Introductory Remarks on Metastatistics for The Practically Minded Non–Bayesian Regression Runner

John DiNardo*

July 22, 2008

1 Introduction

"Everything has already been said, but perhaps not *b*y everyone and *t*o everyone."[1]

The purpose of the somewhat silly title of this essay is to warn the reader what not to expect. This is not intended as a "proper" introduction to metastatistics, which I could not write, of which there are several very good ones.[2] Given the enormous amount of writing on the subject, it is not surprising then that none of the ideas or arguments will be original. [3]

An even sillier title that some might use to describe the following is: "A jaundiced appraisal of some extreme Bayesian views by someone who just doesn't get it."

That is, of course, not my intent. Rather, I think that it is sometimes useful for the practically–minded non–Bayesian regression runner (like myself) to consider some of the basic "philosophical" issues at the heart of statistics and econometrics. My purpose is also to bring some of the issues debated in metastatistics or the "philosophy of induction" literature "back to earth" from the somewhat airy realms in which they often dwell and toward the more messy realms of the low sciences addressed to an audience, like myself, who aren't philosophers but don't think that philosophy is necessarily a synonym for "useless." It is true that the discussion is often very mathematical, sometimes filled with obscure polysyllabic words of Greek or Latin origin, and pages and pages of definitions where the reader is expected to suspend disbelief before something of practical import seems to enter the discussion. Partly as a consequence, I will talk about ideas that a philosopher would define with much more precision: if you have philosophical inclinations, consider yourself forewarned.

That discussion about metastatistics often dwells in the "airy realms" is unfortunate. First, many of the issues discussed in this literature are of practical import. In economics, Bayesian approaches are becoming increasing popular and in the United States, the Food and Drug Administration (FDA) issued a

*NBER and University of Michigan.

[1]The original author(s) of the quote are unknown. "The very model of the anonymous aphorism" Koenker (2007).

[2]The number of Bayesian discussions are too numerous to list; nearly every book by a Bayesian has some discussion of metastatistics. A few books I found helpful:Berger and Wolpert (1988), deFinetti (1974), Earman (1992), Good (1983), Howson and Urbach (1993), Joyce (1999), Keynes (1921), Savage (1972). Economists in particular may find Poirier (1995) useful for it's comparative approach, as well as Zellner (1984). Non–Bayesian discussions are not nearly as numerous but there are still many useful ones. Hacking (1965, 2001) are excellent introductions, as are Mayo (1996) and Venn (1888). Mayo (1996) has helped inspire a large literature trying, among other things, to provide a "philosophy of experiment" Useful articles with a broad focus include Freedman (1995) and LeCam (1977). The former includes some nice examples where economists and sociologists come off rather badly.

Occasionally all sides agree to get together and sometimes even agree to discuss issues. The famous "Savage Forum" (Savage et al., 1962) is a nice introduction to a lot of the issues. Kyburg Jr. and Thalos (2003) has a nice collection of different approaches.

[3]Even the term "introductory" is not mine. Hacking (1983) wrote "Introductory topics should be clear enough and serious enough to engage a mind to whom they are new, and also abrasive enough to strike sparks off those who have been thinking about these things for years."

call for comments on a proposal about increased use of Bayesian methods (Food and Drug Administration, U.S. Department of Health and Human Services, 2006). Second, it seems to me that much of the debate among practically minded researchers is rooted in (frequently) unstated assumptions about the underlying philosophical justification for statistical procedures being debated. Consider the following statement of the advantages of adopting a Bayesian approach to FDA testing:

1. If we turn to Bayesian methods, difficult issues will be discussed in the right way by the right people.
2. Some of the dilemmas that FDA decision makers face are artifacts of the (non–Bayesian) statistical methods they use, and not due to demands of the scientific method.
3. The Bayesian perspective provides the best way to think about evidence (Goodman, 2004).
4. [In contrast to the usual approach] the Bayesian approach is ideally suited to adapting to information that accrues during a trial, potentially allowing for smaller, more informative, trials and for patients to receive better treatment. Accumulating results can be assessed at any time, including continually, with the possibility of modifying the design of the trial: for example, by slowing (or stopping) or expanding accrual, imbalancing randomization to favour better-performing therapies, dropping or adding treatment arms, and changing the trial population to focus on patient subsets that are responding better to the experimental therapies (Berry, 2006).

