_t - Y(t)$ might be suitable measures. This last one has a nice interpretation. It is the most that the bookie can lose on the one particular gamble. It measures a gamble by its extreme value in the same spirit as Dutch Book measures incoherence in terms of an extreme value (the minimum payoff to the gambler) of a combination of gambles. Alternatively, if we think of the gambler and bookie as adversaries with regard to this one gamble Y, the gambler might want to be sure that the bookie will be able to pay up when the bet is settled. We could imagine that the gambler requests that the bookie place funds in escrow to cover the maximum possible loss. So, for the remainder of the paper, we will call $e(Y) = \sup_t - Y(t)$ the *escrow for gamble* Y. Note that

https://doi.org/10.1086/341850 Published online by Cambridge University Press

$e(cY) = ce(Y)$ for all $c > 0$. We use the escrow to measure the size of the gamble Y.

Example 1. Let A be an arbitrary event which is neither certain to occur nor certain to fail. Suppose that a lower prevision p is given, and consider the gamble $Y(t) = \alpha(I_A(t) - p)$ with $\alpha > 0$. Then $\sup_t -Y(t) = \alpha p$, and the escrow is $e(Y) = \alpha p$. If an upper prevision q is given and $\alpha < 0$, then $\sup_t -Y(t) = -\alpha(1 - q) = e(Y)$, where $\alpha < 0$.

When we consider more than one gamble simultaneously, we need to measure the size of the entire collection. We assume that the size of (escrow for) a collection of gambles is some function of the escrows for the individual gambles that make up the collection. That is $e(Y_1, \ldots, Y_n) = f_n(e(Y_1), \ldots, e(Y_n))$. In order for a function to be an appropriate measure of size, we have a few requirements. First,

$$f_n(cx_1, \ldots, cx_n) = cf_n(x_1, \ldots, x_n), \quad \text{for all } c > 0,\ x_1, \ldots, x_n. \tag{2}$$

Equation (2) says that the function f_n must be homogeneous of degree 1 in its arguments so that scaling up all the gambles by the same amount will scale the escrow by that amount as well. Second, since we are not concerned with the order in which gambles are made, we require

$$f_n(x_1, \ldots, x_n) = f_n(y_1, \ldots, y_n), \quad \text{for all } n,\ x_1, \ldots, x_n \text{ and all permutations } (y_1, \ldots, y_n) \text{ of } (x_1, \ldots, x_n). \tag{3}$$

Third, in keeping with the use of escrow to cover bets, we will require that, if a gamble is replaced by one with higher escrow, the total escrow should not go down:

$$f_n(x_1, \ldots, x_n) \text{ is nondecreasing in each of its arguments.} \tag{4}$$

If a gamble requires 0 escrow, we will assume that the total escrow is determined by the other gambles:

$$f_{n+1}(x_1, \ldots, x_n, 0) = f_n(x_1, \ldots, x_n), \quad \text{for all } x_1, \ldots, x_n \text{ and all } n. \tag{5}$$

Since nobody can lose more than the sum of the maximum possible losses from all of the accepted gambles, we require that

$$f_n(x_1, \ldots, x_n) \le \sum_{i=1}^{n} x_i \quad \text{for all } n \text{ and all } x_1, \ldots, x_n. \tag{6}$$

Small changes in the component gambles should produce only small changes in the escrow, so we require that

$$f_n \text{ is continuous for every } n. \tag{7}$$

https://doi.org/10.1086/341850 Published online by Cambridge University Press

Finally, since we have already decided how to measure the size of a single gamble, we require

$$f_1(x) = x. \tag{8}$$

So, if $Y_1, \ldots, Y_n$ is a collection of gambles, we can set $e(Y_1, \ldots, Y_n) = f_n(e(Y_1), \ldots, e(Y_n))$ for some function f_n satisfying (2)–(8) and call $e(Y_1, \ldots, Y_n)$ an *escrow for the collection of gambles.* Every sequence of functions $\{f_n\}_{n=1}^{\infty}$ that satisfy (2)–(8) leads to its own way of defining escrow. Such a sequence is called an *escrow sequence.* Each function in the sequence is an *escrow function.*

We can find a fairly simple form for all escrow sequences. Combining (8), (4), and (5), we see that $f_n(x_1, \ldots, x_n) \geq \max\{x_1, \ldots, x_n\}$. From (3), we conclude that f_n is a function of the ordered values $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$ of $x_1, \ldots, x_n$. That is, $f_n(x_1, \ldots, x_n) = f_n(x_{(1)}, \ldots, x_{(n)})$. Combining these results with (6), we get

$$0 \leq f_n\left(x_{(1)},\ldots,x_{(n)}\right) - x_{(n)} \leq \sum_{i=1}^{n-1} x_{(i)}. \tag{9}$$

Let $\lambda_n(x_{(1)}, \ldots, x_{(n)}) = (f_n(x_{(1)}, \ldots, x_{(n)}) - x_{(n)})/x_{(n)}$ so that

$$f_n(x_{(1)}, \ldots, x_{(n)}) = x_{(n)}\,(1 + \lambda_n(x_{(1)}, \ldots, x_{(n)})). \tag{10}$$

