Hypothesis Testing, “Dutch Book” Arguments, and Risk

Daniel Malinsky*†

“Dutch Book” arguments and references to gambling theorems are typical in the debate between Bayesians and scientists committed to “classical” statistical methods. These arguments have rarely convinced non-Bayesian scientists to abandon certain conventional practices (like fixed-level null hypothesis significance testing), partially because many scientists feel that gambling theorems have little relevance to their research activities. In other words, scientists “don’t bet.” This article examines one attempt, by Schervish, Seidenfeld, and Kadane, to progress beyond such apparent stalemates by connecting “Dutch Book”–type mathematical results with principles actually endorsed by practicing experimentalists.

1. Introduction. Bayesianism in philosophy of science is most often associated with the “degree of belief” interpretation of probability or with a particular normative theory of belief revision. In statistics, “Bayesian methods” refers to a collection of statistical tools and procedures that stem from a systematic decision-theoretic foundation for statistical practice. According to the Bayesian viewpoint, these methods are preferable alternatives to the ubiquitous “classical” methods popular among experimentalists; Bayesians think that experimentalists ought to interpret their findings in light of the proper methods of belief revision and that they ought to use certain procedures (the Bayesian ones, not the classical ones) in analyzing their data. The normative justification for using Bayesian methods and eschewing (some) classical methods often takes the form of a so-called Dutch Book argument. There are many variants of the Dutch Book argument that I will not get into

*To contact the author, please write to: Department of Philosophy, Baker Hall 135, Carnegie Mellon University, Pittsburgh, PA 15206; e-mail: malinsky@cmu.edu.

†I would like to thank Liam Kofi Bright, David Danks, Greg Gandenberger, Clark Glymour, and Teddy Seidenfeld for enlightening discussions on the topic of this article and comments on earlier drafts.

Philosophy of Science, 82 (December 2015) pp. 917–929. 0031-8248/2015/8205-0016$10.00

here,[1] but generally speaking the Dutch Book argument compels rational agents to conform their degrees of belief over some set of possible events to the Kolmogorov probability axioms (nonnegativity, normalization, and finite additivity). The first clear and explicit examples of such arguments can be found in the work of Frank Ramsey (1926/1931) and Bruno de Finetti (1937/1964). De Finetti shows that if an agent's previsions on some set of possible events do not conform to the Kolmogorov probability axioms, they are *incoherent*, that is, there exists a number of bets, each of which is acceptable to the agent, the combination of which guarantees a sure loss for the agent regardless of the state of the world.[2] But this is a result about the rational behavior, or rational degrees of belief, of an idealized gambling agent, which seems to have little direct connection to decisions made by scientists in experimental settings—what relevance does it have to statistical practice? A statistical procedure is incoherent if the statisticians' implied prior probabilities (i.e., their implied degrees of belief) over the hypothesis space are incoherent. Bayesians sometimes argue that scientists should abandon certain classical statistical procedures, like fixed-level null hypothesis significance testing, on the ground that the procedures are incoherent. In other words, the procedure implies that the scientist's degrees of belief over the possible outcomes of the experiment are incoherent, and rational agents with incoherent beliefs are liable to be made sure losers by a clever gambler. Thus, the argument goes, the scientist should not use this procedure. But the experimentalists studying psychological response theories, or searching for new fundamental particles, are not entering into any bets, so why should they care that their procedures are incoherent? Should the mere fact that a procedure is judged to be incoherent provide sufficient reason for a rational scientist to abandon the procedure and use a coherent one instead? I am using the generic terms "scientist" and "experimentalist" here to refer to any agent who makes decisions about what methods to use in an experimental setting; this can be a laboratory supervisor, a technician, an evaluator at a research funding agency, and so on. I assume the scientist is not already a Bayesian, meaning that scientists are not already convinced that they ought to act as if their prior probabilities conform to the axioms.

I consider this question by looking at a paper by Schervish, Seidenfeld, and Kadane (2002), in which they relate gambling outcomes to the trade of risk functions. In that work, they present a Dutch Book–style mathematical result that could naturally be interpreted as lending normative support

[1] In particular, I will not get into the differences between the synchronic and diachronic Dutch Book arguments.

