Stopping rules matter to Bayesians too

April 15, 2010

Abstract

This paper considers a key point of contention between classical and Bayesian statistics—the issue of *stopping rules*, or more generally, *outcome spaces*, and their influence on statistical analysis. Firstly, a working definition of classical and Bayesian *statistical tests* is given, which makes clear that i) once a test has been conducted and an outcome recorded, only the classical approach to inference depends on the full outcome space for the test, and ii) full outcome spaces are nevertheless relevant to *both* the classical and Bayesian approaches, when it comes to planning/choosing a test. The latter commonality between the approaches undermines at least one argument against classical statistics. But it also undermines what may have been a compelling argument against the Bayesian approach—the Bayesian indifference to persistent experimenters and their *optional stopping rules*. Indeed, the final section of the paper offers three Bayesian error theories for the pro-classical 'optional stopping intuition'.

1 Introduction

This paper engages in an old dispute within statistical analysis: a dispute about the relative merits of the classical and Bayesian approaches. It should be made clear from the outset, however, that there is no attempt to survey/tackle this whole debate. The field of statistics is large, and includes a variety of problems of differing complexity. Moreover, it is not clear what *are* the full set of defining features of either the classical or Bayesian methods. The aim of the paper, is, rather, to consider a basic point of contention between the two approaches, in a restricted setting in which classical statistics arguably looks its best.

The following section (Section 2) describes the domain of interest—dichotomous tests of two *simple* statistical hypotheses—and provides a working definition of classical and Bayesian versions of such tests. The question is whether, in this domain at least, the classical approach has merit, despite lacking the generality and the neat connection to a theory of practical decision that is enjoyed by the Bayesian model. Sections 2 and 3 introduce the main issue of the paper that bears on this question—the role of outcome spaces, here examined in terms of *stopping rules*, for both classical and Bayesian statistical tests.

Bayesians criticize classical statistics for the fact that inference depends not just on what happened, but what might otherwise have happened. Arguments to this effect require some nuance, however, since the full outcome space is important in Bayesian statistics too, during experimental design. Section 4 elaborates

on this point. Furthermore, the classical statistician has an interesting positive argument for their approach. There is at least one kind of stopping rule—the stopping rule of the 'persistent experimenter'—that suggests full outcome spaces deserve a larger role in inference than what the Bayesian model affords. The details of the persistent-experimenter case, and what we should think about our intuitions regarding such tests, are addressed in the final two sections of the paper.

2 Classical and Bayesian statistical tests

Let us initially define both a classical and a Bayesian statistical test, for the case where we are comparing two disjoint and exhaustive *simple* hypotheses. (What is meant by a 'simple' hypothesis will be discussed shortly.) Although differences between the tests will become apparent, one of the aims of this section is to highlight the extent to which classical and Bayesian methods are similar.

By way of example, imagine there are just two possibilities regarding the chance of heads for a coin: hypothesis 1 (H_1) is that the chance of heads is 0.5, and hypothesis 2 (H_2) is that the chance of heads is 0.6:

$H_1 : ch = 0.5$
$H_2 : ch = 0.6$

Our example experiment involves tossing the coin 20 times, such that the outcome space is the set $X = \{H, T\}^{20}$, with H designating a result of heads on a single toss, and T a result of tails. Each hypothesis specifies a probability distribution over the outcome space, known as the *likelihood* function for that hypothesis. This is background knowledge and is not itself under test, i.e., we assume a probability model Pr for which the likelihood $Pr(x|H_i)$ is specified for all outcomes x in X and all hypotheses H_i under consideration (here just H_1 and H_2).[1] For our coin example, the probability distributions over X are uncontroversial—the coin tosses are independent and on each toss, there is an identical probability distribution over $\{H, T\}$, as specified by the chance hypotheses.[2] H_1 and H_2 are *simple* hypotheses, at least in this context, because they specify a precise probability distribution over the outcome space, as opposed to *composite* hypotheses which specify a set of possible probability distributions over the outcome space. This paper is confined to the comparison of two simple hypotheses.

Different experiments may yield different outcome spaces and corresponding likelihood functions for the hypotheses under consideration. The *stopping rule* of an experiment has an obvious effect on the outcome space. For instance, instead of tossing a coin a fixed number of times, the stopping rule might specify

[1] Classical statistics does not recognize probabilities for hypotheses, so in this setting the likelihoods $Pr(x|H_i)$ are not equivalent to the familiar ratio of unconditional probabilities. The likelihoods must instead be primitive, but this is a detail that need not concern us.

[2] Note that the outcomes of an experiment are often summarized in terms of a *test statistic*, for example, '4 heads in 20 tosses' as opposed to a precise description of the ordered sequence of tosses. I will not use test statistics in this paper, however, nor will I discuss the criteria for determining what is an appropriate test statistic.

tossing the coin until 6 tails are recorded. The outcome space for this experiment includes samples of all sizes that have a tail on the last toss and 6 tails in total in the full sequence. This clearly differs from the outcome space for the fixed-20-toss experiment. In what follows, we focus on stopping rules and how they matter to the Bayesian versus the classical statistician. This is just one way of expressing the more general issue, however, which is the role of full outcome spaces and likelihood functions for the two statistics camps. Outcome spaces can differ between experiments for all sorts of reasons, for instance, the way the outcomes are measured by the equipment and interpreted by the experimenter (see Edwards 1972 for the classic case of the Voltmeter).

Given a particular experimental outcome, $x \in X$, we want to make an assessment of the hypotheses, where the possible assessments are defined by the set D. In the usual classical hypothesis-testing scenario there are only two possibilities for this final assessment: either H_1 is accepted and H_2 rejected, or H_1 is rejected and H_2 accepted. We can denote these possibilities a and r respectively, so $D = \{a, r\}$.[3] Bayesian tests can also be conceived in this way, as will become clear shortly. That is, the following is a general characterization of a statistical test:

For a given outcome space X, a statistical test (t_X) is a function from the test outcome to an assessment of the hypotheses:

$t_X : X \rightarrow D$

In effect, a statistical test with $D = \{a, r\}$ specifies which subset of experimental outcomes, $R \subset X$, lead to the rejection of H_1.[4] Of course, statistical tests are not foolproof; there is always the possibility of making the wrong judgment. We call the probability of falsely rejecting H_1, or in other words, the probability of rejecting H_1 when it is in fact true, the type I error (α) for the test. (This is just the total probability, given H_1, of the outcomes in the subset R.) Likewise, the probability of accepting H_1 when H_2 is in fact true is the type II error (β).

