The Bayesian and Frequentist Approaches to Inference

Matthew Kotzen
kotzen@email.unc.edu
UNC Chapel Hill Department of Philosophy

Draft of September 26, 2011

1 Introduction

Bayesianism and Frequentism are two very different approaches to statistical inference and, I think, to inference more generally. Of course, there are many different versions of each view, often with very substantial differences. But there do seem to me to be some core philosophical differences between these two approaches to statistical inference, which have been under-explored in epistemology generally. In this paper, I will be criticizing the epistemological theses that are characteristic of the Frequentist approach. But my target is not any particular "Frequentist" statistician or epistemologist; I am, rather, interested in the arguments for and against the particular epistemological theses that I identify. To the extent that a "Frequentist" rejects the epistemological theses that I criticize, I have no objection to her approach.

In this paper, I aim to do two things. First, in Sections 2–4, I will explain what I take to be the epistemological cores of both the Bayesian and the Frequentist approaches by giving both analyses of the same simple coin case. Second, in Sections 5–8, I will identify what I take to be the four most salient philosophical differences between Bayesians and Frequentists, and argue that in each case the Bayesian's position is more defensible.

Many of the points that I will raise against the Frequentist approach are not new; in particular, much of my discussion will be guided by that of Howson and Urbach 1993. The reason for this is that one of my goals is to give a broad overview of the relative philosophical merits of the Bayesian and Frequent approaches, and I am not the first person to investigate this question; in particular, Howson and Urbach 1993 discuss some objections to Frequentism that arise from issues in Sections 6 and 7. But I am not aware of any systematic philosophical discussion of the epistemological theses that separate Bayesians and Frequentists, and I hope that one will be a welcome contribution. In addition, I think

that there is much that is novel in Sections 5–8, in particular: the quantum particle type of case in Section 5, the discussion of relativism/contextualism in Section 6, the discussion of the relevance of the Requirement of Total Evidence in Section 6, the discussion of Parfit's Two Lotteries in Section 7, and the point about the hypocrisy of the Frequentist's charge of subjectivism in Section 8.

2 Setup

Suppose that we have a coin and we want to collect evidence about its probability of landing heads on any particular flip (call this value c_h). Let's suppose that we have some sort of guarantee that this probability is constant; we know that, whatever its probability of landing heads on the first flip is, it has the same probability of landing heads on the second flip, and the third flip, and so on. And let's also suppose that we have a guarantee that, once you fix the value of c_h, the coin landing one way on one flip makes it no more or less likely that the coin will also land that way on any other flip.[1]

To gather evidence about the value of c_h, we decide to flip the coin 20 times and record the outcome. The outcome is: T, T, T, H, T, H, T, T, T, H, T, T, T, T, T, H, H, T, T, H. (There are 6 heads in this sequence and 14 tails.)

3 The Bayesian Approach

According to the Bayesian Approach, it is an inescapable fact of our epistemic lives that the credences we ought to have toward propositions after taking some piece of evidence into account depend on the credences that we had toward those propositions *before* taking the evidence into account; these earlier credences are known as our **prior probabilities**.

So, according to the Bayesian approach, the credences that I ought to attach to propositions of the form $c_h = x$ after observing the 20 coin flips depend on the credences that I attached to those propositions before observing the coin flips. Bayesians differ on how to think about prior probabilities, but in this case, it seems like it would be at least fairly reasonable to think that:

1. This coin is randomly sampled from the population of coins in the world.

2. Most of the coins in the world are pretty close to fair

[1] Of course, the coin landing heads on the 1st flip is some evidence that the coin is somewhat heads-biased, and hence that it's also going to land heads on the second flip. So, the "guarantee" here that we have is not: that a particular outcome on one flip can't make it more or less likely that the coin will have that outcome on another flip. Rather, the guarantee is that, any effects here are "screened off" by the value of c_h; in other words, the outcome of one flip only gives us information about the outcome of another flip *in virtue of* giving us information about the value of of c_h.

3. Heads-biasing and tails-biasing are at least approximately symmetrically distributed in the coin population; in other words, there are roughly the same number of slightly heads-biased coins as slightly tails-biased coins, and roughly the same number of very heads-biased coins as very tails-biased coins, etc.

4. As the bias of a coin increases, there are fewer and fewer coins with that bias; in other words, there are fewer moderately biased coins than slightly biased coins, and fewer very biased coins that moderately biased coins, etc.[2]

These considerations motivate[3] a prior credence distribution over possible biases of the coin that is normal (i.e., symmetrically bell-shaped) and peaks at $c_h = .5$. In order to avoid the mathematical complications that come with continuous distributions, however, let's use a discrete approximation of a continuous normal distribution:

$p(c_h = .2) = .05$
$p(c_h = .4) = .1$
$p(c_h = .5) = .7$
$p(c_h = .6) = .1$
$p(c_h = .8) = .05$

So, my prior probability that the coin will land heads on any particular flip is $(.2)(.05) + (.4)(.1)+(.5)(.7)+(.6)(.1)+(.8)(.05) = .01+.04+.35+.06+.04 = .50$. This value, of course, is unsurprising; since my credence distribution over possible values of c_h is symmetric about $c_h = .5$, there is nothing in my prior credence distribution to favor heads or tails, so my credence that the coin will land heads on any flip should be (and is) .5.

The next step in the Bayesian analysis of the coin is to calculate how likely the actual outcome (i.e., TTTHTHTTTHTTTTTHHTTH) is on the assumption of the various possible values for c_h.

This is straightforward to do. Since the coin flips are independent and the value of c_h is constant throughout the flips, the likelihood of the actual outcome is just the likelihood of the coin landing tails on the first flip, times the likelihood of the coin landing tails on the second flip, times the likelihood of the coin landing tails on the third flip, times the likelihood of the coin landing heads on the fourth flip, and so on for each of the 20 flips. On the assumption of a particular value for c_h, this quantity is just $(c_h)^6(1 - c_h)^{14}$:

[2]Of course, this might not be quite true; for instance, some novelty shops might sell trick coins that are heavily heads-biased, and perhaps they sell comparatively few mildly heads-biased coins. But I don't want to get detained on the details of coins; in a wide range of circumstances, a normal distribution seems to be a good candidate for a reasonable prior probability distribution, and I'm going to assume that this case is one of them.

[3]Of course, there are other distributions that are symmetrically decreasing as you move away from $c_h = .5$, such as an "inverted vee" with its peak at $c_h = .5$. But the Central Limit Theorem provides some strong motivations for using a normal distribution in this sort of case. See, e.g., Brosamler 1988.

$p(\text{TTTHTHTTTHTTTTTHHTTH}|c_h = .2) = (.2)^6(.8)^{14}$
$p(\text{TTTHTHTTTHTTTTTHHTTH}|c_h = .4) = (.4)^6(.6)^{14}$
$p(\text{TTTHTHTTTHTTTTTHHTTH}|c_h = .5) = (.5)^{20}$
$p(\text{TTTHTHTTTHTTTTTHHTTH}|c_h = .6) = (.6)^6(.4)^{14}$
$p(\text{TTTHTHTTTHTTTTTHHTTH}|c_h = .8) = (.8)^6(.2)^{14}$

Bayes' Theorem says that the posterior probability for a hypothesis H_1, conditional on evidence E, is just the prior probability of H_1, times the likelihood of E on the supposition of H_1, divided by the prior "expectedness" of E (which is just the weighted sum of: prior probabilities in the each hypothesis times the likelihood of E on the supposition of each hypothesis):

$$\text{Bayes's Theorem: } p(H_1|E) = \frac{p(H_1)p(E|H_1)}{p(E)} = \frac{p(H_1)p(E|H_1)}{p(H_1)p(E|H_1)+p(H_2)p(E|H_2)+...+p(H_n)p(E|H_n)}$$