[2] "Previsions" are associated with the agent's "degrees of belief" over the possible future states of the world. Previsions are not necessarily probabilities in the sense that previsions might violate the probability axioms.

to the Bayesian criticism of experimentalists who engage in fixed-level hypothesis significance testing. A Bayesian statistician might use the result Schervish et al. present in their paper as the basis of a Dutch Book argument against the non-Bayesian scientist; I argue that such an argument should fail to convince a non-Bayesian scientist to change her methods because certain features of the theorem are uninterpretable in the context of actual laboratory procedures.[3] However, I also note that Schervish et al.'s work on the connections between risk and incoherence points to a different criticism of classical statistical practice, one that makes no reference to either gambling or any "nonclassical" statistical concepts. In particular, their work makes it clear that fixed-level hypothesis testing violates the classical statistician's preference for lower-risk functions, at least when that preference is meant to apply to combinations (or averages) of hypothesis tests that might be done over the course of an experiment. This will be made precise below.

2. Background. A Dutch Book theorem (of the synchronic de Finetti/Ramsey type) generally has the following form: "if an agent's previsions violate the probability axioms, then *there exists* a combination of individually acceptable bets such that entering into those bets guarantees a sure loss for the agent, regardless of the state of the world." A Dutch Book argument, then, which compels an agent to conform to the probability axioms, will generally include as a premise that the agent would actually accept each of the individually acceptable bets that make her a sure loser. There are a number of interesting criticisms of Dutch Book arguments in the literature, many of which center on this premise (or something like it) and consider to what extent the idealized gambling model is appropriate for representing agents in the real world. See Kyburg (1978), Glymour (1980, 71–72), and Hájek (2005) for examples of such criticisms and related discussion. In contrast with those authors, I am interested in whether the premises of Dutch Book arguments, which are generally interpreted in terms of gambling or forecasting scenarios, have plausible analogues in the context of statistical decisions a scientist might make over the duration of an experiment. If they do, then Dutch Book arguments may provide compelling reasons for a scientist to abandon incoherent methods and adopt Bayesian methods instead. If they do not, then Dutch Book arguments seem to be of little relevance to experimentalists. Of course, in thinking that this is a worthwhile question to ask I am implicitly taking a controversial stance on the issue of what can constitute a "compelling reason." The mere fact that an agent is

[3] It should be mentioned that Schervish et al. (2002) do not present such an argument and never claim that their primary aim is to compel scientists to change their methods. Rather, Schervish et al.'s aim in that paper relates to quantifying and measuring degrees of incoherence; this will be explicated in sec. 3.

incoherent—that is, the fact that *there exists* a set of gambles with such-and-such properties—does not by itself compel an agent to revise her beliefs or change her methods. There are compelling reasons to change methods if and only if the actual commitments and preferences of the agent, when taken seriously, cause the agent to engage in self-defeating behavior. There are some Bayesians who might disagree; for them, the fact that an agent's beliefs are incoherent constitutes a compelling reason to revise them (e.g., Christensen 1996). I think this is a misguided view, but for considerations of space I will not elaborate on my reasons here.

The following discussion will focus on a particularly controversial but common statistical practice: fixed-level null hypothesis significance testing. Most of the Bayesian criticisms leveled at the practice of hypothesis testing involve consideration of prior probabilities or mixed tests. I eschew discussion of these concepts here because many practicing scientists who have been trained in classical methods claim they do not make use of these concepts. What principles do scientists trained in the classical paradigm actually endorse, then, when it comes to null hypothesis significance testing?