The similarity between classical and Bayesian tests runs even deeper: both obey Neyman and Pearson's requirement that a test have minimal β, given α and the sample space X (see Hacking 1965, p. 93). A test satisfies the Neyman-Pearson condition just in case the rejection class R consists of those outcomes that are most 'distant' from H_1, where outcome x is more distant from H_1 than outcome x' if and only if the inverse of the likelihood ratio for x is greater. In other words, both Bayesian and classical tests can be written in the following form, where t is a threshold value greater than zero:

$$t_X(x) \quad = \quad \begin{cases} a & \text{if } \frac{Pr(x|H_2)}{Pr(x|H_1)} \leq t; \\ r & \text{if } \frac{Pr(x|H_2)}{Pr(x|H_1)} > t \end{cases}$$

[3] The set of possible hypothesis assessments, D, might be something other than $\{a, r\}$, but this paper considers only the simple accept/reject scenario.

[4] This is the terminology of Hacking (1965 p. 78). The subset of outcomes that lead to the rejection of H_1 is also referred to as the *critical region* (see Kendall and Stuart 1991 p. 795).

The difference between the classical and Bayesian tests lies in the way the threshold value t, and thus the rejection subset of outcomes, R, is determined. For both statistical camps, t and R presumably depend on the context, that is, the epistemic situation and the consequences of making right and wrong decisions. For the Bayesian, however, this context fixes t, and depending on the precise experiment/outcome space, we can then derive R and the error probabilities α and β. For the classical statistician, the situation is reversed: the context fixes a term involving error probabilities, commonly α, and for the particular experiment/outcome space at hand, R and t can then be derived. Our question in subsequent sections is whether the classical approach has some merit.

First a few more details regarding classical and Bayesian tests. The hypothesis assessments in D are considered actions, and so the Bayesian will 'accept' H_1 just in case this act has higher expected utility than 'rejecting' H_1.[5]

Table 1

	H_1 true	H_2 true
Accept H_1 (a)	u_{a1}	u_{a2}
Reject H_1 (r)	u_{r1}	u_{r2}

To calculate the expected utility of the acts according to Table 1, we need to know the new (posterior) probabilities for the hypotheses, p'_1 and p'_2, given the outcome of the test, x. These are calculated using Bayes' formula, and we get the following characterisation of a Bayesian test:

$$t\_Bayes_X(x) = \begin{cases} a & \text{if } p'_1(x)\cdot u_{a1} + p'_2(x)\cdot u_{a2} \geq p'_1(x)\cdot u_{r1} + p'_2(x)\cdot u_{r2}; \\ r & \text{if } p'_1(x)\cdot u_{a1} + p'_2(x)\cdot u_{a2} < p'_1(x)\cdot u_{r1} + p'_2(x)\cdot u_{r2} \end{cases}$$

$$= \begin{cases} a & \text{if } \frac{Pr(x|H_1)\cdot p_1}{Pr(x)}\cdot u_{a1} + \frac{Pr(x|H_2)\cdot p_2}{Pr(x)}\cdot u_{a2} \geq \frac{Pr(x|H_1)\cdot p_1}{Pr(x)}\cdot u_{r1} + \frac{Pr(x|H_2)\cdot p_2}{Pr(x)}\cdot u_{r2}; \\ r & \text{if } \frac{Pr(x|H_1)\cdot p_1}{Pr(x)}\cdot u_{a1} + \frac{Pr(x|H_2)\cdot p_2}{Pr(x)}\cdot u_{a2} < \frac{Pr(x|H_1)\cdot p_1}{Pr(x)}\cdot u_{r1} + \frac{Pr(x|H_2)\cdot p_2}{Pr(x)}\cdot u_{r2} \end{cases}$$

$$= \begin{cases} a & \text{if } \frac{Pr(x|H_2)}{Pr(x|H_1)} \leq \frac{p_1\cdot(u_{a1}-u_{r1})}{p_2\cdot(u_{r2}-u_{a2})}; \\ r & \text{if } \frac{Pr(x|H_2)}{Pr(x|H_1)} > \frac{p_1\cdot(u_{a1}-u_{r1})}{p_2\cdot(u_{r2}-u_{a2})}. \end{cases}$$

We see from the final formulation above that the threshold value t depends on the prior probabilities of the hypotheses and the decision utilities, that is, it depends explicitly on the epistemic and pragmatic context. Clearly t does not change across different experiments with different outcome spaces.

[5] The act of 'accepting' versus 'rejecting' a hypothesis may mean different things in different contexts. For instance, an academic group 'accepting' a claim in theoretical physics may amount to reporting it in an elementary text book, while government scientists 'accepting' that a food is safe may mean permitting its general distribution.

Classical tests are more difficult to define because the term 'classical' refers to a collection of ideas/principles, notably Fisher's and Neyman and Pearson's. The typical sort of classical test fixes α, otherwise known as the *significance level*.[6] Given a particular outcome space X and a choice of α,[7] the rejection region satisfying Neyman and Pearson's condition can be determined, as well as the threshold t for defining the test in the manner of $t_X(x)$ above. It is more usual, however, to define the classical test in terms of *p-values*. The *p-value* of an experimental outcome $x \in X$ is the probability of this outcome or an outcome more *distant* with respect to H_1. We might in fact associate two p-values with any given outcome in outcome space X: $p\_value_X(x) = (v_1, v_2)$ where v_1 is relative to H_1 and v_2 is relative to H_2:

$$\begin{aligned} p\_value_X(x) &= (v_1, v_2) \\ &= \left(\sum_{\{x' \in X : d_{H_1}(x') \geq d_{H_1}(x)\}} Pr(x'|H_1), \sum_{\{x' \in X : d_{H_2}(x') \geq d_{H_2}(x)\}} Pr(x'|H_2) \right) \end{aligned}$$

where $\quad d_{H_1}(x) = \frac{Pr(x|H_2)}{Pr(x|H_1)} \qquad d_{H_2}(x) = \frac{Pr(x|H_1)}{Pr(x|H_2)}$

The typical classical test can then be defined as follows:

$$t\_class1_X(x) = \begin{cases} a & \text{if } v_1 \geq \alpha; \\ r & \text{if } v_1 < \alpha. \end{cases}$$

If we are liberal in our characterization of classical tests, the following might be considered a more palatable classical test than the above:

$$t\_class2_X(x) = \begin{cases} a & \text{if } \frac{v_1}{v_2} \geq k; \\ r & \text{if } \frac{v_1}{v_2} < k \end{cases}$$

The idea here is that the epistemic/pragmatic context fixes, not α, but rather k, which is a threshold value for the ratio of p-values of an outcome. This test can be shown to satisfy Neyman and Pearson's condition[8] and thus could be written in the manner of the general $t_X(x)$ above. Note that t may differ depending on the particular experiment/outcome space, and so the test is not Bayesian.