This then allows us to calculate posterior probabilities for each of the values of c_h:

$$p(c_h = .2|\text{TTTHTHTTTHTTTTTHHTTH}) = \frac{(.05)(.2)^6(.8)^{14}}{(.05)(.2)^6(.8)^{14}+(.1)(.4)^6(.6)^{14}+(.7)(.5)^{20}+(.1)(.6)^6(.4)^{14}+(.05)(.8)^6(.2)^{14}} \approx .123$$

$$p(c_h = .4|\text{TTTHTHTTTHTTTTTHHTTH}) = \frac{(.1)(.4)^6(.6)^{14}}{(.05)(.2)^6(.8)^{14}+(.1)(.4)^6(.6)^{14}+(.7)(.5)^{20}+(.1)(.6)^6(.4)^{14}+(.05)(.8)^6(.2)^{14}} \approx .281$$

$$p(c_h = .5|\text{TTTHTHTTTHTTTTTHHTTH}) = \frac{(.7)(.5)^{20}}{(.05)(.2)^6(.8)^{14}+(.1)(.4)^6(.6)^{14}+(.7)(.5)^{20}+(.1)(.6)^6(.4)^{14}+(.05)(.8)^6(.2)^{14}} \approx .585$$

$$p(c_h = .6|\text{TTTHTHTTTHTTTTTHHTTH}) = \frac{(.1)(.6)^6(.4)^{14}}{(.05)(.2)^6(.8)^{14}+(.1)(.4)^6(.6)^{14}+(.7)(.5)^{20}+(.1)(.6)^6(.4)^{14}+(.05)(.8)^6(.2)^{14}} \approx .011$$

$$p(c_h = .8|\text{TTTHTHTTTHTTTTTHHTTH}) = \frac{(.05)(.8)^6(.2)^{14}}{(.05)(.2)^6(.8)^{14}+(.1)(.4)^6(.6)^{14}+(.7)(.5)^{20}+(.1)(.6)^6(.4)^{14}+(.05)(.8)^6(.2)^{14}} \approx .000$$

So, our posterior probability that the next flip of the coin will be heads $\approx (.123)(.2) + (.281)(.4) + (.585)(.5) + (.011)(.6) + (.000)(.8) \approx .436$. Again, this is not surprising; an outcome consisting of 14 tails and only 6 heads is some reason to become more confident that the coin is tails-biased and less confident that the coin is heads-biased, and hence reason to lower (from .5) our credence that the next flip of the coin will be heads.

4 The Frequentist Approach

There are actually several different Frequentist approaches, but I'm only going to focus on one: namely, Fisherian significance tests. Though I think that much of what I will say applies to, e.g., Neyman-Pearson tests, I will not be addressing them specifically.[4]

The general strategy of a Fisherian significance test is as follows:

1. Choose a hypothesis H_0, referred to as the Null Hypothesis, which you are going to see whether you have good grounds to *reject*.

2. Figure out the possible outcomes of the experiment, and assign a likelihood to each outcome on the assumption that H_0 is true.

3. Once you obtain the actual outcome, calculate the likelihood that that outcome or an outcome at least as unlikely would occur by summing the likelihoods of each outcome that is at least as unlikely as the actual outcome (including the actual outcome itself).

4. Use this sum (called a p-value) as a guide to the rejection of H_0. In other words, if the p-value is $\leq \alpha$, then your results are statistically significant at level α and you may reject H_0 at that significance level. The lower the value of α is, the "stronger" the rejection of H_0 is.

The intuitive thought here is that when a Null Hypothesis says that it was very unlikely that an outcome as least as "extreme" as the actual outcome would occur, and then the actual outcome does occur, that's reason to reject that Null Hypothesis. And, the less likely that an outcome at least as extreme as the actual outcome is, according to the Null Hypothesis (i.e., the lower the p-value of the experiment), the stronger grounds we have to reject that Null Hypothesis. By contrast, if a Null Hypothesis says that it was fairly likely that an outcome as least as extreme as the actual outcome would occur (i.e., if the p-value of the experiment is relatively high), then we have comparatively weak reason to reject the Null Hypothesis.

Let's apply the procedure above to the case of the coin. First, let's let H_0, the hypothesis that we're seeing whether we have good grounds to reject, be the hypothesis that the coin is fair (i.e., that $c_h = .5$). Second, we need to think about the possible outcomes of our 20-flip experiment. One natural way to think about the possible outcomes is in terms of the number of times that the coin landed heads.[5] On this understanding, there were 21 possible outcomes of our experiment: {0 heads, 1 head, 2 heads, ..., 20 heads}. On the assumption that H_0 is true (i.e., that $c_h = .5$), it is straightforward to assign a likelihood to each of these 21 outcomes. The likelihood of a 1-head outcome, for instance, is the likelihood of any particular 1-head outcome (say, TTTTTTTTTTHTTTTTTTTT), times

[4] For a discussion of Neyman-Pearson tests, see Howson and Urbach 1993.

[5] Though this is complicated—see Section 6 below.

the number of 1-head outcomes (in this case 20, since the one H could occur on any of the 20 flips). In general, on the assumption that $c_h = .5$, the likelihood of any particular r-heads outcome (indeed, of *any* particular sequence of H's and T's) is just $(.5)^{20}$. And in general, the number of different possible r-heads outcomes is ${}^{20}C_r = \frac{20!}{r!(20-r)!}$. (This of course yields the result above that the number of different possible 1-head outcomes is $\frac{20!}{1!(19!)} = \frac{20!}{19!} = 20$.) Thus, on the assumption of H_0, the likelihood that we would observe some r-heads outcome or other is: $(.5)^{20}\frac{20!}{r!(20-r)!}$. The approximate values of this likelihood for all 21 values of r are given in the table below:

r	$p(r$ heads$\|H_0)$	r	$p(r$ heads$\|H_0)$
0	.000	11	.160
1	.000	12	.120
2	.000	13	.074
3	.001	14	.037
4	.005	15	.015
5	.015	16	.005
6	.037	17	.001
7	.074	18	.000
8	.120	19	.000
9	.160	20	.000
10	.176		

Now, since the actual outcome (TTTHTHTTTHTTTTTHHTTH) contained 6 heads and 14 tails, the observed value of r is $r = 6$, which has a likelihood on the assumption of H_0 of (approximately) .037. There are 14 outcomes which have at least as low a likelihood (on the assumption of H_0) as $r = 6$: $r = 0, r = 1, r = 2, r = 3, r = 4, r = 5, r = 6, r = 14, r = 15, r = 16, r = 17, r = 18, r = 19$, and $r = 20$. Summing the likelihoods for these outcomes, we obtain a p-value of:

$$p(r = 0|H_0) + p(r = 1|H_0) + p(r = 2|H_0) + p(r = 3|H_0) + p(r = 4|H_0) + p(r = 5|H_0) + p(r = 6|H_0) + p(r = 14|H_0) + p(r = 15|H_0) + p(r = 16|H_0) + p(r = 17|H_0) + p(r = 18|H_0) + p(r = 19|H_0) + p(r = 20|H_0) = .000 + .000 + .000 + .001 + .005 + .015 + .037 + .037 + .015 + .005 + .001 + .000 + .000 + .000 \approx .116$$

Since the p-value of this experiment is .116, the Null Hypothesis H_0 that the coin is fair can be rejected at any significance level greater than or equal to .116 (thus, H_0 could *not* be rejected at the .10, .05, or .01 levels).

If the actual outcome had instead contained 4 heads ($r = 4$) (for instance, if it had been TTTHTTTTTHTTTTTTHTTH), then the p-value of the experiment would have been:

$$p(r = 0|H_0) + p(r = 1|H_0) + p(r = 2|H_0) + p(r = 3|H_0) + p(r = 4|H_0) + p(r = 16|H_0) + p(r = 17|H_0) + p(r = 18|H_0) + p(r = 19|H_0) + p(r = 20|H_0) = .000 + .000 + .000 + .001 + .005 + .005 + .001 + .000 + .000 + .000 \approx .012$$

Then, H_0 could have been rejected at the .05 (or any $\alpha > .012$) level, but not at the .01 level (since $.012 > .01$). Again, lower p-values correspond to "stronger" grounds to reject H_0. We got a lower p-value for $r = 4$ than we did for $r = 6$ (.012 vs. .116), which is not surprising; again, the intuition is that more "extreme" outcomes consisting of lots of heads or lots of tails are good reason to reject the hypothesis that the coin is fair, whereas outcomes that consist of more balanced numbers of heads and tails are not. Thus, the further away that the observed value of r is from $r = 10$ (i.e., the value of r corresponding to an outcome with a perfectly balanced number of heads and tails), the better reason we have to reject H_0, and hence we should expect a lower p-value. Since 4 is obviously further away from 10 than 6 is, we should expect a lower p-value for the $r = 4$ outcome than we found with the $r = 6$ outcome, and this is indeed precisely what we find.

5 Likelihood Values vs. Likelihood Ratios

The first main difference between Bayesians and Frequentists is that Frequentists care about likelihood *values*, whereas Bayesians care about likelihood *ratios*. Let me briefly explain how this difference played out in the analyses from Sections 3 and 4.