In order to satisfy (5), we need $\lambda_n(0, x_{(2)}, \ldots, x_{(n)}) = \lambda_{n-1}(x_{(2)}, \ldots, x_{(n)})$. In order to satisfy (2), λ_n must be invariant under common scale changes for all of its arguments. That is

$$\lambda_n(cx_{(1)}, \ldots, cx_{(n)}) = \lambda_n(x_{(1)}, \ldots, x_{(n)}).$$

Every such function can be written as

$$\lambda_n\left(x_{(1)},\ldots,x_{(n)}\right) = \gamma_n\left(\frac{x_{(1)}}{x_{(n)}},\ldots,\frac{x_{(n-1)}}{x_{(n)}}\right).$$

In order to satisfy (4), we must have γ_n nondecreasing in each of its arguments. In order to satisfy (9), we must have

$$0 \leq \gamma_n(y_1,\ldots,y_{n-1}) \leq \sum_{i=1}^{n-1} y_i.$$

In summary, every escrow sequence satisfies

$$f_n(x_1,\ldots,x_n) = x_{(n)}\left[1 + \gamma_n\left(\frac{x_{(1)}}{x_{(n)}},\ldots,\frac{x_{(n-1)}}{x_{(n)}}\right)\right], \tag{11}$$

https://doi.org/10.1086/341850 Published online by Cambridge University Press

for some sequence $\gamma_1, \gamma_2, \ldots$ of continuous functions where $\gamma_1 \equiv 0$ and for $n > 1$ the functions satisfy the following properties:

- $\gamma_n(y_1, \ldots, y_{n-1})$ is defined and continuous for $0 \leq y_1 \leq y_2 \leq \ldots \leq y_{n-1} \leq 1$,
- γ_n is nondecreasing in each argument,
- $0 \leq \gamma_n(y_1, \ldots, y_{n-1}) \leq \sum_{i=1}^{n-1} y_i$
- $\gamma_n(0, y_2, \ldots, y_{n-1}) = \gamma_{n-1}(y_2, \ldots, y_{n-1})$,
- $x[1 + \gamma_n(y_1/x, \ldots, y_{n-1}/x)]$ is nondecreasing in x for all $y_1 \leq \ldots \leq y_{n-1} \leq x$.

(The last condition is equivalent to f_n being nondecreasing in $x_{(n)}$.) It is straightforward to show that every sequence that meets this description satisfies (2)–(8), hence we have characterized escrow sequences.

One simple collection of escrow sequences consists of all sequences in which $\gamma_n(y_1, \ldots, y_{n-1}) = \gamma \sum_{i=1}^{n-1} y_i$ for some common constant $\gamma \in [0,1]$. In this case, we get a family of escrow functions:

$$f_{\gamma,n}(x_1,\ldots,x_n) = x_{(n)} + \gamma \sum_{i=1}^{n-1} x_{(i)}, \tag{12}$$

for each $0 \leq \gamma \leq 1$. Another example is $\gamma_n(z_1, \ldots, z_{n-1}) = z_{n-1}$ for $n > 1$. This one makes the total escrow equal to the sum of the two largest individual gamble escrows. Other functions are possible, but we will focus on $f_{\gamma,n}$ for $0 \leq \gamma \leq 1$. It is easy to see that the two extreme escrow functions correspond to $\gamma = 0$ and $\gamma = 1$:

$$f_{0,n}(x_1,\ldots,x_n) = x_{(n)},$$
$$f_{1,n}(x_1,\ldots,x_n) = \sum_{i=1}^{n} x_i.$$

We now propose to measure the incoherence of a collection of incoherent previsions based on a normalization by an escrow. For a combination of gambles $Y = \sum_{i=1}^{n} \alpha_i(X_i - y_i)$ define the *guaranteed loss* to be $G(Y) = -\min\{0, \sup_{t \in T} Y(t)\}$. So, Dutch Book can be made if there exists a combination of acceptable gambles whose guaranteed loss is positive. The *rate of guaranteed loss* relative to a particular escrow function f_n is

$$H(Y) = \frac{G(Y)}{f_n\left(e(Y_1),\ldots,e(Y_n)\right)}, \tag{13}$$

https://doi.org/10.1086/341850 Published online by Cambridge University Press

where $Y_i = \alpha_i(X_i - y_i)$. Notice that the rate of guaranteed loss is unchanged if all α_i are multiplied by a common positive number. Also, the rate of guaranteed loss is interesting only when Dutch Book is made, otherwise the numerator is 0. The denominator $f_n(e(Y_1), \ldots, e(Y_n))$ is 0 if and only if $e(Y_i) = 0$ for all i. This will occur if and only if the agent who is required to escrow cannot lose any of the individual gambles, in which case the numerator is 0 as well, and we will then define the rate of guaranteed loss to be 0 (since we cannot guarantee loss). The *extent of incoherence* relative to an escrow and corresponding to a collection of previsions will be the supremum of $H(Y)$ over all combinations Y of acceptable gambles. If the previsions are incoherent then the maximum rate of guaranteed loss is positive, otherwise it is 0.