Such arguments are becoming increasingly common in domains outside of medicine well and are most easily understood with some of the metastatistical background.

1.1 Life, Death, and Statistical Philosophy: An Example

The issue of whether to use Bayesian or non–Bayesian methods has sometimes quite literally involved life or death issues. The case of ECMO (extracorporeal membrane oxygenation) is a useful example for those skeptical of the potential importance of the debate.

ECMO was a therapy developed for use with infants with persistent pulmonary hypertension: an ECMO machine circulates blood through an artificial lung back into the bloodstream. The idea is described as providing adequate oxygen to the baby while allowing time for the lungs and heart to rest or heal. The mortality rate using conventional therapy was believed to be 40% (Ware, 1989), although there is debate about whether that number was reasonable.[4] A possibly important consideration is that the notion of providing additional oxygen for infants was not obviously "safe." See the British Journal of Opthamology (1974) and Silverman (1980), for example, for a discussion of the case of "oxygen therapy" for infants which, far from being harmless, caused blindness.[5]

Concern about the ethics of a conventional randomized trial (RCT), where half the patients are randomized into treatment and half to control, led the surgeons who had developed the therapy to use a "randomized

[4]Paneth and Wallenstein (1985) observe, for example, that the survival rate among the 34 children who were considered for the trial, but did not enter because of a failure to meet one of the threshold criteria, was 100%.

[5]It is also helpful to observe that the "prior" view of most ophthalmologists was that supplemental oxygen therapy was not a potential cause of Retrolental Fibroplasia (RLF) (now referred to as Retinopathy of Prematurity). From British Journal of Opthamology (1974):

> In the early days of research into the cause of RLF it was not uncommon at any meeting where oxygen was suggested as the cause, for an indignant ophthalmologist to rise from the floor and report a typical case where to his certain knowledge no supplemental oxygen was given. He would then sink back convinced that he had delivered the coup de grace to the oxygen theory. Equally challenging were those who claimed to have seen the condition in full-term infants, which seemed to deny any special vulnerability of growing retinal vessels. Although we now know these claims to have been valid, at the time they were stumbling blocks to the early acceptance of the vital importance of prematurity and oxygen.

Much like the case of ECMO, the debate continues, as does the need for randomized controlled trials. Also, like ECMO, the debate has moved to more subtle questions, for example, about the appropriate threshold for starting oxygen in very low birth weight children(Askie and Win, 2003, Davis et al., 2004, Hansmann, 2004, Shah, 2005, Silverman, 2004, Vanderveen et al., 2006).

play–the–winner" statistical method to evaluate the treatment.The purpose of this convoluted randomization scheme was to evade "the ethical problem aris[ing] from the fact that during a 'successful' randomized clinical trial (i.e., one that demonstrates a significant advantage to one treatment) about half of the trial subjects will receive a treatment which, at the end of the trial, will be known to be inferior. The recipients of the inferior treatment are individuals whose own outcomes are, in some sense, sacrificed to the greater good of knowing, with far more certainty than before the trial, the value, lack of value, or actual harm of the treatments under investigation."(Paneth and Wallenstein, 1985)

The randomization procedure is too elaborate to be described fully, but this gloss should be sufficient.[6] The essence of their "modified randomized play–the–winner" method is that "the chance of randomly assigning an infant to one treatment or the other is influenced by the outcome of treatment of each patient in the study. If one treatment is more successful, more patients are randomly assigned to that treatment"Bartlett et al. (1985).

Call ECMO "Treatment A" and conventional treatment "Treatment B". Initially, a group of biostatisticians prepared a sequence of blinded random treatment assignments. When the outcome of a treatment was known, this information would be sent to the biostatisticians, who would then create another sequence of blinded random treatment assignments; the probability of being assigned to A or B, however, was now a function the success or failure of the treatment. In their study,

1. The first infant – with even odds – was randomly assigned to ECMO and survived.
2. The second infant – again with even odds – was randomly assigned to conventional treatment and died.
3. The third infant – with better than even odds in favor of being placed in ECMO as a result of the first two experiences – was randomized to ECMO and survived.
4. With now even higher odds the next infant was randomized to ECMO and survived.

This continued until there were a (prespecified) total of 12 events. The result of this unusual randomization was that only one child was randomized to the conventional treatment and the 11 others received the ECMO treatment.