[6] The way in which the context governs the value of α is not made explicit.

[7] In fact, for discrete probability distributions (as per our example), and barring the use of tests with some randomization, not all choices of α are possible, but that is not important here.

[8] The outcomes whose conditional probabilities are added to get v_2 are *almost* the complement of those outcomes whose conditional probabilities are added to get v_1, *almost* in the sense that one outcome—the experimental outcome in question, x—is common to both. Thus if there is some outcome x_i in R because it makes $\frac{v_1}{v_2} < k$, then all x_j that are more 'distant' than x_i relative to H_1 will also be in R, because these will yield an even smaller value for $\frac{v_1}{v_2}$. Thus Neyman and Pearson's condition will be satisfied.

3 A note about experimental design

The above definitions make clear that, once an outcome x has been recorded, the Bayesian does not care about the full outcome space, whereas the classical statistician does. That is, the following is true of Bayesian tests, but the corresponding expression is not necessarily true of classical tests:

$$t\_Bayes_X(x) = t\_Bayes_{X'}(x) \qquad \forall X, X', \forall x \in X \bigcap X'$$

The dependence of classical inference on the outcome space is regarded a major downfall by Bayesians, as will be discussed in the next section. But outcome spaces and thus stopping rules *do* matter to the Bayesian as well, at least when it comes to experimental design, or choosing amongst tests, if not when drawing an inference from a given outcome. This is well appreciated by statisticians, but worth discussing because it is relevant to the rest of the paper.

Assume an 'ideal' decision setting in which the cost of information is free. Basically, the classical statistician chooses the test that minimizes error probabilities subject to constraints, whereas the Bayesian chooses the test that maximizes expected utility. For the first type of classical test considered, the type I error is fixed, so the optimal test is the one whose stopping rule/outcome space yields the smallest type II error. The choice is trickier when neither α nor β is fixed (as per the second type of classical test considered) and the available tests are such that one test has minimal type I error while another has minimal type II error. In such cases there may be a number of admissible tests; the classical approach does not offer an explicit method for determining the optimal balance between type I and type II errors.

For the Bayesian, tests are assessed according to their expected utility. The expected utility (EU) of a test is calculated by considering all the possible outcomes and how probable they are, and what decision about the hypotheses would be made given each of these outcomes. More precisely, we sum the probability of the experimental outcome and the expected utility of the act that would be chosen if this outcome occurred, for all possible outcomes (here assuming information is free):

$$\begin{aligned} EU(t\_Bayes_X) &= \sum_{\{x \in X\}} Pr(x) \cdot \max j\{p_1'(x) \cdot u_{j1} + p_2'(x) \cdot u_{j2}\} \\ &= \sum_{\{x \in X\}} Pr(x) \cdot \max j\{\frac{Pr(x|H_1) \cdot p_1}{Pr(x)} \cdot u_{j1} + \frac{Pr(x|H_2) \cdot p_2}{Pr(x)} \cdot u_{j2}\} \\ &= \sum_{\{x \in X\}} \max j\{Pr(x|H_1) \cdot p_1 \cdot u_{j1} + Pr(x|H_2) \cdot p_2 \cdot u_{j2}\} \end{aligned}$$

Let us refer to the set of outcomes for which the maximum j is a (accepting H_1) as X_a and the set of outcomes for which the maximum j is r (rejecting H_1) as X_r. ($X_a \cup X_r = X$ and $X_a \cap X_r = \emptyset$) Now:

$$
\begin{aligned}
EU(t\_Bayes_X) &= \sum_{\{x \in X_a\}} [Pr(x|H_1) \cdot p_1 \cdot u_{a1} + Pr(x|H_2) \cdot p_2 \cdot u_{a2}] \\
&+ \sum_{\{x \in X_r\}} [Pr(x|H_1) \cdot p_1 \cdot u_{r1} + Pr(x|H_2) \cdot p_2 \cdot u_{r2}] \\
&= p_1 \cdot \left(u_{a1} \cdot \sum_{\{x \in X_a\}} Pr(x|H_1) + u_{r1} \cdot \sum_{\{x \in X_r\}} Pr(x|H_1) \right) \\
&+ p_2 \cdot \left(u_{a2} \cdot \sum_{\{x \in X_a\}} Pr(x|H_2) + u_{r2} \cdot \sum_{\{x \in X_r\}} Pr(x|H_2) \right)
\end{aligned}
$$

We can express the above using error probabilities α and β (as per the equation below, and depicted in Figure 2):

$$
EU(t\_Bayes_X) = p_1 \cdot (u_{a1} \cdot (1 - \alpha) + u_{r1} \cdot \alpha) + p_2 \cdot (u_{a2} \cdot \beta + u_{r2} \cdot (1 - \beta))
$$

So we see that error probabilities (and consequently the stopping rule/outcome space for a test) are important to the Bayesian when it comes to experimental design, since the error probabilities in a sense determine the expected utility and thus the choice-worthiness of a test. In general, the smaller the error probabilities, the greater the EU of the test (so long as the utility function favours accepting true over false hypotheses).

4 Classical statistics on the back foot

The following sections consider the merit, relative to the Bayesian approach, of classical statistics and its focus on error probabilities. Recall our limited setting: comparing two simple hypotheses. From the discussion thus far, Bayesian tests surely stand on firmer ground than their classical counterparts, simply because they are less ad hoc. The Bayesian model makes explicit how tests should depend on the epistemic and pragmatic context, and offers a clear decision method for choosing amongst tests. Classical statistics leaves us in the dark about the finer points of such matters. Ad hoc need not mean inconsistent or absurd, however, and indeed this section defends classical tests against more damning criticism along these lines.