Suppose the experimentalist is interested in testing a simple null hypothesis H_0 against a simple alternative H_1. Let the independent and identically distributed random variables $X_1, \ldots, X_n \sim N(\theta, \sigma^2)$ with known σ^2 and consider $H_0 : \theta = 0$ versus $H_1 : \theta = 1$. After observing $X_1, \ldots, X_n$ the experimentalist either decides to reject the null hypothesis or fails to reject the null hypothesis.[4] The negative consequences of making such a decision d can be summarized in the loss function $L(\theta, d)$:

$$L(\theta,\ d) = \begin{cases} c_0 & \text{if } q = 0 \text{ and } d = \text{reject } H_0, \\ c_1 & \text{if } q = 1 \text{ and } d = \text{fail to reject } H_0, \\ 0 & \text{otherwise,} \end{cases} \tag{1}$$

with $c_0, c_1 > 0$. Following Schervish et al. (2002) one can set $c_1 = 1$ and define the risk function for a sequence of decisions δ (also called a *decision rule*) as the expected loss:

$$R(\theta,\ \delta) = \mathbf{E}[L(\theta,\ d)] \begin{cases} c_0\alpha & \text{if } q = 0, \\ \beta & \text{if } q = 1, \end{cases} \tag{2}$$

where α is the probability of type I error (or the *size* of the test) and β is the probability of a type II error ($1 - \beta$ is the *power* of the test). In general the experimentalist aims to minimize the risk function by choosing δ appro-

[4] If we were considering an alternative hypothesis like $\theta \neq 0$ or $\theta > 0$ (or if we left the alternative "unspecified"), this would bring us into the realm of composite hypothesis testing, which introduces additional complications. I will not deal with composite testing in this article.

priately. In a classical hypothesis-testing scenario it is standard to accept the following two principles:[5]

> P1. At a given sample size n, if $R(\theta, \delta_1) \leq R(\theta, \delta_2)$ for all θ, and the inequality is strict for some θ, prefer δ_1.
>
> P2. At a given sample size n, prefer the $\alpha = .05$ test from among the class of admissible universally most powerful tests.

P2 is not properly a principle, but rather it is an extension of the Neyman-Pearson theory of hypothesis testing, and the .05 level is a convention adopted by a wide range of practitioners. In some situations α is fixed at a lower conventional level, like .01. For a given sample size, the Neyman-Pearson theory defines an admissible set of universally most powerful (α, β) pairs, where the risk function of a particular (α, β) pair is not dominated by any other; that is, each test is constructed to be the most powerful (lowest β) for a given α. Fixing the α value amounts to preferentially controlling the probability of type I error.

In section 5, I discuss the circumstances under which P1 and P2 might be mutually unsatisfiable. First, let us examine Schervish et al.'s (2002) analysis.

3. Gambling and Risk Functions. Schervish et al.'s (2002) stated aim is to provide a way of moving beyond the binary judgment of the coherence or incoherence of a particular method; coherence can come in degrees, and it is particularly useful to have a measure of how much bookies stand to lose when they expose themselves to sure loss in de Finetti's sense.[6] The "rate of incoherence" then quantifies how badly the bookie is sure to do in the worst case, the case in which her incoherent previsions are maximally exploited by some clever gambler (or market of gamblers). With such a measure, the bookie can make a more informed decision about how or whether to become coherent, since becoming coherent is often "expensive" to the bookie. Bookies are assumed to have finite resources, so one way of assessing their incoherence is by examining the fraction of their resources they stand to lose (in the worst case) with their stated previsions. Much of Schervish et al.'s paper is devoted to working out the technical details of this normalization.

Following de Finetti, Schervish et al. (2002) consider gambling payoffs of the form $\alpha(X - x)$, where X is a random variable, α is the number of

[5] See Lehmann and Romano (2005, 9–10) for a statement similar to P1.

[6] In some discussions, it is the "gambler" and not the "bookie" who stands to lose money. I am following Schervish et al.'s terminology here.