The outcome of this experiment was that the 11 infants randomized to ECMO treatment survived; the one infant randomized to conventional treatment died. The debate revolved around whether the evidence from that trial and the previous history of non–randomized studies was "sufficient" or whether any other studies involving randomization were necessary. The researchers were reluctant to conclude that the single trial and the previous studies using "historical controls" were enough. Ware (1989), among others, observed that the randomization wasn't satisfactory and that one couldn't rule out other explanations for the observed outcomes. For instance, the sole infant not randomized to treatment, was coincidentally, the most severely ill patient in the study. The implication was that, had this one patient been randomized to ECMO, it is quite likely the child wouldn't have survived either.

Berry (1989), an advocate of Bayesian methods, harshly condemned the decision to continue further study as unethical.[7]

Most of the debate focused on the structure of the randomization, and revolved around a very narrow "binary" question: did ECMO "work" or, possibly, "what was the probability that ECMO works?" Both sides focused on whether the answer was "yes" or no. The debate did not include, for example, a heated discussion about the necessary prerequisites to be considered "eligible" for treatment. Even if the researchers had used a more conventional randomization scheme, the study would not have been able to provide a good answer to *that* much more difficult question.[8]

[6]See Bartlett et al. (1985), Wei and Durham (1978) and Zelen (1969) for a complete description of the variant of the "randomized play-the-winner" statistical method used.

[7]See the several comments in *Statistical Science* Volume 4, Number 4, 1989 and Ware's rejoinder in that issue. See Bartlett (2005) for a review of some of the history by one of the surgeons. The ethical issues don't end there, see also Couzin (2004): "Some companies seek out Berry Consultants [a small company founded by Bayesian advocate Donald Berry and his son] in the wild hope that a drug or device that's performed poorly in traditional trials can somehow undergo a Bayesian resurrection. (Such a 'rescue analysis' is rarely a possibility, both Berrys agree.)"

[8]Indeed, while ECMO is used much more liberally today, *who* should get ECMO is a subject of considerable controversy

1.2 The Metastatistics Literature

It is likely that much of the philosophical discussion on "induction" or "metastatistics" is somewhat unfamiliar to regression runners – it was to me. Moreover, often the metastatistics debate seems to involve few participants of the practical sort. As a consequence, many of the case study examples debated by philosophers of induction or statistics are drawn from physics; I am sure this is true in large part because physics has had some success – it is easier to debate "how to get the answer right" in a science when a consensus exists that, at some point, someone got it right. Such cases are rare (non–existent?) for low sciences like medicine and economics. Part and parcel of this general tendency, the types of problems considered in the metastatistics literature often seem far removed from the types of problems confronted by economists of my stripe – the "practically minded regression runner."[9]

When (by accident) I began reading about the philosophy of statistics I was surprised to discover

1. the vehemence of the debate, and
2. the almost near–unanimous consensus that almost everything someone like me – a "practically minded non–Bayesian regression runner" – understands about statistics is wrong or profoundly misguided at best.

As to (2), consider the "stupid" inferences people like myself are "supposed to draw" on account of not adopting a Bayesian point of view. One example is inspired by an example from Berger and Wolpert (1988). Consider computing the standard error of a measurement that occurs in the following way:

Flip a fair coin.

- If heads, use measuring device A for which the measurement is distributed normally with variance one and expected value equal to the truth.
- If tails, use measuring device B which has zero measurement error.

What is the *right* standard error if B is chosen? Although I had not given the matter a lot of thought before, it seemed obvious to me, a non–Bayesian, that the answer would be zero. Thus it came as a surprise to learn that, on some accounts, a non-Bayesian is "supposed" to give an answer of $\frac{1+0}{2}$.[10] By way of contrast, the Bayesian is described as someone who "naturally" avoids this inference, being "allowed" to "condition" on whether the measurement was made with machine A or B.[11]

As to the vehemence of the debate, LeCam (1977), a thoughtful non–Bayesian, prefaced his (rare) published remarks on "metastatistics" [12] by observing:

> Discussions about foundations are typically accompanied by much unnecessary proselytism, name calling and personal animosities. Since they rarely contribute to the advancement of the debated discipline one may be strongly tempted to brush them aside in the direction of the appropriate philosophers. However, there is always a ghost of a chance that some new development

(Allan et al., 2007, Lequier, 2004, Thourani et al., 2006). ECMO is now frequently employed but is still considered risky: "ECMO can have dangerous side effects. The large catheters inserted in the baby's neck can provide a fertile field for infection, resulting in fatal sepsis"Groopman (2007). See Groopman (2007) for a case study where ECMO was begun, but then stopped because it was the "wrong treatment".