In the Bayesian analysis from Section 3, the actual outcome (TTTHTHTTTHTTTTTH-HTTH) was quite unlikely on the supposition of any of the hypotheses under consideration. To reiterate,

$$p(\text{TTTHTHTTTHTTTTTHHTTH}|c_h = .2) = (.2)^6(.8)^{14} \approx 2.81 \times 10^{-6}$$
$$p(\text{TTTHTHTTTHTTTTTHHTTH}|c_h = .4) = (.4)^6(.6)^{14} \approx 3.21 \times 10^{-6}$$
$$p(\text{TTTHTHTTTHTTTTTHHTTH}|c_h = .5) = (.5)^{20} \approx 9.54 \times 10^{-7}$$
$$p(\text{TTTHTHTTTHTTTTTHHTTH}|c_h = .6) = (.6)^6(.4)^{14} \approx 1.25 \times 10^{-7}$$
$$p(\text{TTTHTHTTTHTTTTTHHTTH}|c_h = .8) = (.8)^6(.2)^{14} \approx 4.29 \times 10^{-11}$$

Obviously, each of these likelihoods is quite low; the highest likelihood (conditional on $c_h = .4$) is just over three chances in a million. But for a Bayesian, what matters is not the *values* of the various likelihoods, but rather their *ratio*. For the actual outcome of TTTHTHTTTHTTTTTHHTTH is very unlikely *regardless* of which hypothesis about the value of c_h is true. For a Bayesian, what matters is how the values of the relevant likelihoods compare *with each other*. Before the coin was flipped, the likelihood of the actual outcome was approximately:

$$(.05)(2.81 \times 10^{-6}) + (.1)(3.21 \times 10^{-6}) + (.7)(9.54 \times 10^{-7}) + (.1)(1.25 \times 10^{-7}) + (.05)(4.29 \times 10^{-11}) \approx 1.14 \times 10^{-6}$$

After the flips, on the Bayesian approach, those hypotheses with likelihoods higher than this number (i.e., $c_h = .2$ and $c_h = .4$) will receive a positive "bump" from the data and will thus increase in probability, and those hypotheses with likelihoods lower than this number (i.e., $c_h = .5$, $c_h = .6$, and $c_h = .8$) will receive a negative "bump" from the data and will thus decrease in probability (compare the prior and posterior probabilities for each hypothesis in Section 3 to verify this). Thus, for example, it is irrelevant that that the likelihood of the actual outcome was very low on the supposition that $c_h = .2$; since the likelihood of the actual outcome on the supposition that $c_h = .2$ is *higher than* the weighted average of the likelihoods on the various hypotheses about the value of c_h, the actual outcome is some evidence for the hypothesis that $c_h = .2$.[6]

By contrast, on the Frequentist approach, it's the *values* of the relevant likelihoods that matter; a p-value is just the sum of the likelihood *values* that are at least as low as the actual outcome's likelihood on the supposition of the Null Hypothesis. There are two unfortunate consequences of this fact. First, on the Frequentist approach, it is possible to reject a hypothesis when something unlikely has happened, even if that unlikely thing has nothing to do with the hypothesis. Second, and relatedly, on the Frequentist approach, it is possible to reject each member of a partition of hypotheses. The following example will illustrate both.

Suppose that a quantum particle is in a state that makes it 99% likely to emerge from the left aperture of a box, and 1% likely to emerge from the right aperture of a box. Nonetheless, let's suppose that the particle in fact emerges from the right aperture.

Now, consider the Null Hypothesis that grass is green. On the supposition of the Null Hypothesis, the likelihood that something at least as unlikely as the actual outcome would occur is just .01, since the only outcome that has at least as low a likelihood on the supposition of the Null is the actual outcome (which has a likelihood on the Null of .01). Thus, relative to the Null Hypothesis that grass is green, we calculate a p-value of .01, and hence can reject the Null Hypothesis at any $\alpha > .01$. But this is absurd; the behavior of our quantum particle has absolutely nothing to do with the color of grass, and hence isn't relevant to whether or not we should reject the hypothesis that grass is green. But since the value of the likelihood (conditional on the Null Hypothesis) that something at least as unlikely as the actual outcome would occur is low, the Frequentist methodology mandates a rejection of the Null.[7]

Relatedly, consider a partition[8] of hypotheses $\{H_1, H_2, ..., H_n\}$ about the color of grass. Since the likelihood that something at least as unlikely as the actual outcome would occur is .01 regardless of which hypothesis about the color of grass we're assuming, the same considerations as above will lead us to reject each of the H_n's. But this is odd, since

[6] For a Bayesian, E is evidence for H iff E is likelier on the supposition of H than on the supposition of $\neg H$.

[7] Maybe a Frequentist could impose additional requirements on what's allowed to be a Null, but it's not at all clear how that would go.

[8] A partition is just a set of mutually exclusive and jointly exhaustive hypotheses.

we're now in a position where we've rejected each member of a partition of hypotheses, even though we have a logical guarantee (since the hypotheses form a partition) that one member of the partition is true. This is especially odd if the partition is small (say, if there were only three colors), since we would have a logical guarantee that one out of the small number of hypotheses that we're rejecting at a low significance level is in fact true.

In fact, neither of the above points really depends essentially on the H_n's being totally irrelevant to the likelihood of the observed outcome. Let's go back to a case where we're going to flip a coin 20 times, and where we're considering a partition of hypotheses $\{H_1, H_2, ..., H_n\}$ about the value of c_h. Now, let's imagine that the outcome consists of 5 heads, 5 tails, and 10 flips where the coin lands precisely balanced on its side. This outcome, of course, is wildly unlikely regardless of the value of c_h (though it is of course slightly more likely on the supposition of $c_h = .5$ than it is on the supposition of $c_h = .1$ or $c_h = .9$, since the numbers of heads flips and tails flips were equal). Thus, regardless of which H_n we're assuming, the likelihood of an outcome at least as unlikely as the actual outcome will be staggeringly low. This will allow us to again reject any H_n, and this is *not* a case where the value of c_h is irrelevant to the likelihood of the actual outcome (since, again, the outcome is more likely on the supposition of $c_h = .5$ than it is on the supposition of $c_h = .1$ or $c_h = .9$). And, again, this will allow us to reject *each* of the H_n's, so that we have again rejected each member of a partition of hypotheses.

The Bayesian response to these problems, of course, is that the thing to focus on is the likelihood ratios. Since our quantum particle is no *more* likely to emerge from the right aperture of the box if grass is green than if grass is any other color, the fact that the quantum particle did in fact emerge from the right aperture of the box is no evidence for or against grass being green as opposed to any other color. And since the coin is no *more* likely to land on its side 10 times regardless of the value of c_h, the fact that the coin did land on its side 10 times is no evidence for or against any particular hypothesis about the value of c_h (though of course the results of the *other* ten flips may well be evidence for or against some such hypotheses).[9]

6 Describing the Data

Another issue that arises for the Frequentist approach to statistical inference is how to describe the data. In the Frequentist analysis of the coin case above, a central move was to distinguish 21 possible outcomes of the experiment, one corresponding to each possible number of times that the coin landed heads in 20 flips.

[9]Of course, in actuality, maybe the value of c_h does affect the likelihood of the coin landing on its side; perhaps, for example, very biased coins are unlikely to land perfectly balanced, whereas fairer coins are somewhat more likely to land perfectly balanced. For simplicity, I've assumed that this isn't true, but even if it is true, the point remains that the likelihood of the actual outcome is staggeringly low regardless of which H_n we're assuming. If you prefer, imagine that in ten of the flips, the coin randomly bursts into a million pieces.

But, of course, this is not the only way to distinguish possible outcomes of the experiment. For example, if we individuated outcomes not by the number of times that the coin landed heads, but rather by the precise ordered sequence of heads and tails that occurred in the 20 flips, we would end up with 2^{20} (=1,048,576) outcomes, rather than 21. Moreover, individuating outcomes this way, each outcome has *precisely the same* likelihood on the supposition of H_0—namely, $(.5)^{20}$. Thus, individuating outcomes this way, *whatever* outcome we observe, it will have a likelihood on H_0 of $(.5)^{20}$, and hence the set of outcomes with at least as low a likelihood on the supposition of H_0 as the actual outcome will contain *all of the possible outcomes*. Thus, regardless of which outcome we observe (including, for example, observing the all-tails outcome TTTTTTTTTTTTTTTTTTTT or the all-heads outcome HHHHHHHHHHHHHHHHHHHH, each of which would seem intuitively to provide strong reason to reject H_0), we will calculate a p-value of 1, and we will never be able to reject H_0 at any significance level.

Of course, there are numerous other ways that we could individuate outcomes of the 20-flip experiment. We could characterize outcomes by the greater of the number of heads and the number of tails, which would leave us with 11 outcomes. Or we could decide that we'll individuate outcomes by the number of heads, with the exception that sequences containing 13 heads and sequences containing 14 heads will be counted as a single outcome; this will leave us with 20 outcomes (for symmetry, we could also collapse 6 heads and 7 heads into one outcome, leaving 19 outcomes). Each of these individuations of outcomes could lead to different p-values being calculated on the basis of precisely the same observed sequence of coin flips.

It seems that there are two different things that the Frequentist might say about the fact that there are multiple ways to describe the same data.