“units” of the bet up for sale, and x is a prevision for the random variable X. To be more precise, the bookie can actually announce *upper* previsions x for X by indicating that all bets with payoff $\alpha(X - y)$, where $\alpha < 0$, $y > x$, are acceptable gambles, or she can announce *lower* previsions x for X by indicating that all bets with payoff $\alpha(X - y)$, where $\alpha > 0$, $y < x$, are acceptable gambles. The bookie can announce only upper previsions or only lower previsions or both. The random variable X might be the indicator function I_A on event A, which is 1 when A occurs and 0 when A does not occur. Schervish et al. (2002, S250) state the following definition of incoherence:

> 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
>
> $$\sup_{t\varepsilon T} \sum_{i=1}^{n} \alpha_i(X_i(t) - y_i) < -\varepsilon, \qquad [3]$$
>
> in which case we say that a *Dutch Book* has been made against the bookie.

The ε in equation (3) is the total guaranteed loss that the bookies face when they accept the set of gambles, but it is not by itself an adequate measure of the degree of incoherence because it is dependent on the α_i (the number of “units” bought or sold). Schervish et al. (2002) want a measure that reflects the stated previsions, not the number of gambles undertaken. To accomplish this, they characterize the bookies’ escrow, that is, the quantity that the bookies can be expected to set aside or commit themselves to betting with when entering into a gambling contract. They subsequently define the bookies’ “maximum rate of guaranteed loss” with respect to this escrow. The mathematical details do not affect the discussion in this essay, but the reader should consult Schervish et al. (2002) for the full technical presentation.

Schervish et al. (2002) apply their measure of incoherence to the case of fixed-level hypothesis testing. They interpret an agent’s preference for a test δ_1 over test δ_2 to mean that the agent would prefer to suffer the loss from δ_1 rather than the loss from δ_2, that is, the agent would rather “pay out” $R(\theta, \delta_1)$ than $R(\theta, \delta_2)$. They call $R(\theta, \delta_2) - R(\theta, \delta_1)$ an acceptable gamble. Schervish et al. use $\alpha_\delta(\sigma)$ and $\beta_\delta(\sigma)$ to denote the size and power of test δ conditional on σ (equivalently, conditional on sample size n). The conventional .05 test is δ_{CL}, with $\alpha_{\delta_{CL}}(\sigma) = .05$ and $\beta_{\delta_{CL}}(\sigma)$ equal to whatever is the highest power possible under the Neyman-Pearson theory for that σ. Schervish et al. show that the following gamble is acceptable to the scientist who prefers the .05-level test:

$$R(\theta, \delta) - R(\theta, \delta_{CL}) = \begin{cases} (\alpha_\delta(\sigma) - .05)c_0 & \text{if } \theta = 0, \\ \beta_{\delta_{CL}}(\sigma) - \beta_\delta(\sigma) & \text{if } \theta = 1, \end{cases} \tag{4}$$
$$= a(I_A - b),$$

where

$$\begin{aligned} A &= \{\theta = 0\}, \\ a &= (\alpha_\delta(\sigma) - .05)c_0 + \beta_\delta(\sigma) - \beta_{\delta_{CL}}(\sigma), \\ b &= \frac{\beta_\delta(\sigma) - \beta_{\delta_{CL}}(\sigma)}{(\alpha_\delta(\sigma) - .05)c_0 + \beta_\delta(\sigma) - \beta_{\delta_{CL}}(\sigma)}. \end{aligned} \tag{5}$$

So, the preference for test δ_{CL} amounts to partaking in a "trade" of risk functions that has the form of a de Finetti gamble on an indicator function taking the value 0 when $\theta = 1$ and 1 when $\theta = 0$. The constants a and b are meant to be uncontroversial combinations of sizes and powers—purely classical statistical concepts with no explicit dependence on subjective prior probabilities.