There is a slightly simpler way to compute the extent of incoherence corresponding to a finite set of previsions than directly from the definition.

Theorem 1. Let $x_1, \ldots, x_n$ be a collection of incoherent upper and/or lower previsions for $X_1, \ldots, X_n$. Define

$$g(\alpha_1,\ldots,\alpha_n) = \sup_{t\in T} \sum_{i=1}^{n} \alpha_i \left(X_i(t) - x_i\right),$$
$$h(\alpha_1,\ldots,\alpha_n) = f_n\left(e\left(\alpha_1[X_1 - x_1]\right),\ldots,e\left(\alpha_n[X_n - x_n]\right)\right).$$

Then the rate of incoherence is

$$\sup_{\alpha_1,\ldots,\alpha_n} \frac{-g(\alpha_1,\ldots,\alpha_n)}{h(\alpha_1,\ldots,\alpha_n)} \tag{14}$$

or equivalently

$$-\inf_{\alpha_1,\ldots,\alpha_n} g(\alpha_1,\ldots,\alpha_n), \text{ subject to } h(\alpha_1,\ldots,\alpha_n) \le 1. \tag{15}$$

The supremum and infimum are taken over those α_i that have the appropriate signs.

As with all of the more lengthy proofs in this paper, the proof of Theorem 1 is in Schervish et al. (1999). Theorem 1 allows us to ignore the fact that the gamble $\alpha(X - x)$ might not be acceptable when x is a lower or upper prevision for X if we are proving results concerning the rate of incoherence.

Note that if a collection of gambles satisfies the escrow condition $h(\alpha_1, \ldots, \alpha_n) \le 1$, then every subcollection also satisfies the escrow condition because of (5). Also, note that, since every escrow function f_n is

https://doi.org/10.1086/341850 Published online by Cambridge University Press

between $f_{0,n}$ and $f_{1,n}$, the maximum and minimum possible rates of incoherence correspond to these two escrows.

When we use the bookie's escrow with $e(Y) = \sup_t - Y(t)$ for each individual gamble Y, we call the extent of incoherence *the maximum rate of guaranteed loss,* since the extent of incoherence is the maximum rate at which the bookie can be forced to lose relative to the particular escrow chosen. We focus on the family of escrows $f_{\gamma,n}$ defined in (12). The corresponding maximum rates of guaranteed loss will be donoted ρ_γ.

4. Testing Simple Hypotheses At a Fixed Level. Lindley (1972, 14) argues that it is incoherent to test all hypotheses at the same level, such as .05. (See also Seidenfeld, Schervish, and Kadane 1990.) Cox (1958) gave an example of how testing all hypotheses at the same level leads to inadmissibility. In this section, we show how this incoherence and inadmissibility can be measured using the measure of incoherence ρ.

Consider the case of testing a simple hypothesis against a simple alternative. Let f_0 and f_1 be two possible densities for a random quantity X, and let f be the "true" density of X. Suppose that we wish to test the hypothesis $H_0 : f = f_0$ versus the alternative $H_1 : f = f_1$. To write this as a decision problem, let the parameter space and the action space both be $\{0, 1\}$ where action $a = 0$ corresponds to accepting H_0 and action $a = 1$ corresponds to rejecting H_0. Also, parameter i corresponds to $f = f_i$ for $i = 0, 1$. Let the loss function have the form

$$L(i,a) = \begin{cases} c_i & \text{if } f = f_i \text{ and } a = 1 - i, \\ 0 & \text{otherwise,} \end{cases} \tag{16}$$

with $c_0, c_1 > 0$. The Neyman-Pearson lemma says that the most powerful tests of their sizes and the Bayes rules with respect to all priors have the form: For some constant k, choose $a = 1$ if $f_1(x) > kf_0(x)$, choose $a = 0$ if $f_1(x) < kf_0(x)$, and do whatever you wish (even randomization) if $f_1(x) = kf_0(x)$. Now, suppose that someone chooses a value of k and declares that he/she prefers the corresponding test to all other tests. One could infer from this choice an "implied prior" over the two possibilities f_0 and f_1. If $\Pr(f = f_0) = p$ and $\Pr(f = f_1) = 1 - p$, then the Bayes rule is to choose $a = 1$ if $pc_0f_0(x) < (1-p)c_1f_1(x)$, which corresponds to $k = pc_0/[(1-p)c_1]$. So $p = c_1k/(c_0 + c_1k)$.