To claim that inference should not depend on outcomes that might have, but did not, in fact, occur, is not an argument for the Bayesian position; it is just a restatement of that position. More substantial arguments have been offered, however, to undermine the classical dependence on stopping rules/outcome spaces. For instance, Howson and Urbach (1989) tell the story of an experimenter doing some trials (let's say coin tosses, to test for the chance of heads

for the coin), where they are prepared to stop just as soon as they get bored and hungry and so leave to go for a meal. This is a case where the stopping time depends not just on the chancy outcomes of the trials already performed, but also on other chancy aspects of the world, like the mood of the experimenter, which may itself be affected by all sorts of factors. The problem is that it is not clear how one should take account of these varied factors in listing the different possible stopping times and thus the outcome space for the experiment. In other words, this is a case where one cannot distinguish whether the test performed under the rogue stopping rule is t_A or t'_A, A and A' being two different outcome spaces. (Of course, more than two outcome spaces may be possible candidates for the stopping rule.) As we have seen, the distinction is important for classical inference: it may be the case that $t\_class_A \neq t\_class_{A'}$.

This criticism of the classical approach is somewhat overstated, since we have seen that stopping rules/outcome spaces matter to Bayesians too in designing experiments. Odd stopping rules that are difficult to map to an outcome space frustrate the Bayesian's plans. Such rules make it unclear what are the set of acts under consideration. To continue with the example above, it may not be clear whether the rogue stopping rule makes t_Bayes_A or $t\_Bayes_{A'}$ an available option.

One might suggest that both the classical and Bayesian camps should restrict the stopping rules an experimenter can legitimately entertain or plan for. But there will remain, of course, a genuine asymmetry between the two approaches. Experiments need not proceed as planned. Trialling might be stopped spontaneously in some 'illegitimate' way, and this will be a problem for the classical approach because inference as well as experimental design depends on the outcome space of the test. It is arguably this possibility that Howson and Urbach wished to draw attention to with their example of the experimenter running off for lunch.

There may yet be a way to alleviate this problem, however. Perhaps only some features of a stopping mechanism are statistically important and should be considered random variables that affect the outcome space. Arguably, what matters regarding the experimenter's stopping mechanism is how it relates to the trials themselves. Aspects of the greater environment that are independent of the trial outcomes (like when the experimenter gets hungry) would simply be treated as constants. For example, according to this rationale, Howson and Urbach's stopping rule above would be treated as a fixed-trial test; it is as if the experimenter was always going to stop after the number of trials that was conducted before they became too hungry. This is one suggestion. The point is that the classical statistician owes us an account of how to map from an everyday kind of stopping rule, like 'stop when the experimenter gets hungry', to a particular outcome space. We should not rule out the possibility that such an account can be given, however.

Of course, even if there is a principled way to map apparently ill-defined stopping rules to a corresponding well-defined stopping rule/ outcome space, classical inference has a curious subjectivity. The stopping mechanism depends in part on the intentions of the experimenter, and what's more, the experimenter

may not even know their own intentions. So inference from a test outcome is dependent on subjective beliefs about an experimenter's subjective intentions! Furthermore, this is not the kind of subjectivity that may be 'washed out' by repeated testing, as per Bayesian priors (which come under attack by classical statisticians on account of their subjectivity).[9] Each new test, rather, brings a whole new element of subjectivity to the evaluation of hypotheses. This may not be such a bad thing, but it would be a mistake to regard the classical approach as the *objective approach* to statistical inference.

5 The classical comeback: 'persistent' experimenters

Thus far the classical statistician is depicted as being on the defensive, but they do have an interesting positive case for the dependence of inference on stopping rules. (In fact the following is inspired by a 2001 paper by Mayo, who is a proponent of 'error statistics'.) The example above comparing the 20-coin-toss test with the one stopping on 6 tails may not be moving, one way or the other. There are certain kinds of stopping rules, however, that many intuitively think *should* matter to inference. For instance, imagine that a pharmaceutical company stands to gain much profit if their new drug X is shown to be a certain amount more effective than the existing drug Y. The company decides to run a test and are prepared to be persistent—they trial however many patients it takes for the results to sufficiently favour drug X being more effective. (There may be a maximum number of trials, m, that can feasibly be run.) This is referred to as an *optional stopping rule*: sampling will cease just as soon as the outcome has desired evidential implications, or after the maximum m trials, if there is such a maximum.

The original investigations of the 'optional stopping problem' for the Bayesian approach were concerned with an unbounded number of trials. The question was whether the Bayesian could be practically certain to get some desired evidence if they were persistent enough, i.e. if they were were prepared to keep doing trials until the outcome yielded a suitable likelihood ratio for the hypotheses. Consider the case where the experimenter wants the posterior probability of H_1 (regarding the chance of heads) to be sufficiently lowered, relative to H_2. Can they be 'guaranteed' this (probability equal to one) if they are prepared to toss the coin indefinitely, even if H_1 is in fact true? (The type I error would be equal to one.) This would mean stopping the trial only when the likelihood ratio $\frac{Pr(x|H_1)}{Pr(x|H_2)}$ is less than some value c, where $c \leq 1$ (since the probability for H_1 must be lowered). Happily for the Bayesian method, the probability that such an experiment will end when H_1 is in fact true has been shown to be, not 1 at all, but less than or equal to c. This means that the smaller the posterior probability for H_1 and thus the smaller the likelihood ratio one requires (i.e. the smaller the value of c), the smaller the maximum probability that this will be achieved and the experiment stopped, if H_1 is in fact true.[10]

[9] I am referring to the Bayesian 'convergence theorems'. See Earman (1992).

[10] This result is proved by Robbins (1970), as reported in Sober (2008, p. 77). Kadane et al. (1996) prove a similar result, which they contrast with more problematic cases that

An optional stopping rule could be framed in terms of any desired likelihood ratio c (or it might make reference to a number of critical likelihood ratios), but given this paper has focused on tests that yield an accept/reject result, let us continue with this theme. Define an *optional stopping test* as one in which trialling continues until the desired test result is achieved (either a or r) or after the maximum m trials (if there is a maximum). We might refer to such a test as *self-referential*: the stopping rule for the test depends on the results of the test itself, determined progressively during trialling. In fact, if we define an optional stopping test in this way, then there cannot be such a classical test; the results of a classical test depend (by definition, at least in this paper) on the stopping rule/outcome space, so it cannot also be the case that the stopping rule/outcome space also depends on the results of the test. We will return to this point shortly, after considering optional stopping tests in the Bayesian setting.