First, she might opt for a *relativist* understanding of p-values. A p-value can already be thought of as a two-place function from an experiment and a null hypothesis to a number; the relativist understanding would have it that a p-value is a *three*-place function from an experiment, a null hypothesis, *and a test statistic* to a number. A test statistic is essentially just some way of describing the data; one test statistic is the number of times that the coin lands heads, and another test statistic is the precise sequence of results of the 20 flips.[10] Moreover, this relativist view would have it that there's no one "correct" test statistic (or even a set of "correct" test statistics that is a proper subset of all test statistics). Rather, the relativist view is that, just as Special Relativity entails that it only makes sense to talk about simultaneity relative to one of many "equally correct" reference frames, so too does it make sense to talk about p-values only relative to one of many "equally correct" test statistics. Thus, on the relativist view, the thing to say about the p-value that we calculated above is that it's the p-value of the 20-flip experiment relative to the null hypothesis that the coin is fair *and* relative to the choice of "number of heads" as the test statistic, but that

[10]For the formal definition of "test statistic," see, e.g., Berger and Casella 2001, p. 374.

we could have just as "correctly" calculated a different p-value, relative to a different test statistic.

But if p-values are to have any epistemic force at all, this is not a very satisfying story. What we want to know is whether we have reason to reject the null hypothesis, and how strong that reason is. We want to be able to read a study in a medical journal in which the null hypothesis that some drug is ineffective against cancer is rejected at a low significance level, so that we have some reason to believe that the drug is effective against cancer. Since we're specifically considering whether or not to reject a given null hypothesis, it of course makes sense to relativize p-values to the null hypotheses that they're characterizing. But relativizing p-values to test statistics is a completely different matter. What, for instance, are we supposed to conclude from the fact that an experiment warrants the rejection of some null hypothesis at a low significance level relative to one test statistic, but not relative to a different test statistic? It is just overwhelmingly intuitive that some test statistics matter more than others. With regard to our problem, the number of heads matters; if you flip a coin 20 times and it lands tails all 20 times, you have really good reason to reject the hypothesis that that coin is fair. And with regard to our problem, the precise sequence of heads and tails doesn't matter; when the coin lands tails 20 times, it is irrelevant that you were certain to observe some precise sequence of heads and tails with at least as low a likelihood on the supposition that the coin is far. The relativist view simply cannot accommodate this.

A defender of the relativist view might respond that Contextualism provides an important precedent in epistemology for the view that attribution of some epistemic state (such as knowledge or justification) to a subject always takes place relative to some sort of standard; on the Contextualist view, if two attributers are in different contexts, then they might both be speaking truly even if one of them says "S is justified in believing p" and the other one says "S is not justified in believing p." And merely pointing out that consequence of Contextualism is no objection to the view; the whole point of the Contextualist view is that attributions of epistemic states can be evaluated only once a standard is somehow provided. Similarly, the Relativist might claim, attribution of an epistemic property like "rejectable with a p-value of .05" to a null hypothesis can be evaluated only once a contextual parameter is provided, and that that contextual parameter is provided by the test statistic.

But even if Contextualism is true (which is of course highly controversial), not all contextual parameters are created equal. If S has a level of justification to believe p that is "good enough" by everyday standards by not "good enough" by some heightened standard, then it's somewhat plausible that "S is justified in believing p" can be true in an ordinary context and false in some heightened-standard context. But even if this is correct, this is not a story according to which the truth of sentences like "S has a such-and-such amount of justification to believe p" or "S is justified in believing p to degree .96" varies from context to context; what varies is just whether the fixed amount of justification that S has meets the contextually-determined threshold for "justified belief." It would be a very implausible

version of Contextualism, for example, that entailed that attributions of justification could only be evaluated once one of many equally good "evidential standards"—one referring to the scientific method and another referring to Tarot cards—was provided. If you are really interested in the question of whether S is justified in believing p, the scientific method standard (or something like it) matters and the Tarot card standard doesn't. Similarly, if an attribution of "rejectable with a p-value of .05" to a null hypothesis is supposed to have any real epistemic force, we are going to need a way to distinguish the test statistics that matter from the ones that don't.

The second (and in my view much more reasonable) thing for the Frequentist to say is that, even though p-values can of course be calculated only once a choice of a test statistic is made, there are better and worse choices of test statistics. The most extreme version of this view is that once you specify the experiment and the null hypothesis, there is precisely one correct test statistic that guides legitimate rejection of the null hypothesis, and all of the other test statistics are useless. Less extreme versions of this view might say that there is a spectrum of test statistics, some of which are better than others, but many of which are in some sense useful or relevant to the question of whether we should reject the null hypothesis. I won't worry too much about the distinction between the various versions of this view. I'll instead be focusing on the question of what might plausibly be taken to be better-making features of one test statistic over another.

In order to address that question, we should think about how we ought to describe the evidence that we have in general; it would be very surprising if one set of rules applied to descriptions of evidence in general, but then another set of rules applied to the statistical tests that are supposed to formally capture the significance of that evidence.

It's very clear that we can sometimes get ourselves into trouble by ignoring evidence. Suppose you learn that there's a friendly bear approaching you. That's no evidence that you're about to be killed—friendly bears don't ever hurt anyone. But suppose you had ignored the "friendly" part and just focused on a less informative description of your evidence: there's a bear approaching you. That *is* some evidence that you're about to be killed—not all bears are friendly, and the unfriendly ones sometimes attack and kill people. But if you know that there's a friendly bear approaching you, you'd go wrong describing your evidence as "there's a bear approaching me," as you'd unjustifiedly conclude that you have reason to think that you're about to be killed. In fact, you have no such reason, and you're unjustified in thinking that you have such reason. Moral of the story: you sometimes will draw the wrong conclusions by "throwing away" information that you have access to.

Can you ever go wrong in the other direction—i.e., can you ever go wrong by *failing* to throw away information? Well, in the case above, you probably have a good deal more evidence than just that there's a friendly bear approaching you. You know (let's assume) what color socks you're wearing, and you know what you ate for breakfast three days ago, and you know the exact size in square miles of Antarctica. You certainly don't run into any trouble by ignoring all of that information, and that information is just as much a part of your total evidence as is the fact that there is a bear running toward you or the fact that the

bear is friendly. But there is a big difference between, on the one hand, failing to run into trouble by ignoring some ignoring some information and, on the other, running into trouble by failing to ignore some information. As it happens, ignoring the evidence you have about the color of your socks was harmless; the color of your socks happens (in this situation) to be irrelevant to the question of whether you're about to be killed, and so you got lucky by suffering no epistemic harm as a result of failing to take your sock color into account. Or, more likely, you knew in some implicit way that your sock color was irrelevant, and so you did, at least implicitly, take your sock color into account by appropriately allowing it to have no effect one way or the other on your final belief about whether you were about to be killed.

Cases like this one are often used to motivate the so-called "Requirement of Total Evidence." Here is one recent representative statement of the Requirement: "to the extent that what it is reasonable to believe depends on one's evidence, what is relevant is the bearing of one's total evidence."[11] The idea is simple enough: since you can sometimes run into trouble by ignoring evidence you have and since you can't ever run into trouble by taking account of evidence you have, you should always consider all of the evidence that you have.

But if, in general, we ought to always take account of all of our evidence, then one might naturally think that we should always use the most specific, "informative" test statistic at our disposal, since that is the one that is going to capture all of the evidence that we have. We've been considering a case where the actual 20-flip outcome was TTTHTHTT-THTTTTTHHTTH. This description captures "all the information we have" about the outcome of the experiment. If we were, for example, to describe our evidence merely as "a 14 tails, 6 heads outcome," that would be to throw away information. For example, the latter description throws away the information that the first flip was a tails flip; the more informative description entails that fact, whereas the less informative description leaves it open. But, as already noted, the test statistic that assigns a different value to each possible 20-flip sequence (rather than "collapsing" several 20-flip sequences into one value, as the "number of heads" statistic does) leads to the result that the null hypothesis that $c_h = .5$ can never be rejected at any significance level regardless of the outcome of the experiment.

One move made by some Frequentists in trying to solve this problem is to insist that the test statistics that should be used in significance tests are the "minimal-sufficient" statistics.[12] A statistic t is sufficient relative to H_0 iff, on the supposition that H_0 is true, all of the more specific outcomes compatible with t are equally likely. So, for example, the "number of heads" statistic is sufficient in our coin case; on the supposition that H_0 is true, each 1-head outcome is equally likely (i.e., $20 \times (.5)^{20}$), and so too for each other n-heads outcome. The precise sequence of heads and tails is a sufficient statistic too; each value

[11] Kelly 2008. For a discussion of how to implement the Requirement of Total Evidence, see Kotzen forthcoming b.