Of course, a classical statistician who refuses to use prior and posterior probabilities will not acknowledge the implied prior. However, incoherence will arise if two tests about the same parameter imply different priors. We illustrate this with a version of the example of Cox (1958). Since the only part of the loss function that matters is c_0/c_1, let $c_1 = 1$. As an ex-

https://doi.org/10.1086/341850 Published online by Cambridge University Press

ample, let f_0 and f_1 be normal distributions with different means θ but the same variance σ^2. Suppose that the hypothesis is $H_0 : \theta = 0$ versus $H_1 : \theta = 1$ with $c_0 = 1$. (The phenomenon we illustrate here applies more generally as shown in Theorem 2). Suppose that either $\sigma = 1$ or $\sigma = 0.3$ will be true, but we will not know which until we observe the data. That is, the data consist of the pair (X, σ). Let $\Pr(\sigma = 1) = 0.5$, so that σ is ancillary. A classical statistician who prefers level 0.05 tests whenever available might think that, after observing σ a conditional level 0.05 test should still be preferred to a test whose conditional level given σ is something else. The most powerful conditional level 0.05 test is to reject $H_0 : \theta = 0$ if $X > 1.645\sigma$. The most powerful marginal level 0.05 test rejects H_0 if $X > 0.5 + 0.9438\sigma^2$ and is the Bayes rule with respect to the prior $\Pr(\theta = 0) = 0.7199$. The marginal power of the Bayes rule is 0.6227, while the marginal power of the conditional level 0.05 test is 0.6069. Since both tests have the same level, the conditional test is inadmissible.

To see how this inadmissibility translates into incoherence, we interpret the preference of one test δ_1 to another δ_2 as a preference for suffering a loss equal to the risk function of δ_1 to suffering a loss equal to the risk function of δ_2. The risk function of a test δ is

$$R(\theta,\delta) = \begin{cases} c_0 \text{ times size of test } \delta & \text{if } \theta = 0, \\ \text{one minus power of test } \delta & \text{if } \theta = 1. \end{cases}$$

To say that δ_1 is preferred to δ_2 means that $R(\theta, \delta_2) - R(\theta, \delta_1)$ is an acceptable gamble. In our example, let $\alpha_\delta(\sigma)$ and $\beta_\delta(\sigma)$ denote the size and power of test δ conditional on σ. Also, let $\beta_{cl}(\sigma)$ denote the power of the most powerful level 0.05 test. Then, for each σ, the classical statistician prefers the level 0.05 test to every other test. So, for each σ and all δ that are not the most powerful level 0.05 test, the following gamble is acceptable, even favorable:

$$\begin{aligned} R(\theta,\delta) - R(\theta,\delta_{cl}) &= \begin{cases} \big(\alpha_\delta(\sigma) - .05\big)c_0 & \text{if } \theta = 0, \\ \beta_{cl}(\sigma) - \beta_\delta(\sigma) & \text{if } \theta = 1 \end{cases} \\ &= a(I_A - b), \end{aligned} \tag{17}$$

where

$$\begin{aligned} A &= \{\theta = 0\}, \\ a &= \big(\alpha_\delta(\sigma) - .05\big)c_0 + \beta_\delta(\sigma) - \beta_{cl}(\sigma), \\ b &= \frac{\beta_\delta(\sigma) - \beta_{cl}(\sigma)}{\big(\alpha_\delta(\sigma) - .05\big)c_0 + \beta_\delta(\sigma) - \beta_{cl}(\sigma)}. \end{aligned}$$

https://doi.org/10.1086/341850 Published online by Cambridge University Press

In other words, b is an upper or lower prevision for A depending on whether $a < 0$ or $a > 0$.

We can make use of the construction in (17) to obtain a general result. Theorem 2 has a technical condition (concerning risk sets) that is known to be satisfied for problems of testing simple hypotheses against simple alternatives using fixed sample size and sequential tests. For more detail on risk sets, see Sections 3.2.4 and 4.3.1 of Schervish (1995).

Theorem 2. Let θ be a parameter and let the parameter space Ω consist of two points $\{0, 1\}$. Consider two decision problems D_0 and D_1 both with the same parameter space Ω and with nonnegative loss functions L_0 and L_1. Let the data in problem D_i be denoted X_i. Suppose that the risk sets for the two decision problems are closed from below. Suppose that an agent prefers the admissible decision rule δ_i to all others in problem D_i for $i = 0, 1$. For each decision rule ψ in problem D_i, let $R_i(\theta, \psi)$ denote the risk function. Let $A = \{\theta = 0\}$ and define

$$a_i(\psi) = \left[R_i(0,\psi) - R_i(0,\delta_i)\right] + \left[R_i(1,\delta_i) - R_i(1,\psi)\right],$$
$$b_i(\psi) = \frac{R_i(1,\delta_i) - R_i(1,\psi)}{\left[R_i(0,\psi) - R(0,\delta_i)\right] + \left[R_i(1,\delta_i) - R_i(1,\psi)\right]}. \tag{18}$$

If ψ is admissible in problem D_i and is not equivalent to δ_i, then

$$R_i(\theta, \psi) - R_i(\theta, \delta_i) = a_i(\psi)(I_A - b_i(\psi)). \tag{19}$$

If δ_0 and δ_1 are not Bayes rules with respect to a common prior, then there exist real numbers d_0 and d_1 and decision rules ψ_0 (in problem D_0) and ψ_1 (in problem D_1) such that the two gambles $d_0a_0(\psi_0)$ $(I_A - b_0(\psi_0))$ and $d_1a_1(\psi_1)$ $(I_A - b_1(\psi_1))$ are both acceptable, but

$$d_0a_0(\psi_0)(I_A - b_0(\psi_0)) + d_1a_1(\psi_1)\,(I_A - b_1(\psi_1)) < 0. \tag{20}$$

Also,

$$\rho_\gamma = \sup_{p_0 > p_1} \frac{p_0 - p_1}{\max\{p_0, 1 - p_1\} + \gamma \min\{p_0, 1 - p_1\}}, \tag{21}$$

where the supremum is over all $p_0 > p_1$ such that either δ_i is a Bayes rule with respect to prior p_i for $i = 0, 1$ or δ_i is a Bayes rule with respect to prior p_{1-i} for $i = 0, 1$.