It follows from the result above that a Bayesian falls rather short of being 'guaranteed' (having probability equal to one) of, say, rejecting H_1 when it is in fact true, even if they are prepared to trial indefinitely.[11] This is a significant result, but, arguably, it does not close the optional stopping case. We may still have concerns about a drug company seeking to substantially raise the chances of finding evidence to support their case—rejection of the hypothesis that their drug has no effect—via persistent trialling. The company might be prepared to run some large maximum number of trials, stopping anytime the evidence is in their favour. Mayo (2001, p. 393) urges that optional stopping is a problem for the Bayesian, even if they cannot be 'guaranteed' strong evidence against a true hypothesis. She states that the problem is "rather the fact that ignoring stopping rules can lead to a high probability of error..."

In order to investigate this issue, let us leave drug companies aside, and return to the more mundane coin-tossing scenario. Consider a very simple example where the hypothesis space concerns the chance of heads for a coin, as per Section 2 ($H_1 : ch(H) = 0.5; H_2 : ch(H) = 0.6$). The test is stopped when H_1—that the coin is fair—is rejected, or after 7 tosses, whichever happens first. While this experiment is not very striking, given that it concerns the trivial matter of the bias of a coin and involves a maximum of only 7 tosses, one can easily draw an analogy to more dramatic cases like drug trials. Assume that the background to the test is as follows: the two hypotheses are considered equally likely, and the decision utilities are such that $u_{r2} - u_{a2} = u_{a1} - u_{r1}$. This means that H_1 will be rejected (r chosen) just in case the likelihood ratio for the experimental outcome, $\frac{Pr(x|H_1)}{Pr(x|H_2)} < 1$. The 'trick' with the optional stopping test is to

involve a probability function over a continuum of hypotheses that does not satisfy countable additivity. Armitage (1962) first posed the optional stopping problem for Bayesian statistics in the latter kind of setting, as reported by Mayo (2001).

[11] If the desired test result is r, then trialling continues until:

$$\frac{Pr(x|H_1)}{Pr(x|H_2)} < c \qquad \text{where } c = \frac{p_2 \cdot (u_{r2} - u_{a2})}{p_1 \cdot (u_{a1} - u_{r1})}$$

As per the above, if $c \leq 1$, the probability that the experiment will end is less than or equal to c.

monitor the trials/sampling with the test result in mind. A sample continues, $x = (t_1, t_2, \ldots)$ where $t_i \in \{H, T\}$ until the first t_n such that $t\_Bayes_X(x) = r$. If the experiment has not already stopped before the maximum m trials, then it is stopped at this point, regardless of the test result. For our example with priors and decision utilities stipulated above, and where $m = 7$, the outcome space X_O includes the following outcomes:
$X_O = \{(H), (THH), (THTHH), (TTHHH), (THTHTHH), \ldots\}$
The missing outcomes are sequences of 7 coin tosses. Given the nature of the stopping rule, the outcomes that involve less than 7 tosses all result in r—the rejection of H_1.

What might be the problem with such a test? Well, in loose terms, it seems that the test is unfairly stacked against H_1. After all, the stopping rule gives H_1 every chance of being rejected, since the experiment is stopped as soon as a sample yields this result. This means that, for our example case, it is more likely than not that H_1 will be rejected, *whether or not it is true*. In fact, if H_1 is true, it nonetheless has a considerably high probability of being rejected: the type I error probability (α) = 0.727. And this test has a maximum of only 7 tosses. The probability of rejecting H_1 when it is false is also high: $1 - \beta = 0.855$. (Refer to the appendix—Table A1—for a partial explanation of these calculations.)

The optional stopping test is supposed to push our intuitions that error probabilities, which depend on the stopping rule/outcome space for a test, should indeed 'matter' to the inferences that we draw about hypotheses. If a drug company presents some results to us—"a sample of n patients showed that drug X was more effective than drug Y"—and this sample could i) have had size n fixed in advance, or ii) been generated via an optional stopping test that was 'stacked' in favour of accepting drug X as more effective—do we care which of these was the case? Do we think it is relevant to ask the drug company what sort of test they performed when making our final assessment of the hypotheses? If the answer to this question is 'yes', then the Bayesian approach seems to be wrong-headed or at least deficient in some way. Returning to our coin example: imagine that the trials did in fact stop before the maximum number, because the desired result was achieved. Perhaps the sequence $x = (TTHHH)$ was recorded. Do we intuitively want to treat this outcome differently from the same outcome produced by a test that was fixed at 5 tosses (with outcome space X_F)? In other words, should the stopping rule with its associated outcome space and error probabilities affect inference? Note that the error probabilities for the fixed 5-toss test are as follows: $\alpha = 0.5$ (which is considerably less than 0.727 for the optional stopping test) and $\beta = 0.317$. (Refer to table A2 in the appendix for details.)

Recall from Section 3 that the result of a Bayesian statistical test is independent of the outcome space for the test, X, even if, in the peculiar case of optional stopping tests, the outcome space is itself dependent on the results of the test. Arguably, intuitions are more in line with the classical approach when it comes to optional stopping rules. In the classical setting, the same outcome may lead to different inferences or test results, depending on whether it was the result of a fixed-trial test or a test designed to stop when the likelihood ratio of the outcome satisfied some criteria. For instance, consider the typical classical

test, $t\_class1_X(x)$ for our two different stopping rules/outcome spaces: X_O and X_F. Recall that the stopping rule associated with X_O is to stop after 7 trials or when the likelihood ratio is less than 1. For this test, the relevant p-value, v_1, of the outcome $x = (TTHHH)$, is 0.688, while for the fixed-trial test, v_1 is 0.5.[12] Thus, if the appropriate value for the type I error, α, was deemed just greater than 0.5, then the outcome $(TTHHH)$ would lead to a rejection of H_1 for the fixed-trial test with outcome space X_F, and yet it would lead to the acceptance of H_1 for the test with outcome space X_O.[13] This might seem a reasonable state of affairs: a particular outcome speaks less for the rejection of a hypothesis if the test was designed to stop as soon as the experimenter found evidence lowering the probability of that hypothesis, as compared to the outcome being produced by a fixed-trial test.

We might also compare the results of the other sort of classical test mentioned above: $t\_class2_X(x)$. Here too, the outcome $x = (TTHHH)$ will lead to different decision results for our two stopping rules X_O and X_F, for certain values of k. For instance, $k = 1$ might be deemed appropriate to the context. This would mean rejecting H_1 just in case the ratio of p-values, $\frac{v1}{v2}$ is less than 1. This is indeed the case for our example fixed trial test: $v_1 = 0.5 < v_2 = 0.663$. Thus H_1 would be rejected. For the optional stopping test, however, $v_1 = 0.688 > v_2 = 0.256$ and so there would not be sufficient evidence to reject H_1.