[12] A suggestion along these lines was first made to me by John Roberts. For a discussion of sufficiency and minimal-sufficiency, see Howson and Urbach 1993, pp. 189–192. For a defense of the minimal-sufficiency move, see Seidenfeld 1979.

of that statistic is compatible with only one possible outcome, so sufficiency is trivially secured. A *minimal*-sufficient statistic, then, is a sufficient statistic such that any loss of information would destroy its sufficiency. The precise sequence of heads and tails isn't minimal-sufficient, since the "number of heads" statistic contains less information and is still sufficient. The idea behind this move seems to be that, since a minimal-sufficient statistic partitions the outcome space into equivalence classes of outcomes that have equal probability on the supposition of H_0, it contains "all the information we need" when we're trying to test the truth of H_0.

One worry about this move is that, as argued above, the Requirement of Total Evidence seems to mandate using the *most* informative information at our disposal, and minimal-sufficient statistics are designed to be *less* than maximally informative. The Requirement of Total Evidence is compatible with its sometimes being *harmless* to use a less-than-fully informative description of our evidence, but it is incompatible with its ever being *harmful* to use a maximally informative one, which is what the minimal-sufficiency move entails. Moreover, even in cases where using a minimal-sufficient statistic generates the results that the Frequentist wants, there doesn't seem to be any *independent* justification (analogous to the justification provided by the Requirement of Total Evidence for using maximally informative statistics) for using a minimal-sufficient statistic to partition the outcome space—i.e., a justification that doesn't just rely on the legitimacy of the Frequentist methodology. Why should we partition the outcome space into equivalence classes of outcomes that are equally likely on the supposition of H_0? Why not partition the outcome space into equivalence classes on the basis of some other consideration? Presumably, because what matters about an outcome is its "extremeness," where "extremeness" is being identified with the outcome's probability on the supposition of H_0. But that answer relies on the claim that an outcome's "extremeness" relative to other outcomes is the right way to think about the evidential impact of that outcome, which is precisely what the Frequentist assumes. Thus, the Frequentist is unable to offer any non-question-begging justification for the use of minimal-sufficient statistics. Finally, minimal-sufficiency seems to work correctly in the coin case only if H_0 is the hypothesis that the the value of c_h is something other than .5. On the supposition that $c_h = .6$, say, each of the n-heads outcomes has the same probability, so "number of heads" is a sufficient (indeed: minimal-sufficient) statistic. But on the supposition that $c_h = .5$, each sequence of heads and tails has the same probability, so the only minimal-sufficient statistic is the trivial one that contains no information at all.[13]

For Bayesians, describing the data is no problem; Bayesians obey the Requirement of Total Evidence and always conditionalize on the strongest statement of the evidence that they have access to. Thus, Bayesians never run into trouble by "throwing out" evidence. As already observed, it is often the case that much of your evidence is irrelevant, but here

[13]Moreover, though this point is merely *ad hominem*, Pratt 1965 (pp. 169–170) points out that many of the most prominent Frequentist statistics—χ^2, t, and F, for example—are not even sufficient, never mind minimal-sufficient.

the Bayesian is again rescued by her reliance on likelihood ratios rather than likelihood values. For suppose that my total evidence in the coin case is that the actual outcome was TTTHTHTTTHTTTTTHHTTH *and* that I'm wearing blue socks. Since the color of my socks and the outcome of coin flips are probabilistically independent, the likelihood of my total evidence on the supposition that $c_h = x$ is just

$$p(\text{TTTHTHTTTHTTTTTHHTTH}|c_h = x) \times p(\text{I'm wearing blue socks}|c_h = x).$$

And since the likelihood that I'd be wearing blue socks is independent of the value of c_h, we can express this quantity as

$$p(\text{TTTHTHTTTHTTTTTHHTTH}|c_h = x) \times k$$

for some constant k (which of course does not depend on x). Thus, the *ratio* of likelihoods of the actual 20-flip outcome, conditional on various hypotheses about the value of x, will be unaffected by the inclusion of the fact that I'm wearing blue socks in my total evidence; the constant k will cancel out of the relevant likelihood ratios, and those ratios will thus have the same value that they would have had if the information about my sock color had not been included in the statement of my total evidence. For a Bayesian, this is what it is for some fact in your total evidence to be irrelevant to your credence in various hypotheses about the value of c_h.

7 Actual and Non-Actual Likelihoods

Another one of the main differences between the Bayesian and the Frequentist approaches to statistical inference is that Bayesians think that only the likelihoods of the *actual* outcome on various hypotheses matter, whereas Frequentists think that the likelihoods of various *non-actual* outcomes can matter too. In the Bayesian analysis of our coin case above, the only likelihoods that mattered were the likelihoods of the *actual* outcome, TTTHTHTT-THTTTTTHHTTH, on the supposition of the various hypotheses under consideration. By contrast, for the Frequentist, a p-value is the probability, assuming the Null, that an outcome *at least as improbable as the actual outcome* would occur; thus, for the Frequentist, it matters how likely various *non-actual* outcomes (namely, those non-actual outcomes which were at least as improbable on the supposition of the Null as the actual outcome was) were on the supposition of the Null.

Let's call likelihoods of the actual outcome on various hypotheses **actual likelihoods**, and let's call likelihoods of non-actual outcomes on various hypotheses **non-actual likelihoods**. One odd consequence for the Frequentist of the relevance of non-actual likelihoods is that there can be cases where the "same data" justifies different conclusions by different investigators. Suppose, for instance, that Anne and Bob start off in the same epistemic

situation with respect to the coin, and they decide that they are going to collect evidence relevant to the value of c_h. But they disagree about how to collect that evidence; Anne wants to flip the coin 20 times, and Bob wants to flip the coin until it has landed heads 6 times. Anne and Bob can't come to an agreement on when to stop flipping the coin, so they each decide that they'll just stop observing coin flips once they've collected all the data that they each want. In other words, if the coin lands heads 6 times before the 20^{th} flip, Bob will stop watching (since he's only interested in data up until the 6^{th} head) and Anne will keep flipping the coin until it has been flipped 20 times. If, on other hand, the coin still hasn't landed heads 6 times after 20 flips, Anne will stop watching (since she's only interested in data up to the 20^{th} flip) and Bob will keep flipping the coin until it has landed heads 6 times.[14]

So, Anne and Bob start flipping the coin, and as things turn out, both observe the following 20 flips (as before): TTTHTHTTTHTTTTTHHTTH. Since the 20^{th} flip of the coin is also the flip on which the coin lands heads for the 6^{th} time, both Anne and Bob stop collecting data at the same time, at which point they seem to have exactly the same evidence: namely, that the first (and only) 20 flips of the coin landed in the sequence TTTHTHTTTHTTTTTHHTTH.[15]

[14] You might worry about the fact that this process could go on forever, since Bob might keep flipping forever without ever seeing six heads. But there's nothing to worry about here; if you want, just imagine that Bob decides to stop flipping after observing one billion tails if he still hasn't seen 6 heads. This process will definitely end (after at most 1,000,000,005 flips), but the analysis I give here of the infinite case will be so close to the analysis of this case so as to make practically no difference.

[15] As Marc Lange points out to me, the Frequentist might object to the characterization of Anne and Bob as having the "same evidence"; she might claim that whereas Anne has the evidence "I observed that the coin landed TTTHTHTTTHTTTTTHHTTH *when I decided to observe 20 flips of the coin*," Bob has the evidence "I observed that the coin landed TTTHTHTTTHTTTTTHHTTH *when I decided to observe flips up until the* 6^{th} *head*," and that those two pieces of evidence are distinct. Compare: suppose Anne and Bob had both sampled balls from an urn containing large and small balls, but that Anne *randomly* sampled whereas Bob used a net with mesh that was too wide to catch large balls. Even though they're both in possession of the evidence "I drew a large ball from the urn," it's obvious that they really have *different* evidence, and that Anne's selection really is evidence that the urn contains mostly large balls (since she easily could have drawn a small ball if one had entered her net), whereas Bob's isn't (since he couldn't have drawn a small ball). But in the case with the balls in the urn, even the Bayesian will agree that Anne and Bob possess different evidence, since even once we fix the composition of the urn, the agent's sampling method *affects how likely she is to draw a large ball*. By contrast, the agent's stopping rule does not affect how likely she is to collect the relevant evidence; regardless of whether you have Anne's stopping rule or Bob's, you are equally likely to observe the sequence TTTHTHTTTHTTTTTHHTTH. And it seems as though we need a principled story about when two pieces of evidence are distinct. Suppose that someone had a view according to which the outcome TTTHTHTTTHTTTTTHHTTH yielded different conclusions based on whether the coin-tosser had blonde or brown hair; when confronted with the objection that her view leads to the "same data" rationalizes different conclusions for different people, she insists that her view doesn't have that consequence, since TTTHTHTTTHTTTTTHHTTH *observed by a blonde-haired person* isn't the same outcome as TTTHTHTTTHTTTTTHHTTH *observed by a brown-haired person*. Clearly, something has gone wrong, and Bayesian can say precisely what has gone wrong: outcomes don't count as distinct if they have the same likelihood on the supposition of the relevant hypotheses, and once we fix the truth of the relevant hypotheses, a blonde-haired person is just as likely to observe TTTHTHTTTHTTTTTHHTTH

Since the Bayesian approach doesn't take non-actual likelihoods into consideration, the fact that Anne and Bob had planned to use different "stopping rules" is irrelevant. They started off in the same epistemic situation with respect to the coin (which presumably entails having the same prior credence distribution over possible values of c_h), and they collected the same data, which had precisely the same likelihood for both of them on the supposition of any possible value for c_h. So, on the Bayesian approach, Anne and Bob end up in the same epistemic situation with respect the coin. And this is the intuitive result; notwithstanding the fact that they *would have* collected different data if things had gone differently, they *actually* collected precisely the same data. Thus, if they started off in the same epistemic situation, and they acquired precisely the same evidence, the intuitive thought is that they should end up in the same epistemic situation too.