As an example of Theorem 2, return to the test of $H_0 : \theta = 0$ versus $H_1 : \theta = 1$ where $X \sim N(\theta, \sigma^2)$ with σ being one of two known values.

https://doi.org/10.1086/341850 Published online by Cambridge University Press

Let σ_i be the value of σ for problem D_i for $i = 0, 1$. The implied prior probability of $\{\theta = 0\}$ is $k/(c_0 + k)$ where k comes from the form of the Neyman-Pearson test and c_0 comes from the loss function (16). If a classical statistician decides to test the hypothesis $H_0 : \theta = 0$ at level 0.05 regardless of the value of σ, this will often require choosing two different values of k in the Neyman-Pearson lemma. In fact, $k = \exp(1.645/\sigma - 0.5/\sigma^2)$, and the implied prior is

$$p(\sigma) = \left[1 + c_0 \exp\left(\frac{-1.645}{\sigma} + \frac{0.5}{\sigma^2}\right)\right]^{-1}, \tag{22}$$

not a one-to-one function of σ, but highly dependent on σ. In this problem, the lower boundary of the risk set is a strictly convex differentiable function so that every point on the lower boundary has a unique support line and hence a unique prior such that the corresponding rule is Bayes with respect to that prior. This means that the sup in (21) is unnecessary and

$$\rho_\gamma = \frac{p(\sigma_0) - p(\sigma_1)}{\max\{p(\sigma_0), 1 - p(\sigma_1)\} + \gamma \min\{p(\sigma_0), 1 - p(\sigma_1)\}},$$

if $p(\sigma_0) > p(\sigma_1)$ with a similar formula if $p(\sigma_0) < p(\sigma_1)$. For the case of $\gamma = 1$, this simplifies to

$$\rho_1 = \frac{|p(\sigma_0) - p(\sigma_1)|}{1 + |p(\sigma_0) - p(\sigma_1)|},$$

In this case, the degree of incoherence is a simple monotone function of how far apart the implied priors are for the two level 0.05 tests. Theorem 2 would also apply if one or both of the two decision problems were a sequential decision problem in which the loss equals the cost of observations plus the cost of terminal decision error.

It is interesting to examine the relationship between ρ_γ and the pair (σ_0, σ_1). For example, suppose that $\sigma_i = 2/\sqrt{n_i}$ for two different values n_0 and n_1. This would correspond to the data consisting of a sample $X_1, \ldots, X_{n_i}$ of independent normal random variables with mean θ and variance 4, where the sample size is n_i in decision problem D_i. Figure 1 is a plot of ρ_γ as a function of (n_0, n_1) for the case $\gamma = 1$ and $c_0 = 19$. Other values of γ produce plots with similar appearances. Of course the values of ρ_γ are higher for $\gamma < 1$. We chose $c_0 = 19$ to correspond to the classical choice of $\alpha = 0.05$. Each curve in Figure 1 corresponds to a fixed value of n_0 and lets n_1 vary from 1 to 150. Notice that each curve touches 0 where n_1

https://doi.org/10.1086/341850 Published online by Cambridge University Press

$= n_0$ since there is no incoherence in that case. Some of the curves come close to 0 in another location as well. For example, the $n_0 = 27$ curve comes close to 0 near $n_1 = 2$ and the $n_0 = 2$ curve comes close to 0 near $n_1 = 27$. The reason is that the implied priors corresponding to $\sigma = 2/\sqrt{2}$ and $\sigma = 2/\sqrt{27}$ are nearly the same (0.7137 and 0.7107 respectively), making these two level 0.05 tests nearly coherent. Indeed, the entire curves corresponding to $n_0 = 2$ and $n_0 = 27$ are nearly identical for this same reason. Another interesting feature of Figure 1 is that all of the curves are rising as $n_1 \to \infty$ but not to the same level. As $n_1 \to \infty$, the implied prior on $A = \{\theta = 0\}$ converges to 0. But if n_0 is large also, then the implied prior corresponding to $\sigma = 2/\sqrt{n_0}$ is also close to 0. For example, with $n_0 = 100$, the implied prior is 7.3×10^{-4}. There is not much room for incoherence between 0 and 7.3×10^{-4}, so the curve corresponding to $n_0 = 100$ will not rise very high. On the other hand, with $n_0 = 11$, the implied prior is 0.1691, leaving lots of room for incoherence. In fact, since 0.1691 is the largest possible implied prior in this example, all of the other curves