So much for how the classical tests play out for the particular outcome spaces X_F and X_O. In general, there cannot be a classical optional stopping test, because we need a nominal sample space in order to calculate nominal p-values for an outcome: $[v_1]$ and $[v_2]$. (These terms are in square brackets to distinguish them from the *actual* p-values.) If the sample is constructed in such a way that the stopping rule is a function of the supposed $[v_1]$ and/or $[v_2]$, then this will induce a new sample space, and each outcome will have new actual p-values that are not necessarily equivalent to the $[v_1]$ and $[v_2]$ referred to in the stopping rule. Thus the test may not yield the same result that was intended by the stopping rule, as per the example just given.

6 A Bayesian error theory for the 'optional stopping intuition'

At the end of the day, it must simply be admitted that Bayesian statistics runs against the intuition that, at least where the optional stopping test is concerned, it should not be the case that:

$$t_X(x) = t_{X'}(x) \qquad \forall X, X', \forall x \in X \bigcap X'$$

Call this the 'optional stopping intuition'. (If the optional stopping test was

[12] The p-values can be deduced from Tables A1 and A2 in the appendix. Recall that more 'extreme' outcomes than x relative to $H_1(H_2)$ are those for which the likelihood ratio is smaller(greater).

[13] This is admittedly a high value for α. Note that the two tests will have different results for $x = (TTHHH)$ for any value of α between 0.5 and 0.688.

designed to favour the rejection of H_1 (r), say, then an outcome from such a test provides less evidence for r/against H_1 than the same outcome produced by a fixed-trial test.) This section aims to mitigate or explain away this apparent problem for Bayesian statistics by offering three different error theories for the optional stopping intuition.

It is worth initially considering a strategy that does not work. In Section 3, it was noted that $t\_Bayes_X(x) = t\_Bayes_{X'}(x)$. But perhaps this was a bit quick, by way of analysing the import of stopping rules to Bayesian inference: we were assuming that the description of the outcome/evidence x is the same, regardless of the stopping rule. Perhaps the above comparison should rather be between $t\_Bayes_X(e)$ and $t\_Bayes_{X'}(e')$, where e amounts to 'sequence of tosses x and full outcome space X', and e' amounts to 'sequnce of tosses x and full outcome space X'''. If it is not generally the case that $Pr(e|H_i) = Pr(e'|H_i)$, then it is not generally the case that $t\_Bayes_X(e) = t\_Bayes_{X'}(e')$, and stopping rules would matter to Bayesian inference after all.

The problem with this strategy is that in the standard statistical setting, where the experimenter knows their own beliefs and preferences (we'll come back to this proviso later), $Pr(e|H_i) = Pr(e'|H_i)$ does generally hold. Let $e = x\&b$ and $e' = x\&b'$, where x is the random variable outcome (e.g. sequence of coin tosses) with chances stipulated by H_i and b and b' refer to the different stopping rules/outcome spaces. We get:

$$\begin{aligned} Pr(e|H_i) &= Pr(x\&b|H_i) = Pr(b|H_i) \times Pr(x|b\&H_i) \\ Pr(e'|H_i) &= Pr(x\&b|H_i) = Pr(b'|H_i) \times Pr(x|b'\&H_i) \end{aligned}$$

$Pr(x|b\&H_i) = Pr(x|b'\&H_i)$ since x is a random variable whose chance is fixed by the hypothesis H_i. Also, $Pr(b|H_i) = Pr(b'|H_i)$, as the stopping rule is independent of the chance hypotheses, once background information pertaining to the beliefs and preferences driving the choice of stopping rule is taken into account.

An error theory looks, then, to be the best response a Bayesian can give regarding the optional stopping intuition. Three error theories will be presented here. By way of an initial attempt, one might argue that the optional stopping intuition is misdirected, in the sense that we are confusing considerations of experimental design and inference. Recall from Section 3 the importance of outcome spaces/error probabilities for the Bayesian when it comes to experimental design, or choosing tests. It can be shown that, *when information is free*, an optional stopping test with maximum m trials is never more choice-worthy than a fixed-m-trial test, which will presumably also be an available option (see Good 1967 on the value of information). Thus, optional stopping tests are in a general sense less choice-worthy, and it might be argued that there is a tendency to confuse this with optional stopping tests providing lesser evidence once an outcome is recorded.

It is not entirely plausible, however, that considerations of choice-worthiness are the source of the optional stopping intuition, which really does seem to

be about inference. We can address inference if we move away from the standard statistical setting; perhaps the optional stopping intuition stems from cases where experimental results are *underdescribed*—only the properties of some portion of an outcome space are reported, rather than an individual outcome. To illustrate, let us return to the coin-tossing experiments from Section 5. Perhaps the experimenter chooses a test, and as per the story above, will reject H_1 when the likelihood ratio is less than 1. The difference is that the inference-maker, for whatever reason, learns only whether or not the outcome falls in the rejection region R, rather than the full coin-toss sequence.[14] As one might suspect, here the error probabilities/outcome space really do/does matter to inference. We can see this by considering the calculation of the posterior probability of H_1:

$$\begin{aligned} Pr(H_1|R) &= \frac{Pr(R|H_1) \times Pr(H_1)}{Pr(R|H_1) \times Pr(H_1) + Pr(R|H_2) \times Pr(H_2)} \\ &= \frac{\alpha \times Pr(H_1)}{\alpha \times Pr(H_1) + (1-\beta) \times Pr(H_2)} \end{aligned}$$

Assume the hypotheses and parameters specified earlier (including equal priors), and a report of 'rejection'. If the inference-maker learns that the optional stopping test (with maximum 7 tosses) was used, they will update their probability for H_1 to ≈ 0.46. If, on the other hand, they learn that some fixed-toss test was used, say 5 tosses, then the probability for H_1 is updated to ≈ 0.42. So the rejection report under optional stopping is here lesser evidence against H_1 than rejection under the fixed trial test. And perhaps people mistakenly carry this lesson over to the standard setting where reports are of individual outcomes.