But that is not how things go on the Frequentist approach. After all, even though the actual outcome of 6-heads-14-tails was a possible outcome for both Anne and Bob, the other possible outcomes for each of them was very different. Anne could have (even though she didn't) observe a 5-heads-15-tails outcome, or a 7-heads-13-tails outcome, since her plan was to observe the first 20 flips of the coin no matter what. So, for Anne, there were 21 possible outcomes, 14 of which were at least as unlikely on the supposition of H_0 as the actual outcome. By contrast, for Bob, the possible outcomes were 6-heads-0-tails, 6-heads-1-tail, 6-heads-2-tails, etc,[16] most of which were at least as unlikely on the supposition of H_0 as the actual outcome. Whereas Anne will calculate a p-value of .1155 (as before), Bob will calculate a p-value of .0318.[17] Thus, Bob can reject H_0 at, for instance, the $\alpha = .05$ level (since $.0318 < .05$), whereas Anne cannot (since $.1155 > .05$). Again, this is an odd result, given that Anne and Bob started off in the same epistemic situation with respect to the coin and observed precisely the same 20 coin flips and nothing else.

A different way to dramatize this same point makes use not of two different stopping rules, but rather of two different ways of "accessing" the data. Suppose that Anne and Bob both know that the coin is going to be flipped exactly 20 times. But suppose that neither of them actually witnesses the coin flip; instead, they each send an assistant to witness the coin flips and report back—Anne sends Albert and Bob sends Betty. Albert and Betty observe the 6-heads-14-tails outcome, and go back to communicate the results

as a brown-haired person is. By contrast, if a Frequentist were to pursue the line that outcomes can be distinct simply because of features of the agent that do not affect the likelihood of those outcomes, it's not at all clear how he can explain what has gone wrong in the appeal to hair color to distinguish outcomes.

[16] If we modified Bob's plan so that he would stop after either the coin landed heads 6 times or landed tails one billion times, then we'd have to add as possible outcomes: 0-heads-one-billion-tails, 1-head-one-billion-tails, ..., 5-heads-one-billion-tails. Of course, each of these outcomes have vanishingly small likelihoods on the supposition of H_0.

[17] The number of 6-heads-n-tails outcomes is $\frac{(5+n)!}{n!5!}$, and the likelihood of a 6-heads-n-tails outcome on the supposition of H_0 is $\frac{(5+n)!}{n!5!}(.5)^{6+n}$. For the actual value of $n = 14$, this likelihood is approximately .0111. All and only outcomes with $n \geq 14$ are at least as unlikely on the supposition of H_0 as the $n = 14$ (i.e., the actual) outcome (the likelihood for $n = 0$ is approximately .0156). The sum of all likelihoods for $n \geq 14$ is approximately .0318.

to their bosses. Anne asks Albert, "How many times did the coin land heads?," to which Albert replies "Six." But Bob, for some reason, is only interested in whether or not the coin landed heads six times; if it didn't land heads precisely six times, he just doesn't care how many times it did land heads (suppose he had a large wager on the proposition that the coin would land heads precisely six times). So when Bob asks Betty about the results, he asks her "Did the coin land heads six times?," to which Betty replies "Yes." As before, Anne could have "observed" any of 21 possible outcomes, since she could have heard Albert reply to her question with any number between 0 and 20 inclusive, and she will calculate a p-value of .1155, as before. But since Bob asked specifically about the six-heads outcome, he could have only "observed" two possible outcomes: "Yes" and "No."[18] Thus, for Bob, the only outcome with at least as low a likelihood as the actual ("Yes") outcome on the supposition of H_0 is the actual outcome, so the p-value that Bob calculates will just be the likelihood of the 6-heads-14-tails outcome on the supposition of H_0, which is approximately .037. Again, Bob can reject H_0 at the $\alpha = .05$ level, say, whereas Anne cannot, even though they started off in the same epistemic situation and have identical information about the results of the coin flips.

The points above were supposed to make trouble for the Frequentist's use of non-actual likelihoods—trouble which does not attach equally to the Bayesian's exclusive use of actual likelihoods. But why might someone think, contra the Bayesian, that there are at least some situations in which non-actual likelihoods can matter too?

I'm not aware of any explicit argument for this thesis that has ever been offered. And perhaps this is just my Bayesian leanings talking, but on the face of it it's a bit odd—why should the likelihoods of events that *did not occur* matter when we're trying to assess the evidential impact of the events that *did* occur? There's a virtually limitless number of things that *didn't* happen. The total body of scientific evidence with regard to the connection between smoking and cancer could have been very different from the way it actually is; for example, the evidence *could* have supported the conclusion that smoking is generally healthful and helps prevent cancer, or that smoking has no effect on health, or that smoking gives you x-ray vision and the power to levitate. But, of course, it doesn't. Why should it matter how likely a hypothesis makes these sorts of non-actual bodies of evidence?

Nonetheless, I think that there are some things that can be said in support of the Frequentist thesis that non-actual likelihoods matter. The cases which provide the most intuitive support for this thesis, I think, are cases with the structure of Derek Parfit's "Two Lotteries."[19] Through Parfit introduced these cases in a very different context and for a

[18] Let's imagine that Betty would never volunteer information that Bob didn't ask for, and instead always replies to a yes-or-no question with either "Yes" or "No"; thus, given that Bob asked the question "Did the coin land heads six times?," it isn't a possibility that Betty might have replied, e.g., "No—it landed heads seven times."

[19] Parfit 1998.

very different purpose,[20] I think that they can be used to provide some prima facie support for the Frequentist approach. In Parfit's First Lottery, I am one of one thousand people facing potential death, and only one person can be rescued. My jailers decide to pick the one survivor by lottery, and I win. In Parfit's Second Lottery, I am the only person facing potential death, and there is a lottery to determine whether or not I will be rescued; if my jailer picks the longest of one thousand straws, I will be rescued, and if he doesn't, I will be killed. Again, I am rescued.

Parfit's judgment in these cases is that in the First Lottery, since "nothing special happened," we have no particular reason to think that the lottery was rigged: "Someone had to win, and why not me?"[21] By contrast, in the Second Lottery, Parfit's judgment is that "the result was special, since, of the thousand possible results, only one would save a life"; thus, "I could be almost certain that, like Dostoyevsky's mock execution, this lottery was rigged."[22] Of course, Parfit's judgments about these two lotteries could be disputed, but I think that they have considerable intuitive force behind them (and for the record, I agree with his judgments).

Let's try to translate what's going on in Parfit's Lotteries to our terms. In each lottery, the Null Hypothesis H_0 seems to be that the relevant lottery was fair. And in each lottery, on the supposition of H_0, the likelihood of the actual outcome (i.e., my being rescued) is .001. But Parfit thinks that in the First Lottery, we have no good reason to reject H_0, whereas in the Second Lottery, we do have good reason to reject H_0 (and to become "almost certain" that the lottery was rigged).

These cases might seem to present a problem for the Bayesian approach; after all, in the two lotteries, the likelihood of the *actual* outcome on the supposition of H_0 was the same. But if rejection of H_0 is warranted in the Second Lottery and not in the First, then it might seem as though the likelihoods of various *non-actual* outcomes are the only things that can make the difference between the two Lotteries, contra the Bayesian.