Schervish et al. (2002) prove a theorem (their theorem 2) that applies to arbitrary decision rules. Rather than introduce the notation required to express this more general result, I will state the analogous result for the special case we have been considering: if δ_0 and δ_1 are not Bayes rules with respect to a common prior,[7] then there exist real numbers d_0 and d_1 and alternative decision rules δ_0^* and δ_1^* such that the two gambles $d_0a_0(\delta_0^*)(I_A - b_0(\delta_0^*))$ and $d_1a_1(\delta_1^*)(I_A - b_1(\delta_1^*))$ are both acceptable, but

$$d_0a_0(\delta_0^*)(I_A - b_0(\delta_0^*)) + d_1a_1(\delta_1^*)(I_A - b_1(\delta_1^*)) < 0, \tag{6}$$

where $a_0(\delta_0^*)$, $b_0(\delta_0^*)$, $a_1(\delta_1^*)$, $b_1(\delta_1^*)$ are the generalized versions of constants a and b above (with respect to the alternative decision rules δ_0^*, δ_1^*). This demonstrates a kind of incoherence, because a combination of acceptable gambles yield a (sure) negative payoff for the bookie. And, since the conventional .05-level testing procedure is not a Bayes rule, that means that using such a rule is equivalent to announcing previsions that expose the scientist to a sure loss. In the next section we ask the question of how this result should be properly interpreted in an experimental setting.

4. The Dutch Book Result. Let us take a closer look at the mathematics, starting with the acceptable gamble $R(\theta, \delta) - R(\theta, \delta_{CL}) = a(I_A - b)$ in equa-

[7] A Bayes rule is a decision rule that minimizes a quantity called the Bayes risk. For current purposes, the formal definition is not crucial. What matters is that an incoherent decision rule (such as the one under discussion, i.e., testing at $\alpha = .05$ regardless of sample size) is not a Bayes rule.

tion (4). In the context of gambling, the constants a and b have unproblematic standard interpretations: b is the prevision on random variable I_A and a is the number of "units" of the gamble bought or sold (in this case, the gamble is presumed to be one-sided, so the sign of a can be fixed appropriately). In the context of statistical decision making, the constants a and b have unproblematic interpretations in terms of combinations of type I and type II error probabilities, as in equation (5). But in the theorem stated by Schervish et al. (2002), the acceptable gambles are $d_0 a_0(\delta_0^*)(I_A - b_0(\delta_0^*))$ and $d_1 a_1(\delta_1^*)(I_A - b_1(\delta_1^*))$. In the gambling context, the real numbers d_0 and d_1 are again unproblematic: it is presumed that if a bookie is willing to accept gambles on I_A with prevision b, gambles of the form $w(I_A - b)$ are acceptable for any real-valued w, so long as the total number of "units" does not exceed the bookie's available resources. In this case, the entire product $d_0 a_0(\delta_0^*)$ acts as the number of "units" bought or sold (of that betting contract). But what is the intended interpretation of the real-valued coefficients d_0 and d_1 in the context of scientific decision making? Maybe looking at equation (6) can provide a hint.

Equation (6) can be rewritten in terms of risk functions:[8]

$$d_0(R(\theta, \delta_0^*) - R(\theta, \delta_{CL})) + d_1(R(\theta, \delta_1^*) - R(\theta, \delta_{CL})) < 0, \tag{7}$$

and can be rearranged to take the form

$$d_0 R(\theta, \delta_0^*) + d_1 R(\theta, \delta_1^*) - (d_0 + d_1)R(\theta, \delta_{CL}) < 0. \tag{8}$$

Again, the theorem states that so long as δ_{CL} is a non-Bayes rule (which it is), there exist real numbers d_0, d_1 and rules δ_0^*, δ_1^* such that the inequality in (8) holds. Is this a compelling reason for a practitioner to abandon the method that uses δ_{CL}? This depends on whether the mathematical statement is interpretable in terms relevant to laboratory practice. A few interpretations seem ostensibly plausible.

One might think, in analogy with taking a real number multiple of "units" of a bet, that the coefficients d_0 and d_1 represent performing the respective tests d_0 or d_1 many times. But this cannot be the case, because the scientist is interested in performing the test one time for each sample size; assuming she takes "accept or reject" decisions seriously, she has nothing to gain by repeating the same test on the same data. Furthermore, it is not clear what it would mean to perform a test a non–whole number of times. Alternatively, the coefficients could represent the probability of performing the associated test, if $d_0, d_1 > 0$ and $(d_0 + d_1) = 1$ (think of flipping a coin to determine whether test δ_0^* or δ_1^* is performed). This does not seem to be the authors' intended interpretation because it would make the normative force