The above is promising as an error theory for the optional stopping intuition, but one might nevertheless think it misses the mark. It could be argued that the optional stopping intuition concerns inference when individual outcomes are reported, and does not rest on mistaking this kind of report with one that gives only the properties of a set of outcomes. The third error theory attempts to accommodate reports of individual outcomes. It appeals to the fact that the choice of stopping rule may be informative. Recall the above proviso regarding the independence of stopping rules and the chance hypotheses under test: this only holds when the beliefs and preferences governing the selection of the stopping rule are known/count as background information. In practice, this may not be the case. For instance, as raised above, the inference-maker may be distinct from the experimenter who chooses the test. There is the possibility that the inference-maker learns something substantial when they learn what sort of stopping rule the experimenter employed. It may indicate something about the experimenter's attitudes, and these attitudes may have a bearing on the truth of the hypotheses under comparison, in a manner that corresponds with the optional stopping intuition. What follows is an attempt to make this possibility seem plausible.

The account rests on the assumption that the cost of running trials is negligible, and centres on the fact that an experimenter who chooses to do an optional stopping test with maximum m trials over a fixed m-trial test in these

[14] I owe the idea of appealing to underdescribed experimental results to xxx.

circumstances most likely does not care about the truth of the hypotheses under consideration. Consider first the case where information is entirely free, in the sense of Good's 1967 theorem. If the (rational) experimenter chooses optional stopping then they must be indifferent between this test and the fixed-trial test, and the most likely explanation for the indifference is that they simply do not care about the truth—one act (say, 'rejecting H_1') dominates the other.[15] If the experimenter actually *prefers* an optional stopping test over rivals, including the fixed-m-trial test, then the new information is not free, but it may be costly for quite subtle reasons. We are assuming that the trialling procedure itself does not incur any costs (in terms of resources required, lost opportunities, etc.), so the cost presumably comes from the bad consequences associated with the presence of new evidence. Consider the drug company, for instance: even if the cost of doing medical trials is negligible, the company may incur a loss by running extra trials if the results of these trials are made public and cause a significant change in the consumer choices of others. One might suspect that this is what is going on if the drug company chooses an optional stopping test: what matters for the drug company is just whether there is *sufficient* evidence against H_1 (that the drug has no positive effect), and not the actual truth of H_1.

This is the sense in which the inference-maker may learn quite a lot when they learn the stopping rule for a test. They may, for instance, learn that the experimenter is unscrupulous, in that they are not interested in the truth of the hypotheses! This may be relevant to the truth of these hypotheses. For example, if one learned the outcome of a drug company's optional stopping test regarding whether their new drug is effective, one might not attribute as high a probability to the drug being effective as would be the case if the evidence was known to come from a fixed-trial test. But this need not be a violation of Bayesian principles, i.e. it need not reflect a concern about the stopping rule/error probabilities for the test. If their choice of stopping rule is evidence that a drug company doesn't care about the efficacy of their drug, then it is plausibly also evidence against the efficacy of that drug! In effect, this may be a case where $Pr(b|H_i) \neq Pr(b'|H_i)$, relative to background knowledge. (Recall that b and b' refer to the stopping rule/outcome space of the test conducted.) In other words, $Pr(H_i|b) \neq Pr(H_i)$: the experimenter's choice of stopping rule is relevant to the truth of the hypotheses, over and above the actual test outcome.

This might seem a rather tenuous Bayesian account of the optional stopping intuition, but optional stopping discussions in the literature do presume a distinction between experimenter and inference-maker, and the questionable character of the experimenter who chooses an optional stopping test is made prominent. For instance, in a brief discussion about optional stopping, Hacking (1965, p. 109) writes: '...although it may be morally deplorable to pretend one had settled n in advance [i.e. settled on a fixed-trial test rather than an optional stopping test], such a lie is statistically innocuous'. While I agree with Hacking that the optional stopping rule is 'statistically innocuous' for Bayesian inference, in the sense that it does not affect the basic evidential logic of the test, the inferred attitudes of the experimenter may, on the other hand, be relevant

[15] Admittedly, another remote possibility is that the experimenter cares about the truth, but their priors and decision utilities surprisingly result in the optional stopping test and the fixed-trial test having equal expected utility.

to the truth of the statistical hypotheses, and these attitudes may be an important determinant of our intuitions about whether or not stopping rules 'matter'.

The character of the optional stopping experimenter is, furthermore, central to a dispute between Mayo (2001) and Berger and Wolpert (1988). In response to the optional stopping problem originally posed by Armitage,[16] Berger and Wolpert apparently state that the Bayesian might raise the probability for H_1, 'perhaps to reflect a suspicion that the agent is using [the optional] stopping rule [favouring the rejection of H_1] because he thinks the null hypothesis [H_1] is true'. According to Mayo, this is to concede to the classical statistician that stopping rules matter, and, in fact, to make this concession two times over. To begin with, Mayo notes that it is very non-Bayesian to alter one's priors depending on what kind of test will be performed. The second point Mayo makes is that, to be suspicious of an optional stopping test is to admit that this is a bad test, presumably due to its high error probabilities. Mayo (2001, p. 398) remarks: 'Equating optional stopping with deception runs counter to Savage's insistence that, because "optional stopping is no sin", any measure that is altered by the stopping rule, such as the significance level, is thereby inappropriate for assessing evidence [Savage 1964, p. 185].'

But we see that there is a Bayesian story for why choosing optional stopping *can* be a sin of sorts, and furthermore, why this may affect confidence in the hypotheses being tested. Mayo is right that it would be entirely unsatisfactory for a Bayesian to suggest, as a solution to a perceived problem with the method, that when the problem arises an agent should simply change their priors. And it does look like this is what Berger and Wolpert are suggesting. The point here, however, is that stopping rules can be informative before we even consider the outcome of a test, because they reveal something of an experimenter's attitudes. We can interpret Berger and Wolpert's suggestion as follows: the inference-maker learns something about the experimenter's beliefs when they learn that optional stopping was chosen, and this changes the inference-maker's own beliefs. They do not offer a narrative for these dependencies in the inference-maker's beliefs, but the basic idea is consistent with the account given here: the inference-maker first conditionalises on what they learn about the experimenter's beliefs from their choice of stopping rule, and then proceeds with the usual Bayesian analysis of the test outcome.

Admittedly, there are some substantial assumptions and jumps in inference along the chain of reasoning we have been discussing, from an experimenter's choice of stopping rule, to their disinterest in the truth of the hypotheses, to what this says about the truth of these hypotheses. But we have seen that optional stopping tends to be discussed in a very suggestive way, which is why one may associate the rule with a particular pattern of reasoning—reduced confidence in the hypothesis favoured by the optional stopping rule. The error is to assume that this pattern of reasoning should apply more broadly, to standard cases where stopping rules are not in fact informative. That is the essence of the third error theory: to the extent that the optional stopping intuition prevails,

[16]Recall that this problem involves a prior probability distribution over a continuum of hypotheses that does not satisfy countable additivity, but in any case, the details need not concern us.

it is because people tend to draw general conclusions about optional stopping from a small number of very suggestive examples where the optional stopping rule is informative, and does matter to inference in a manner consistent with Bayesian logic.