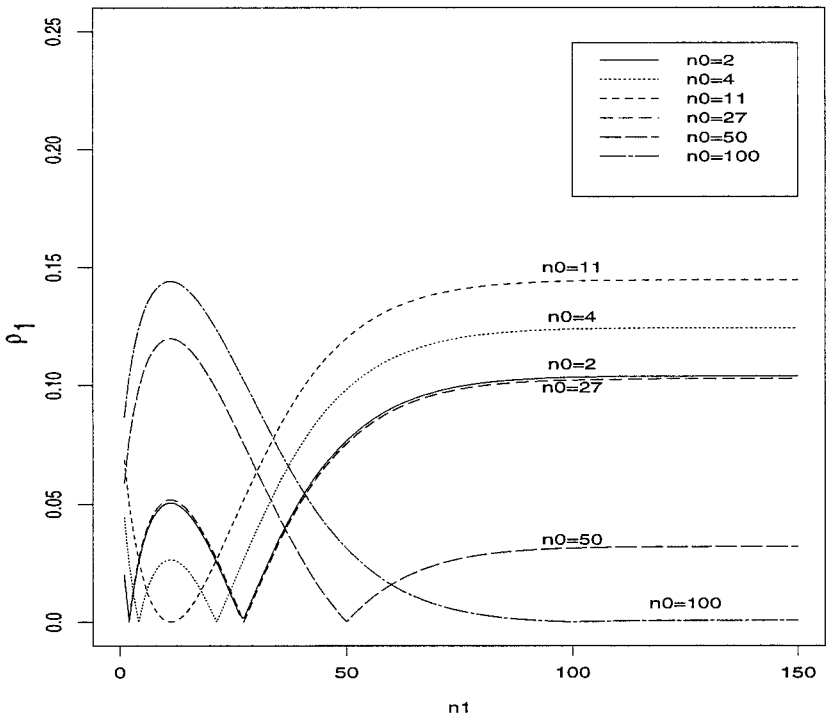


Figure 1. Plot of ρ_1 for level 0.05 testing as a function of n_1 (running from 1 to 150) for various values of n_0 with $c_0 = 19$.

https://doi.org/10.1086/341850 Published online by Cambridge University Press

have local maxima near $n_1 = 11$, and the $n_0 = 11$ curve rises higher than all the others as n_1 increases. Since the limiting implied prior is 0 as $n_1 \to \infty$, the height to which the n_0 curve rises as n_1 increases is

$$\frac{\left|p\left(2/\sqrt{n_0}\right) - 0\right|}{1 + \left|p\left(2/\sqrt{n_0}\right) - 0\right|} = \left[2 + 19\exp\left(-0.8224\sqrt{n_0} + 0.125n_0\right)\right]^{-1}.$$

The curve corresponding to $n_0 = 4$ illustrates the original example of Cox (1958), in which the alternative is that θ equals the larger of the two different standard deviations.

Lehmann (1958) offered a rule of thumb for choosing tests based on both their size and their power. One chooses a postive number λ (such as c_0 in the loss function (16)) and then chooses the test so that the probability of type II error equals λ times the probability of type I error. In our case of testing one normal distribution against another one with the same variance σ^2, this procedure will produce the minimax rule with loss (16) if $\lambda = c_0$. When $\lambda = 1$, it is easy to check that Lehmann's suggestion is the Bayes rule with respect to the prior with $\Pr(\theta = 0) = 1/(1 + c_0)$ for all σ. In this special case $\rho_\gamma = 0$ for all γ. However, when $\lambda \neq 1$, each σ leads to a Bayes rule with respect to a different implied prior. Assuming that the test will be to reject H_0 if $X > y$, one must solve the equation

$$\lambda\Phi\left(-\frac{y}{\sigma}\right) = \Phi\left(\frac{y-1}{\sigma}\right). \tag{23}$$

The implied prior, assuming, still, that the loss is (16), is then

$$p_L(\sigma) = \left[1 + c_0\exp\left(\frac{0.5 - y}{\sigma^2}\right)\right]^{-1}. \tag{24}$$

When $\lambda = 1$, $y = 1/2$ solves (23). Plugging this into (24) yields $p_L(\sigma) = 1/(1 + c_0)$ for all σ as we noted earlier. Two other limiting cases are of interest. If $\sigma \to \infty$, then y/σ must converge to $\Phi^{-1}(\lambda/[1 + \lambda])$ in order for (23) to hold. This would make the type I error probability $1/(1 + \lambda)$, and the limit of $p_L(\sigma)$ would be $1/(1 + c_0)$. It is not difficult to see that the type I error probability is highest for $\sigma = \infty$, so it must be less than $1/(1 + \lambda)$ for all finite σ. If $\sigma \to 0$, then $(y - 1/2)/\sigma^2$ must converge to $\log(\lambda)$ in order for (23) to hold. In this case, $p_L(\sigma)$ converges to $\lambda/(\lambda + c_0)$. For the case of $\lambda = c_0 = 19$, Figure 2 shows the value of ρ_1 with $\sigma_i = 2/\sqrt{n_i}$ for $i = 0, 1$ for various values of n_0 and n_1 in the same spirit (and on the same vertical scale) as Figure 1.