[8] Here I am stating everything with respect to the non-Bayes rule δ_{CL} instead of two arbitrary non-Bayes rules.

of their theorem dependent on the normative force of the controversial ancillarity principle (also known as the conditionality principle) in the presence of mixed tests.[9] That fixed-level hypothesis testing violates dominance in the presence of mixed tests has been known at least since Cox (1958). But the appropriateness of the ancillarity principle and the prescriptions for hypothesis testing in the presence of mixed tests is much debated among statisticians and philosophers of statistics, so it would add no new normative force to recast the issue in the form of a Dutch Book theorem. Furthermore, the theorem as stated by Schervish et al. (2002) would need to be constrained to reflect the fact that d_0 and d_1 are meant to be probabilities and not just some real-valued numbers.

It is possible that the coefficients d_0 and d_1 are not meant to have any special interpretation in the context of laboratory procedures that a scientist might actually carry out. The coefficients might simply be mathematical devices necessary for the formalism that calculates rates of incoherence for the implied priors that one can extract from the use of these testing procedures. This is perfectly reasonable for the purpose of demonstrating the rate of incoherence of the procedure, but it has the unfortunate consequence of making the Dutch Book result uninterpretable in the context of laboratory practice and thus not part of a compelling argument to change methods for any scientist who is not already a Bayesian.

I should note that nothing in the above analysis is meant to preclude the possibility that some other Dutch Book–style result will meet the desideratum I outlined, that is, that the book-making strategy has plausible interpretation in terms of statistical procedures an experimentalist might actually carry out in a laboratory setting. In particular, alternative measures of rates of incoherence might suggest strategies that do not require trades of risk functions to be multiplied by real-valued coefficients; this will depend on the mathematical details of the measure of incoherence (e.g., how the escrow function is defined). I am not aware of any measure that identifies de Finetti gambles with trades of risk functions but that does not require the manipulation of real-valued coefficients in order to guarantee a sure loss to the bookie.

5. Combined (or Average) Risk Functions. One corollary to Schervish et al.'s (2002) result, which is apparent when their result is stated in the form of equation (8), is that for any non-Bayes decision rule like δ_{CL}, there exist

[9] The ancillarity principle states, roughly, that the evidential significance of a statistical test should be calculated conditional on an ancillary statistic, i.e., in the coin-flipping scenario, the evidential significance should depend only on which test was actually performed, not on the probabilities of performing either test. So the coin-flipping procedure should be ignored.

two alternative decision rules δ_0^* and δ_1^* such that the combined risk function of performing δ_0^* on sample n_0 and δ_1^* on sample n_1 will be lower than the combined risk function of performing δ_{CL} on sample n_0 and sample n_1. A *combined risk function* is just the sum of two risk functions for tests performed on two different samples. So, if we index the risk function associated with running test δ_{CL} on a sample n_0 as $R_{n_0}(\theta,\ \delta_{CL})$, then the combined risk function for samples n_0 and n_1 is $R_{\text{comb}}(\theta,\ \delta_{CL}) = R_{n_0}(\theta,\ \delta_{CL}) + R_{n_1}(\theta,\ \delta_{CL})$. More generally one can define the combined risk for a sequence of N tests: $R_{\text{comb}}(\theta,\ \delta) = \sum_i^N R_{n_i}(\theta,\ \delta)$. The exact same could be said replacing combined risk functions with average risk functions, which might be more natural to consider in certain contexts: $R_{\text{avg}}(\theta,\ \delta) = (1/N)R_{\text{comb}}(\theta,\ \delta)$. So, if we extend what I called principle P1 in section 2 to cover combined (or average) risk functions, we now see that the $\alpha = .05$ convention (P2) clashes with another purely classical statistical principle (P1*):

> P1*. For a sequence of samples n_0, n_1, and so on, if $R_{\text{comb}}(\theta,\ \delta_1) \leq R_{\text{comb}}(\theta,\ \delta_2)$ for all θ, and the inequality is strict for some θ, prefer δ_1.