Indeed, non-actual likelihoods are precisely the resources that it would be natural for a Frequentist to appeal to in order to deliver the intuitive verdicts about the two Lotteries. It's fairly natural to think of the First Lottery as having one thousand outcomes: {I am rescued, Prisoner #2 is rescued, Prisoner #3 is rescued, ... , Prisoner #1000 is rescued}.[23] The likelihood of the actual outcome, "I am rescued," on the supposition of H_0, is .001, but so too is the likelihood of *each* possible outcome on the supposition of H_0 equal to .001. This seems to correspond to Parfit's intuition that there's "nothing special" about the outcome in which I'm rescued in the First Lottery; though that outcome has a low likelihood on the

[20] Parfit's interest in the two lotteries was to distinguish cases where low-probability outcomes are no evidence that the process was "rigged" to produce those outcomes (such as his First Lottery) from cases where low-probability outcomes are some evidence that the process was "rigged" (such as his Second Lottery). He doesn't explicitly discuss any appeal to non-actual likelihoods in order to make this distinction.

[21] Parfit 1998, p. 14

[22] Parfit 1998, p. 14

[23] Of course, we might worry about why this is the uniquely correct way to think of the outcome space—see Section 6—but this seems fairly natural here.

supposition of H_0, there are lots and lots of other, non-actual outcomes which have just as low a likelihood on the supposition of H_0 (and hence are just as "special"). To calculate a p-value here, we sum the likelihoods (on the supposition of H_0) of each outcome with a likelihood that is at least as low as .001, and we get 1; *each* possible outcome has at least as low a likelihood as the actual outcome on the supposition of H_0. A p-value of 1 corresponds to a failure to reject H_0 at any significance level, and it's intuitive that that's the right verdict; the null hypothesis that the lottery is fair should not be rejected in the First Lottery.

In the Second Lottery, by contrast, it's fairly natural to think of the lottery as having only two outcomes: {I am rescued, I am not rescued}.[24] And though the likelihood of the actual ("I am rescued") outcome on the supposition of H_0 is only .001, the likelihood of the non-actual ("I am not rescued") outcome on the supposition of H_0 is .999. Thus, the only outcome of the Second Lottery which has at least as low a likelihood as the actual outcome on the supposition of H_0 is the actual outcome itself. So, when we calculate a p-value in the Second Lottery, we get .001, which allows us to reject H_0 at a very significant ($\alpha = .001$) level. Again, this seems to correspond nicely with Parfit's intuitive judgments that the actual result was "special" in the Second Lottery and that we can therefore become almost certain that the Second Lottery was rigged.

If the Frequentist were indeed uniquely able to deliver the intuitive results in Parfit's two Lotteries, I think that this would be a significant mark in the Frequentist's favor. But I will now argue that the Bayesian can give an equally satisfying treatment of the Lotteries. Thus, Parfit's Lotteries do not provide any reason to prefer Frequentism to Bayesianism.

As observed above, in each of Parfit's Lotteries, the likelihood of the actual ("I survive") outcome on the supposition of H_0 (that the relevant lottery is fair) is .001. The Frequentist analysis above makes use of differences in non-actual likelihoods in the two Lotteries to distinguish them and deliver the intuitive result that H_0 is to be rejected in the Second Lottery and not in the First. The Bayesian, obviously, cannot make use of any non-actual likelihoods. Are there any *other* resources that the Bayesian can appeal to, consistently with his view, to distinguish the two cases?

Fortunately for the Bayesian, the answer is yes. For although the Bayesian cannot appeal to the likelihoods of non-actual outcomes on the supposition of H_0, she can appeal to the likelihoods of the actual outcome on the supposition of hypotheses *other than* H_0. Indeed, the analysis from Section 3 makes plain that it's the relative values of the likelihoods of the actual outcome on the supposition of various hypotheses that determines which of these hypotheses are supported by the actual outcome and which ones are counter-supported. And though the likelihood of the actual ("I am rescued") outcome on the supposition of H_0

[24] Though again, the complications from Section 6 arise here, especially since the lottery consisted of the gaoler drawing one of one thousand straws, so it's somewhat natural to think of the Second Lottery too as having one thousand outcomes: {The longest straw is drawn, The 2$^{\text{nd}}$ longest straw is drawn, The 3$^{\text{rd}}$ longest straw is drawn, ..., The 1,000$^{\text{th}}$ longest straw is drawn}.

is the same in each of the two Lotteries, it is *not* the case that the likelihood of the actual outcome on the supposition of *any* hypothesis is the same in each Lottery.

Take, for instance, the hypothesis H_r that the lottery is rigged (which is just the negation of H_0). In the First Lottery, what is the likelihood of the actual outcome on the supposition of H_r? Well, in the First Lottery, even if the lottery is rigged, that still doesn't settle the question of *whom* it's rigged in favor of. Since I'm just one of the one thousand participants in the First Lottery, there's "nothing special" about me, and hence no particular reason to think that the lottery is more likely to be rigged in my favor than in the favor of, say, Prisoner #219, or Prisoner #771. So even if we were to learn before the outcome of the First Lottery that it's rigged, that would be no reason to increase from .001 our credence that I'll be rescued.[25] Instead, it's natural to think that if the lottery is rigged for someone, there's a probability of only .001 that it's rigged for me; hence, in the First Lottery, the (actual) likelihood $p(\text{I'm rescued}|H_r)$=.001. By contrast, in the Second Lottery, I am "special," in that I'm the only person who is facing possible death; the jailer's drawing the longest straw leads to my being saved, whereas his drawing any of the 999 other straws leads to nobody being saved. Thus, if the Second Lottery is rigged, it *is* at least fairly likely that it would be rigged so as to lead to the result that *I* am rescued. Thus, in the Second Lottery, the (actual) likelihood $p(\text{I'm rescued}|H_r)$>.001.

How does this fact affect the Bayesian analysis of the Two Lotteries? In just the right way. In the First Lottery, $p(\text{I'm rescued}|H_r) = p(\text{I'm rescued}|H_0)$, and hence the fact that I'm actually rescued is no evidence for (or against) either H_r or H_0, which is the intuitive result. Thus, we end up with no more reason to believe that the First Lottery is rigged than we started with. In the Second Lottery, however, $p(\text{I'm rescued}|H_r) > p(\text{I'm rescued}|H_0)$, and hence my being rescued *is* evidence in favor of H_r (and against H_0). Of course, *how much* evidence my being rescued is in favor of H_r in the Second Lottery depends on how high $p(\text{I'm rescued}|H_r)$ is in the Second Lottery.[26] But it is certainly consistent with the Bayesian analysis that $p(\text{I'm rescued}|H_r)$ is a *lot* higher than $p(\text{I'm rescued}|H_0)$ in the Second Lottery, and hence that Parfit is right that we should become almost certain that the Second Lottery is rigged.

Again, the point here is not that the Bayesian provides a more compelling analysis of Parfit's Two Lotteries, though you are of course free to be more compelled by it. The point is, rather, that the Bayesian is able to give at least as compelling an analysis of Parfit's Two Lotteries as the Frequentist is able to give. Thus, notwithstanding any initial appearances to the contrary, Parfit's Two Lotteries (and other cases relevantly like it) do *not* provide any support for the Frequentist thesis that non-actual likelihoods matter. Thus, this is a "to the victor go the spoils" situation. Both the Bayesian and the Frequentist can deliver the intuitive results in Parfit's Two Lotteries, so the controversy between them will have to be adjudicated on independent grounds. I've been arguing that these independent grounds

[25] For a discussion of this point in a very different context, see White 2000.

[26] Or, really, on how large the ratio is between $p(\text{I'm rescued}|H_r)$ and $p(\text{I'm rescued}|H_0)$ in the Second Lottery—recall Section 5.

support the Bayesian perspective, and I'm not aware of any *other* sort of argument for the Frequentist's use of non-actual likelihoods.

8 The Eliminability of Subjectivity

The final difference between the Bayesian Approach and the Frequentist Approach that I'll discuss may be the most significant, but is also the difference about which I have the least to say.

Ultimately, what I think troubles so many people about the Bayesian approach is the ineliminability of subjectivity from the Bayesian methodology. Actually, this issue is complicated by the subdivision within Bayesianism into camps of so-called "subjective Bayesians" and "objective Bayesians"; roughly, the former camp thinks that any coherent[27] prior probability function that an agent might have is perfectly rational, whereas the latter camp thinks that there are additional objective constraints on rational prior probability functions that go beyond mere coherence. Obviously, this dispute is beyond the scope of this paper to address.[28] But it has been notoriously difficult to formulate a version of objective Bayesianism that is precise about what the objective constraints on rational prior probability functions are, and in a lot of cases it's very difficult to imagine what they might be.