[9] As it turns out, not even this sentiment is original. From Bickel and Lehmann (2001): "A chemist, Wilson (1952), [considering some issues in inference] pleads eloquently that 'There is a great need for further work on the subject of scientific inference. To be fruitful it should be carried out by critical original minds who are not only well-versed in philosophy but also familiar with the way scientists actually work (and not just with the way some of them say they work).' Wilson concludes pessimistically: 'Unfortunately the practical nonexistence of such people almost suggests that the qualities of mind required by a good philosopher and those needed by a working scientist are incompatible.'"

[10] Of course, a non–Bayesian would feel that $\frac{1}{2}$ is a perfectly good estimator of the variance if you can't know which machine produced the measure.

[11] Apparently, many examples of this specific type of inference can be avoided if one is a "conditional frequentist" (Poirier, 1995, page 344).

[12] The subtitle of LeCam's remarks "Toward Stating a Problem in the Doctrine of Chances" in part was an ironic twist on the title of Bayes' 1763 classic, "Toward Solving a Problem in the Doctrine of Chances."(Bayes, 1958)

might be spurred by the arguments. Also the possibly desirable side effects of the squabbles on the teaching and on the standing of the debated disciplines cannot be entirely ignored. This partly explains why the present author reluctantly agreed to add to the extensive literature on the subject.

It is also a literature which (until recently) seemed almost entirely dominated by "Bayesians" of various stripes – the iconoclastic Bayesian I.J. Good (1971) once enumerated 45,656 different varieties of Bayesianism. There are also "objective" Bayesians and radical subjectivists. We might also choose to distinguish between "full –dress Bayesians" (for whom estimation and testing is fully embedded in a decision–theoretic framework) as well as a "Bayesian approach in mufti" (Good and Gaskins, 1971). As I discuss below, this variety and depth results in part from the view that probability and statistics are tools that *can* and *should* be used in a much broader variety of situations than dreamed of by the usual non–Bayesian regression runner: "Probability is the very guide to life." The non–Bayesian rarely thinks of statistics as being an *all–purpose* way to *think* (See Surprising Idea 3 in section 2).

This is not to suggest that there is *no* non–Bayesian philosophy involving statistics. Most notably, Mayo (1996) has most recently stepped in to present a broader view of philosophical underpinnings on non–Bayesian statistics that I find helpful, especially her notion of "severe testing." And there is an older tradition as well: Peirce (1878a,b) and Venn (1888) are notable examples. The latter still remains an exceptionally clear exposition of non–Bayesian ideas; the articles by Peirce in *Popular Science Monthly* are insightful as well but probably a slightly more difficult read. Nonetheless, such examples are fewer and farther between.

In what follows, when I describe something as "Bayesian" I do not mean to suggest any writer in particular holds all the views so attributed here. There is considerable heterogeneity: some view concepts like "the weight of evidence" as important, others do not. Some view expected utility as important, others do not. This is not intended to be a "primer" on Bayesian statistics. Neither is it intended to be a "critique" of Bayesian views. There are several very good ones, some dating as far back as Venn (1888) (although some of these arguments will appear in what follows.) Indeed, I will admit that, given the types of questions *I* typically find interesting, I don't find Bayesian ideas particularly helpful (and sometimes harmful). On the other hand, I can imagine situations where others might find formal Bayesian reasoning helpful. Indeed, given the prominent role that "models" play in Economics, I am frankly a bit surprised that Bayesian techniques are not more popular than they are.

My purpose is not to do Bayesian ideas justice (or injustice!) but, rather, to try to selectively choose some implications of various strands of Bayesianism and non–Bayesianism for actual statistical practice that highlight their differences so as to be clear to a non–Bayesian perspective.