https://doi.org/10.1086/341850 Published online by Cambridge University Press

The curves in Figure 2 are higher for large n_1 than the corresponding curves in Figure 1. This means that, when $c_0 = 19$, Lehmann's procedure with $\lambda = 19$ is more incoherent (as measured by ρ_1) for large values of n_1 than testing at level 0.05. Lehmann (1958) made his suggestion for testing, not to be more coherent than fixed level testing, but rather to avoid a different problem exhibited by fixed level testing. Testing all hypotheses at the same level, regardless of how much data one has, allows the probability of type II error to become much smaller than the probability of type I error as the sample size increases. This amounts to behaving as if the null hypothesis were not very important compared to the alternative. Indeed, the fact that the implied prior goes to zero as the sample size increases reflects this fact. Lehmann's procedure forces the type I and type II errors to decrease together as sample size increases, thereby making sure that both the null and the alternative remain important as the sample size increases. In fact, the implied prior approaches a value strictly between 0 and 1 as sample size increases. What makes Lehmann's procedure less

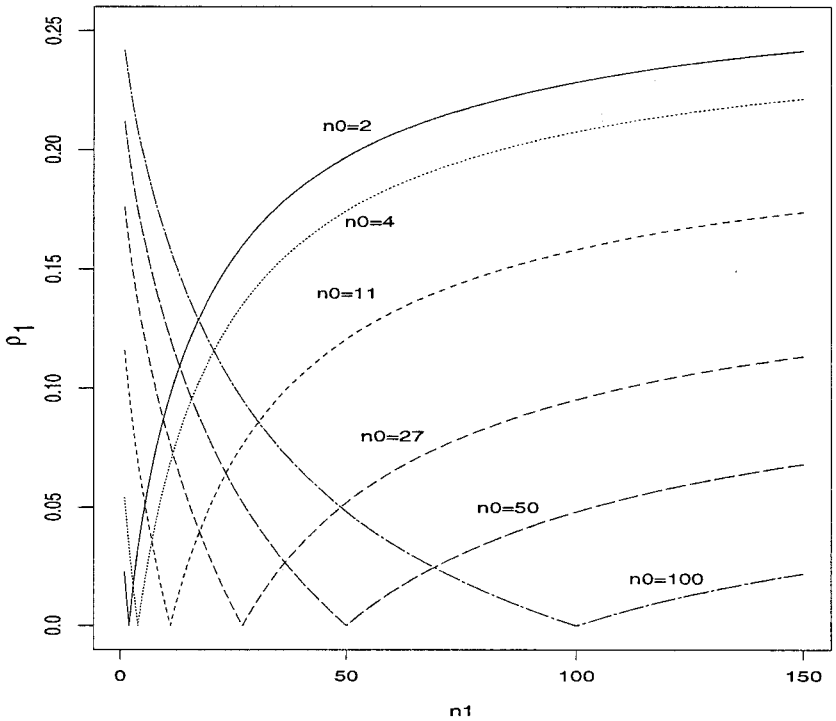


Figure 2. Plot of ρ_1 for minimax rule with $c_0 = 19$ as a function of n_1 (running from 1 to 150) for various values of n_0.

https://doi.org/10.1086/341850 Published online by Cambridge University Press

coherent than fixed level testing is the rate at which the implied prior approaches its limit as sample size increases. For Lehmann's procedure, the implied prior differs from its limit by approximately a constant divided by sample size whereas the implied prior for a fixed level test differs from 0 by approximately $\exp(-cn)$ for some constant c. In this simple testing problem, Lehmann's procedure with $\lambda = 1$ leads to coherent choices of admissible tests for all sample sizes. Lehmann's procedure with $\lambda = 1$ here corresponds to an implied prior for the null hypothesis of $1/(1 + c_0) = 0.05$ when $c_0 = 19$, and an implied prior of 1/2 when losses are equal ($c_0 = 1$). As we noted, Lehmann's rule gives the minimax risk solution for $\lambda = c_0$. However, as Lindley (1972, 14) points out, it is not guaranteed that minimax risk solutions from different families of admissible tests correspond to the same Bayes model. In our testing problem, this is what happens with Lehmann's rule when $\lambda \neq 1$, which explains how it suffers a positive degree of incoherence. An alternative procedure to Lehmann's which also lets type I and type II error probabilities decrease as sample size increases, but which is coherent, is to minimize a positive linear combination of those error probabilities.

5. Summary. In this article we introduce a familiy of indices of incoherence of previsions, based on the gambling framework of de Finetti (1974). When a bookie is incoherent, a gambler can choose a collection of gambles acceptable to the bookie that result in a sure loss to the bookie (and a sure gain to the gambler). That is, the gambler can make a *Dutch Book* against the bookie. Our index of incoherence in the bookie's previsions is the maximum guaranteed rate of loss to the bookie that the gambler creates through his/her choice of coefficients, relative to the bookie's escrow. Throughout, we mean by "escrow" an amount needed to cover the bookie's possible losses as developed in Section 3.