The $\alpha = .05$ convention, when applied to tests at multiple sample sizes, is tantamount to a preference for dominated combined risk functions.

There is nothing mathematically novel in this claim. In fact, the same can be gleaned from discussions in Seidenfeld, Schervish, and Kadane (1990) and Berry and Viele (2008), both of which draw on Cox (1958). The difference is that these discussions employ the terminology of mixed tests, and, as mentioned above, the presence of mixed tests brings to the fore the somewhat controversial ancillarity principle. Since statisticians disagree about the validity of this principle (e.g., Helland 1995), it would be beneficial for practitioners to be able to get some insight into the apparent conflict between the Bayesian and classical recommendations with minimal foray into philosophically troubled waters. In another paper, Schervish, Seidenfeld, and Kadane (2009, 218) prove a theorem (theorem 7) about the risk inadmissibility of non-Bayes decision rules when randomized decision rules are available. Randomized decision rules can be thought of as similar to mixed tests, as discussed above (i.e., the decision maker chooses which test to perform by some random process). Whether a randomized decision rule is in fact available or appealing to the practicing scientist depends on a number of considerations I do not have the space to discuss here. The formulation in terms of combined risk is, I think, advantageous because it does not appeal to any concepts or procedures that might be controversial by the lights of the experimentalist trained in the classical paradigm.

Berry and Viele (2008) show that one can calculate (for the normal distribution) a coherent decision rule in which α is a function of sample size. In summary: a coherent decision rule for null hypothesis testing requires that α

decrease as the sample size increases. The precise functional dependence of α on sample size depends on a number of details, but for simple versus simple the prescription is not complicated. The recipe for compound testing is more involved and requires some perhaps controversial decisions on behalf of the scientist. An interesting implication of this result is that the experimentalist who chooses values of α consistent with Berry and Viele's prescription can arrive at combined risk functions that have both lower α and lower β than δ_{CL}. In other words, we can find combinations of tests that are strictly better than δ_{CL} with respect to combined risk.

6. Future Work: Planning Particle Physics Experiments. While journals in fields like psychology have at times enforced something like the $\alpha = .05$ or .01 convention as a prerequisite for publication, contemporary experimental particle physics has an (unwritten) analogous convention in the so-called 5σ criterion. The 5σ value amounts to an α of approximately 3×10^{-7} (see James 2006, 321). Physics journals will generally only publish claims of "observation of" (as opposed to "evidence for" or something like it) new particles if the null hypothesis—which coincides with a "no new particle" model—is rejected at this extremely low significance level. For the history of this convention and its use, see Franklin (2013). The common justification for the 5σ criterion is a high cost associated with false positives in the discovery of new particles; particle discovery claims at the 3σ and 4σ levels have been frequently overturned by more data and repeated experiments.

The long lifetimes of typical experiments in contemporary particle physics provide plausible contexts for evaluation in terms of average or combined risk functions. The ATLAS and CMS experiments at the Large Hadron Collider, for example, began taking data in 2009 and will continue to do so for several years. In July 2012, ATLAS and CMS both announced the observation of a particle resembling the long-sought-after Higgs boson. The null hypothesis, which consisted of an expected distribution of particle masses that did not include the Higgs, was rejected at the 5σ level in both experiments (see ATLAS Collaboration 2012; CMS Collaboration 2012). Both experiments continue to take data and have presented updated results with much larger data sets, again rejecting the null hypothesis but now (as of mid-2014) at levels well above 5σ. Fortunately, as the experiments collect more data, they report the highest significance level at which they can reject the null hypothesis, and not only the fact that the hypothesis remains rejected. Yet it remains true that many journals in the field do not accept "observation" results below the 5σ level, regardless of sample size. This convention is suboptimal with respect to combined risk. Physicists could do better by adopting a coherent decision rule, and this does not require specifying prior probabilities over competing hypotheses. In future work, I hope to include suggestions for adopting a coherent decision rule, taking the details and complexities of particle physics data into account.