After having surveyed the metastatistics literature, one feels it is almost impossible to use the English language to label or describe the practically minded non–Bayesian regression runner.[13] When not being dismissed as belaboring under fallacious reasoning (Howson, 1997), she has been variously described as a "frequentist" – someone who is congenial to the notion of probability being about "relative frequency" or a NP (Neyman–Pearson) statistician – even though, as Mayo and Spanos (2006) observe, there is a great deal of confusion about what this means. Indeed, in my experience, most regression runners are not entirely sure what it means to be a user of "NP Theory" (which is not surprising given that it is not clear that either Neyman or Pearson practiced or believed NP statistical theory!) Most congenial is Mayo's (1996) term "error statistician" – someone engaged in "severe testing." On the other hand, as a firm adherent of LeCam's Basic Principle Zero – "Do not trust any principle" – I will settle on the term "non-Bayesian."[14]

[13] A term frequently employed instead of "metastatistics" is the philosophy of "induction" and there is even debate on whether it is meaningful to talk about inductive inference. See Neyman (1957) and LeCam (1977), as well as Hacking (2001) and Mayo (1982).

[14] LeCam's Basic Principle Zero (LeCam, 1990) was also intended to apply "in particular to the principles and recommendations listed below and should be kept in mind any time one encounters a problem worth studying." LeCam's principles seem quite sensible to me, and capture a lot of what I think non–Bayesians have in the back of their minds, including problems with the use of asymptotic approximations:

1. Have clear in your mind what it is that you want to estimate.
2. Try to ascertain in some way what precision you need (or can get) and what you are going to do with the estimate when you get it.

My hope is that consideration of some of the underlying metastatistics will make it easier to detect some sources of methodological disagreement. Put differently, one focus of what follows is to consider a claim from Mayo and Kruse (2002), that "principles of inference have consequences" for actual practice.

More on this subsequently, but to ground the discussion, let me list the types of research questions I would like to consider as the aims of the practically minded regression runner:

1. What is the "causal effect" of some new medical treatment?
2. What are the the iatrogenic effects of morphine use? Does the use of pain medicine cause more pain?
3. Does (U.S.) unionization lead to business failures?
4. Do "unions raise wages?"

as well the types of questions I am *not* going to consider

1. What is a good estimate of next quarter's GDP?
2. Does this structural model of the U.S. labor market provide representation adequate enough for the purposes of evaluating potential policies?
3. What are the causes and consequences of black culture?

In my experience, what type of questions one is interested in asking often suggests what type of statistics one finds useful. While both types of questions are routinely asked by economists, the types of problems entailed seem very different to me (even if they do not appear this way to some Bayesians.) This is not to imply that the second set of questions are necessarily illegitimate: I wouldn't want to suggest that people stop trying to estimate next quarter's GDP!

Indeed, when and where probability and statistics are most "useful" is one subject which divides many Bayesian and non–Bayesians and one that we explore in section 3.2.

2 Six Surprising Ideas and One Puzzle

It may seem hard to believe that one's views on the metaphysics of statistics have consequences. In this section I enumerate six "Surprising Ideas" that I think go to the heart of many differences between non–Bayesians and Bayesians. For my purposes, I will focus on suggestions for practice that are most frequently invoked by Bayesians or radical subjectivists that are at furthest remove from *my* own non–Bayesian views. Despite this, my goal isn't to criticize them. Indeed, if they strike *you* as sensible, perhaps you are a (closet) Bayesian!

3. Before venturing an estimate, check that the rationale which led you to it is compatible with the data you have.
4. If satisfied that everything is in order, try first a crude but reliable procedure to locate the general area in which your parameters lie.
5. Having localized yourself by (4), refine the estimate using some of your theoretical assumptions, being careful all the while not to undo what you did in (4).
6. Never trust an estimate which is thrown out of whack if you suppress a single observation.
7. If you need to use asymptotic arguments, do not forget to let your number of observations tend to infinity.
8. J . Bertrand said it this way: 'Give me four parameters and I shall describe an elephant; with five, it will wave its trunk'.

2.1 Six Surprising Ideas

1. The absence or presence of data mining strategies, specification mining, non–random sampling, non–random assignment are (should be) irrelevant to the inference of a set of data. Put differently, what could have happened, but didn't in an experiment, should make no difference to the evidential import of the experiment.

   > . . . considerations about samples that have *not* been observed, are simply not relevant to the problem of how we should reason from the one that has been observed (Jaynes, 1976, page 200).

   > Unbiased estimates, minimum variance properties, sampling distributions, significance levels, power, all depend on something . . . that is irrelevant in Bayesian inference – sample space (Lindley, 1971, page 426).