In Section 4, we apply this idea to identify the degrees of incoherence in two policies for testing simple hypotheses. First, we consider testing at a level that is fixed regardless of the sample size, as in the example of Cox (1958). We show, through a trade of risks, how the gambler can make a "Dutch Book" against a statistician who follows such a testing policy. That is, our index of incoherence coincides with the extent to which the fixed alpha level tests can be dominated by combinations of other tests.

When tests are based on small sample sizes, the degree of incoherence in a fixed-level testing policy is complicated, as illustrated in Figure 1. However, the degree of incoherence between two such tests decreases as the sample sizes for these tests increases. Nonetheless, we do not find this fact sufficient to justify the policy, even with large samples, because the statitstician's near-to-coherent behavior then requires treating one of the hypotheses as practically impossible. That is, the Bayes model that the

https://doi.org/10.1086/341850 Published online by Cambridge University Press

fixed level testing policy approaches with increasing sample size assigns probability 0 to the null hypothesis. Why bother to collect data if that is your behavioral policy? Obviously, mere coherence of a policy is not sufficient to make it also a reasonable one!

A second testing policy that we examine is due to Lehmann (1958), who proposes admissible tests based on a fixed ratio of the two risks involved, i.e., with a fixed ratio of type I and type II errors denoted by his parameter λ. Except for the case in which that ratio is 1, this too proves to be an incoherent policy for testing two simple hypotheses. Figure 2 shows the plot of the degree of incoherence for Lehmann's rule ($\lambda = 19$) applied to tests with differing sample sizes. Surprisingly, even in a comparison of two tests based on large sample sizes, Lehmann's policy is sometimes more incoherent by our standards than the fixed .05 level policy for the same two sample sizes. Thus, in order to gain the benefits of approximate coherence, it is neither necessary nor sufficient merely to shrink the level of tests with increasing sample sizes, as happens with Lehmann's rule. In tests based on increasing sample sizes Lehmann's policy (λ fixed) is approximately coherent against a Bayes model that assigns equal prior probability to each of the two hypotheses, the implied priors converge to 1/2. Of course, for that prior, the choice of $\lambda = 1$ in Lehmann's rule assures exact coherence at all sample sizes!

Our work on degrees of incoherence, illustrated here with an analysis of testing simple statistical hypotheses, indicates the importance of having finer distinctions than are provided by de Finetti's dichotomy between coherent and incoherent methods. We see the interesting work of Nau (1989, 1992) providing useful algorithms for computing the rate of guaranteed loss with the escrow used in this paper. (See Nau 1989, 389).

In conclusion, we believe that approaches like Nau's and those we have developed here and in Schervish, Seidenfeld, and Kadane (1997, 1999) permit a more subtle treatment of such longstanding issues as the debate over coherence versus incoherence of some classical statistical practices. That is not the whole problem. Rather, we need to know how far from coherent a particular policy is after we learn that it is incoherent, and learn how it compares with other incoherent methods that have been adopted in practice. We hope to continue this line of investigation in our future work.

REFERENCES

Cox, D. R. (1958), "Some Problems Connected with Statisitical Inference", *The Annals of Mathematical Statistics* 29: 357–363.

de Finetti, B. (1974), *Theory of Probability,* volume 1. New York: John Wiley.

Freedman, D. A. and R. A. Purves (1969), "Bayes' Method for Bookies", *The Annals of Mathematical Statistics* 40: 1177–1186.

Kadane, J. B. and L. Wolfson (1998), "Experiences in Elicitation", (with comments), *Journal of the Royal Statistical Society,* Ser. D (The Statistician) 47: 3–19.

https://doi.org/10.1086/341850 Published online by Cambridge University Press

Lehmann, E. (1958), "Significance Level and Power", *The Annals of Mathematical Statistics* 29: 1167–1176.
Lindley, D. (1972), *Bayesian Statistics A Review.* Philadelphia: SIAM.
Nau, R. F. (1989), "Decision Analysis with Indeterminate or Incoherent Probabilities", *Ann. Oper. Res.* 19: 375–403.
———(1992), "Indeterminate Probabilities on Finite Sets", *The Annals of Statistics* 20: 1737–1767.
Schervish, M. J. (1995), *Theory of Statistics.* New York: Springer-Verlag.
Schervish, M. J., T. Seidenfeld and J. B. Kadane (1997), "Two Measures of Incoherence: How Not to Gamble if You Must," Technical Report 660, Carnegie Mellon University, Statistics.
Schervish, M. J., T. Seidenfeld and J. B. Kadane (1999), "How Incoherent is Fixed-Level Testing," Technical Report 703, Carnegie Mellon University, Statistics.
Seidenfeld, T., M. J. Schervish and J. B. Kadane (1990), "Decisions without Ordering", in W. Sieg, ed. *Acting and Reflecting,* Dordrecht: Kluwer Academic Publishers, 143–170.
Shimony, A. (1955), "Coherence and the Axioms of Probability", *Journal of Symbolic Logic* 20: 1–28.
Walley, P. (1991), *Statistical Reasoning with Imprecise Probabilities.* London: Chapman and Hall.

https://doi.org/10.1086/341850 Published online by Cambridge University Press