7 Concluding Remarks

This paper set out to compare the logic of classical and Bayesian statistics within a restricted domain: dichotomous tests of two disjoint simple hypotheses. The idea was to keep an open mind about classical statistics, at least within this domain—to consider whether the classical approach might be plausibly defended, and whether there is at least some cause for concern about the Bayesian approach.

Sections 2 to 4 highlighted that there is, in a sense, much commonality between the approaches. This is positive for both. Bayesian statistics can dodge criticism that it does not acknowledge error probabilities—after all, error probabilities are important in Bayesian analysis too because they determine the choice-worthiness of tests. For the same reason, classical statistics can dodge at least some of the criticism regarding ill-specified stopping rules/outcome spaces—such rules also frustrate Bayesian experimental design.

Arguably the best case for the classical approach to inference is our intuitions about persistent experimenters and optional stopping rules. The intuition is that, in these cases, error probabilities, and thus stopping rules, should matter to more than just the choice of tests, but to inference as well, in a manner at odds with the Bayesian model. This was here treated as a challenge for the Bayesian approach that deserves some response. Of course, one might respond that there is no challenge at all, and anyone who has a problem with the Bayesian model is simply misguided. Such a response would be acceptable, but it does nothing to strengthen the Bayesian position. The more constructive response is to account for the 'optional stopping intuition'—that was the goal of the last section. Three different error theories were offered for why we may have the mistaken intuition that at least some stopping rules should affect inference in a manner somewhat in line with classical, but not Bayesian, statistics.

The three error theories of Section 6 each have problems, but collectively they significantly weaken the optional stopping argument for classical statistics. Indeed, if the classical approach has no obvious strength in the simple statistics setting covered in this paper, it is difficult to see how it could stand up to scrutiny in more complicated statistical settings (involving *composite* hypotheses, for starters). Notwithstanding its many varieties, Bayesianism has a simplicity and elegance that is hard to beat—all kinds of inference boil down to the logic of subjective probability and utility. Classical statistics may not be inconsistent, but it has no clear advantage over Bayesianism, and it has many loose ends, particularly regarding the connection between evidence and decision.[17]

[17]Many thanks to...

References

Berger, James O., and Robert L. Wolpert. 1988. *The Likelihood Principle*. 2nd ed. Haywood, CA: Institute of Mathematical Statistics.

Earman, John. 1992. *Bayes or Bust?: A Critical Examination of Bayesian Confirmation Theory* Cambridge, MA: MIT Press.

Edwards, Anthony W. F. 1972. *Likelihood*. 1st ed. Cambridge: Cambridge University Press.

Good, I. J. 1967. On the Principle of Total Evidence. *British Journal of Philosophy of Science* 17:319–321.

Hacking, Ian. 1965. *Logic of Statistical Inference*. London & New York: Cambridge University Press.

Howson, Colin, and Peter Urbach. 1989. *Scientific Reasoning: The Bayesian Approach*. La Salle, III.: Open Court.

Kadane, Joseph B., Mark J. Schervish, and Teddy Seidenfeld. 1996. Reasoning to a Foregone Conclusion. *Journal of the American Statistical Association* 91 (435):1228–1235.

Kendall, Maurice, and Alan Stuart. 1991. *Kendall's Advanced Theory of Statistics*. 5th ed. Vol. II. London: Edward Arnold.

Mayo, Deborah G., and David R. Cox. 2006. Frequentist Statistics as a Theory of Inductive Inference. *IMS Lecture Notes - Monograph Series*.

Mayo, Deborah G., and M. Kruse. 2001. Principles of Inference and Their Consequences. In *Foundations of Bayesianism*, edited by D. Corfield and J. Williamson: Kluwer.

Sober, Elliott. 2008. *Evidence and Evolution*. Cambridge: Cambridge University Press.

Appendix

The following concern the two hypotheses:

H_1: $ch(Heads) = 0.5$
H_2: $ch(Heads) = 0.6$

For the Bayesian tests in Tables A1 and A2, the prior probabilities and decision utilities are such that $\frac{p_2 \cdot (u_{r2} - u_{a2})}{p_1 \cdot (u_{a1} - u_{r1})} = 1$

Table A1: Bayesian Optional Stopping Test (stop when $\frac{Pr(x|H_1)}{Pr(x|H_2)} < 1$, or $n = 7$)

x (summary form)	$Pr(x\|H_1)$	$Pr(x\|H_2)$	$\frac{Pr(x\|H_1)}{Pr(x\|H_2)}$	$t\_Bayes_X(x)$
1 Heads, 0 Tails (1)	0.5	0.6	0.833	r
2 Heads, 1 Tails (1)	0.125	0.144	0.868	r
3 Heads, 2 Tails (2)	0.03125	0.03456	0.904	r
4 Heads, 3 Tails (5)	0.007813	0.008294	0.942	r
3 Heads, 4 Tails (14)	0.007813	0.00553	1.413	a
2 Heads, 5 Tails (14)	0.007813	0.003686	2.120	a
1 Heads, 6 Tails (6)	0.007813	0.002458	3.179	a
0 Heads, 7 Tails (1)	0.007813	0.001638	4.768	a
	Type I error (α)	Type II error(β)		
	0.727	0.145		

Table A2: Bayesian Fixed 5-Trial Test

x (summary form)	$Pr(x\|H_1)$	$Pr(x\|H_2)$	$\frac{Pr(x\|H_1)}{Pr(x\|H_2)}$	$t\_Bayes_X(x)$
5 Heads, 0 Tails (1)	0.03125	0.07776	0.402	r
4 Heads, 1 Tails (5)	0.03125	0.05184	0.603	r
3 Heads, 2 Tails (10)	0.03125	0.03456	0.904	r
2 Heads, 3 Tails (10)	0.03125	0.02304	1.356	a
1 Heads, 4 Tails (5)	0.03125	0.01536	2.035	a
0 Heads, 5 Tails (1)	0.03125	0.01024	3.052	a
	Type I error (α)	Type II error(β)		
	0.5	0.317		