7. Conclusion. Dutch Book arguments in themselves seem to be unconvincing to non-Bayesian experimentalists because they are not relevant (from the perspective of a non-Bayesian) to decisions actually made in laboratories. But luckily, the results of Schervish et al. (2002) point to a way to move forward. By changing the language and focus of the discussion, I have shown how two commonly espoused principles of classical hypothesis testing—the preference for lower risk functions and the convention of using a fixed α-level test at all sample sizes—conflict when the former principle is extended to include combined or average risk. No considerations of mixing tests, ancillarity, or prior probabilities played a key role in identifying this problem or in the proposed solution. The context of contemporary particle physics is offered as a motivating example for the plausibility of considering combined (or average) risk functions in planning experiments. Experimentalists have compelling reasons to take steps analogous to those advocated by Berry and Viele (2008) in order to determine an optimal way to tie the choice of α values to sample size. Thus, we need not find ourselves at an impasse when the experimentalist insists she “doesn’t bet.” We can use the results of Bayesian analysis to illuminate the circumstances in which classical statistical commitments conflict with each other.

REFERENCES

ATLAS Collaboration. 2012. “Observation of a New Particle in the Search for the Standard Model Higgs Boson with the ATLAS Detector at the LHC.” *Physics Letters* B 716 (1): 1–29.

Berry, Scott, and Kert Viele. 2008. “A Note on Hypothesis Testing with Random Sample Sizes and Its Relationship with Bayes Factors.” *Journal of Data Science* 6:75–87.

Christensen, David. 1996. “Dutch-Book Arguments Depragmatized: Epistemic Consistency for Partial Believers.” *Journal of Philosophy* 93 (9): 450–79.

CMS Collaboration. 2012. “Observation of a New Boson at a Mass of 125 GeV with the CMS Experiment at the LHC.” *Physics Letters* B 716 (1): 30–61.

Cox, David R. 1958. “Some Problems Connected with Statistical Inference.” *Annals of Mathematical Statistics* 29:357–63.

de Finetti, Bruno. 1937/1964. *Foresight: Its Logical Laws, Its Subjective Sources*. Trans. H. E. Kyburg Jr. In *Studies in Subjective Probability*, ed. H. E. Kyburg Jr. and H. E. Smokier, 93–158. New York: Wiley.

Franklin, Allan. 2013. *Shifting Standards: Experiments in Particle Physics in the 20th Century.* Pittsburgh: University of Pittsburgh Press.

Glymour, Clark. 1980. *Theory and Evidence*. Princeton, NJ: Princeton University Press.

Hájek, Allan. 2005. “Scotching Dutch Books?” *Philosophical Perspectives* 19:139–51.

Helland, Inge S. 1995. “Simple Counterexamples against the Conditionality Principle.” *American Statistician* 49 (4): 351–56.

James, Frederick. 2006. *Statistical Methods in Experimental Physics*. 2nd ed. Hackensack, NJ: World Scientific.

Kyburg, Henry. 1978. “Subjective Probability: Criticisms, Reflections, and Problems.” *Journal of Philosophical Logic* 7 (1): 157–80.

Lehmann, Erich, and Joseph P. Romano. 2005. *Testing Statistical Hypotheses*. 3rd ed. New York: Springer.

Ramsey, Frank P. 1926/1931. “Truth and Probability.” In *Foundations of Mathematics and Other Logical Essays*, ed. R. B. Braithwaite, 156–98. London: Routledge & Kegan Paul.

Schervish, Mark J., Teddy Seidenfeld, and Joseph B. Kadane. 2002. “A Rate of Incoherence Applied to Fixed-Level Testing.” *Philosophy of Science* 69 (Proceedings): S248–S264.

———. 2009. “Proper Scoring Rules, Dominated Forecasts, and Coherence.” *Decision Analysis* 6 (4): 202–21.

Seidenfeld, Teddy, Mark J. Schervish, and Joseph B. Kadane. 1990. “Decisions without Ordering.” In *Acting and Reflecting*, ed. W. Sieg, 143–70. Dordrecht: Kluwer.