The result is that the Bayesian (to the extent that she accepts anything short of an implausible "fully objective" form of Bayesianism) looks to be committed to the claim that if Scientist #1 and Scientist #2 come to the table with different prior opinions about some subject matter, the evidence that they jointly collect might license Scientist #1 in reacting one way and equally license Scientist #2 in reacting a different way.[29] And this runs counter to a common conception of scientific and statistical inference as domains where the evidence is "all that matters" and where there's a single objective fact about what that evidence does and doesn't support—one that doesn't depend on the prior opinions that the researchers collecting and analyzing the evidence happen to have. The Frequentist methodology, at least in part, is designed to capture that "objective" conception, since subjective prior probabilities do not enter into the calculation of p-values in the way that they clearly do enter into the calculation of Bayesian posterior probabilities.

Of course, this isn't an *argument* against Bayesianism unless we have some reason to think that subjectivity *should* be eliminable from scientific and statistical inference; the mere fact that we'd *like it to be* eliminable might give us a reason to *hope* that Frequentism is true, but not a reason to *believe* that it's true.

[27] i.e., any prior probability function that is an accordance with the axioms of probability theory

[28] For some defenses of subjective approaches, see, e.g., Jeffrey 1992, de Finnetti 1937, and de Finetti 1974. For some defenses of objective approaches, see, e.g., Jaynes 2003 and Rosenkrantz 1981.

[29] Bayesians usually appeal to "washing out of the priors" results here to show that any effect of initial differences of opinion will vanish to zero as more evidence is collected. For discussion see Hawthorne 1994.

I'm not aware of any reasons to believe that subjectivity really is in eliminable in this way, and there are some reasons to think that it's not. Hajek 2011 discusses some cases in which the data is "too good to be true"; specifically, he considers Fisher's accusation that Mendel "cooked the books" in the results of his pea experiment, arguing that the the probability by chance alone that the data would fit Mendel's theory as well as it did was .00003. Whatever the merits of this particular charge, it is clear that, at least in some cases, the fact that data fits a theory too well can be some reason to suspect that the data have been fabricated. And the Bayesian can easily handle this; presumably, we all know that data is sometimes fabricated and thus have some non-zero credence that some particular data was fabricated, and data that fits the theory "too well" is much more likely to be reported on the supposition of fabrication than on the supposition of no fabrication. The Frequentist doesn't seem to be able to accommodate this; if we ignore our (legitimate) subjective suspicion that some reported data might be fabricated, data that fits the theory "too well" simply leads to a lower p-value.

Moreover, as I've already discussed, Frequentism looks to have some subjectivist elements too, though they appear in a different form than they appear in the Bayesian system. Frequentism is not sensitive, as Bayesianism is, to differing prior opinions; even if your and my prior opinions about the the value of c_h in the coin case initially differ, you and I will still calculate the same p-value relative to the null hypothesis that the coin is fair. But as discussed in Sections 6 and 7, this is true *only as long as you and I use the same test statistic and the same stopping rule*, and there are deep difficulties that arise in the Frequentist methodology when it comes to which test statistic and which stopping rule are the "correct" ones to use. Absent some well-motivated principles that determine which test statistic and which stopping rule should be used in any given situation, Frequentism looks to contain elements that are just as "subjective" as Bayesianism; after all, if Scientist #1 and Scientist #2 come to the table with different preferred test statistics and/or different stopping rules, then again they might collect the same evidence and yet be licensed in reacting to it differently.

Again, the *nature* of the subjective elements in the two approaches is different—Bayesianism is sensitive to differences in prior opinion, whereas Frequentism is sensitive to differences in choice of test statistic and stopping rule—but I don't see any reason to find any one of these sensitivities any more or less troubling than the others. So, even if *is* a genuine constraint on a theory of statistical inference that it eliminate "subjective" elements, *both* Bayesianism *and* Frequentism run afoul of this constraint; thus, appeals to subjective sensitivity cannot be used as an argument for Frequentism over Bayesianism.

References

[Berger and Casella 2001] Berger, R. and Casella, G. (2001.) *Statistical Inference, Second Edition*. Duxbury Press.

[Bostrom 2002] Bostrom, N. (2002). *Anthropic Bias: Observation Selection Effects in Science and Philosophy*. Routledge.

[Brosamler 1988] . Brosamler, A. (1988). "An Almost Everywhere Central Limit Theorem," *Mathematical Proceedings of the Cambridge Philosophical Society* 104, pp. 561–574.

[de Finetti 1937] de Finetti, B. (1937). "La prévision: ses lois logiques, ses sources subjectives," Ann. Inst. Henri Poincaré 7, pp. 1–68. Translation reprinted in H.E. Kyburg and H.E. Smokler (eds.) (1980), *Studies in Subjective Probability*, 2nd ed., pp. 53–118. New York: Robert Krieger.

[de Finetti 1974] de Finetti, B. (1974). *Theory of Probability* Vol. 1. New York: John Wiley and Sons.

[Diaconis and Zabell 1982] Diaconis, P. and Zabell, S. (1982). "Updating Subjective Probability," *Journal of the American Statistical Association* 77(380), pp. 822–830.

[Doring 1999] Doring, F. (1999). "Why Bayesian Psychology Is Incomplete," *Philosophy of Science*, Vol. 66, Supplement. *Proceedings of the 1998 Biennial Meetings of the Philosophy of Science Association. Part I: Contributed Papers* (Sep., 1999), pp. S379–S389.

[Gill 2002] Gill, J. (2002). *Bayesian Methods*. CRC Press.

[Gillies 1990] Gillies, D. (1990). "Bayesianism versus Falsificationism," *Ratio (New Series)* III (1), pp. 82–98.

[Hacking 1965] Hacking, I. (1965). *Logic of Statistical Inference*. London: Cambridge University Press.

[Hajek 2011] Hajek, A. (2011.) "A Plea for the Improbable." Available at `http://www.google.com/url?sa=t&source=web&cd=3&ved=0CCsQFjAC&url=http%3A%2F%2Fphilrsss.anu.edu.au%2Fsites%2Fdefault%2Ffiles%2Fdocuments%2FA%2520plea%2520for%2520the%2520improbable.AAP_.final_.ppt&rct=j&q=alan%20hajek%20a%20plea%20for%20the%20improbable&ei=zwdpTqskyIS2B6WXpZsN&usg=AFQjCNHBnZfZA-Bft3EcDyiDOz5zkM3JBA`

[Hawthorne 1994] Hawthorne, J. "On the Nature of Bayesian Convergence," *PSA* Volume 1, 1994, pp. 241–249.

[Horwich 1982] Horwich, P. (1982). *Probability and Evidence*. Cambridge: Cambridge University Press.

[Howson and Urbach 1993] Howson, C. and Urbach, P. (1993) *Scientific Reasoning: The Bayesian Approach, 2nd Edition*. Chicago: Open Court.

[Jaynes 2003] Jaynes, E. *Probability Theory: The Logic of Science*. Cambridge University Press.

[Jeffrey 1992] Jeffrey, R. *Probability and the Art of Judgment* Cambridge: Cambridge University Press.

[Kelly 2008] Kelly, T. (2008). "Evidence: Fundamental Concepts and the Phenomenal Conception," *Philosophy Compass* 3, pp. 933–55.

[Kotzen forthcoming a] Kotzen, M. (forthcoming). "Multiple Studies and Evidential Defeat.," forthcoming in *Noûs*.

[Kotzen forthcoming b] Kotzen, M. (forthcoming). "Selection Biases in Likelihood Arguments," forthcoming in *British Journal for the Philosophy of Science*.

[Pratt 1965] Pratt, J. "Bayesian Interpretation of Standard Inference Statements," *Journal of the Royal Statistical Society* 27B, pp. 169–203.

[Parfit 1998] Parfit, D. (1998). "Why Anything? Why This?," *London Review of Books* 20(2), pp. 24–27; 20(3), pp.22–25. Reprinted in Crane and Farkas, eds., Metaphysics: A Guide and Anthology.

[Rosenkrantz 1981] Rosenkrantz, R. *Foundations and Applications of Inductive Probability*. Atascadero, CA: Ridgeview Publishing.

[Seidenfeld 1979] Seidenfeld, T. *Philosophical Problems of Statistical Inference*. Dordrecht: Reidel.

[Sober 2002] Sober, E. (2002). "Bayesianism—Its Scope and Limits," in Richard Swinburne, ed., *Bayes's Theorem* (Oxford: Oxford University Press): pp. 21–38.

[van Fraassen 1980] van Fraassen, B. (1980). *The Scientific Image*. Oxford: Oxford University Press.

[van Fraassen 1999] van Fraassen, B. (1999). *Laws and Symmetry*. Oxford: Clarendon Press.

[White 2000] White, R. (2000). "Fine-Tuning and Multiple Universes," *Noûs* 34, pp. 260–76.

[Whitehead 1993] Whitehead, J. (1993). "The Case for Frequentism in Clinical Trials," *Statistics in Medicine* 12(15-16), pp. 1405–1413.