2. Pre–specified research design is a waste of time.

   > In general, suppose that you collect data of any kind whatsoever – not necessarily Bernoullian, nor identically distributed, nor independent of each other – stopping only when the data thus far collected satisfy some criterion of a sort that is sure to be satisfied sooner or later [*such as the requirement that a "t–statistic" exceed some critical value*], then the import of the sequence of n data actually observed will be exactly the same as it would be had you planned to take exactly n observations in the first place (Edwards, Lindman and Savage 1962, 238-239). (*I have added the words in brackets.*)

3. The problem of "how to reason" has been solved.

   "Determining which underlying truth is most likely on the basis of the data is a problem in inverse probability, or inductive inference, that was solved quantitatively more than 200 years ago by the Reverend Thomas Bayes"(Goodman, 1999).

   [They are mistaken,] those who have insinuated that the Doctrine of Chances . . . cannot have a place in any serious inquiry . . . [it can] shew what reason we have for believing that there in the constitution of things fixt laws according to which things happen, and that, therefore the frame of the world must be to the effect of the wisdom and power of an intelligent cause; and thus to confirm the argument taken from final causes for the existence of the Deity. It will be easy to show that the problem solved in this essay [by the Reverend Bayes] is more directly applicable to this purpose(Bayes, 1958).

4. Usual (non–Bayesian) practice is very badly wrong.

   . . . almost every frequentist [non–Bayesian] technique has been shown to be flawed, the flaws arising because of the lack of a coherent underpinning that can only come through probability, not as frequency, but as belief (Lindley, 2000).

   Why it is taking the statistics community so long to recognize the essentially fallacious nature of NP [Neyman–Pearson, or non-Bayesian] logic is difficult to say, but I am reasonably confident in predicting that it will not last much longer. Indeed, the tide already seems strongly on the turn (Howson, 1997).

   I explore the historical and logical foundations of the dominant school of medical statistics, sometimes referred to as frequentist statistics, which might be described as error-based. I explicate the logical fallacy at the heart of this system (Goodman, 1999).

5. Randomization rarely makes sense in those contexts where it is most often employed:

   > Physicists do not conduct experiments as Fisher would have them do. For instance, a simple experiment to determine the acceleration due to gravity might, say, require a heavy object to be dropped close to the earth. The conditions would be controlled by ensuring that the air is still, that the space between the object and the ground is free of impediments,

and so on for other factors that are thought to interfere with the rate at which the object descends. What no scientist would do is to divide the earth's surface into small plots and select some of these at random for the places to perform the experiments. Randomizers might take one of two attitudes to this behavior of scientists. They could either say it is irrational and ought to be changed or else claim that experiments in physics and chemistry are, in some crucial respect, unlike those in biology and psychology, neither of which would appear to be very promising lines of defence (Urbach, 1985, page 273).

6. "Probability does not exist.

   The abandonment of superstitious beliefs about the existence of the Phlogiston, the Cosmic Ether, Absolute Space and Time, . . . or Fairies and Witches was an essential step along the road to scientific thinking. Probability, too, if regarded as something endowed with some kind of objective existence, is no less a misleading misconception . . . " (deFinetti, 1974, page 3). [15]

2.2 An Introductory Puzzle

One of the most unusual aspects of metastatistics is that people on different sides of the debate cite *the same example* to make the case that the other side is wrong.

Consider the following example. Mayo (1979) and Mayo and Kruse (2002) have cited it as an example of a flaw in the usefulness of Bayesian reasoning while Bayesians routinely cite such examples (for one example see (Poirier, 1995) for an argument that this is evidence of a flaw in non–Bayesian reasoning! It consists of a comparison of what inferences are justified in two different "experiments."

In both cases, suppose you are interested in the fraction of black balls μ in a huge urn (we ignore the complications arising from issues of sampling with or without replacement) that is "well–mixed" and has only red and black balls. Denote the null hypothesis as $\mathcal{H}_0 : \mu = 0.5$ and $\mathcal{H}_1 : \mu > 0.5$. Denote the random variable "number of black balls" by X and the sample size as n.

Experiment A
Method: Declare in advance that you are going to pick 12 balls randomly from the urn.
Result: 9 of the 12 balls are black. The usual estimate is $\mu = \frac{3}{4}$.

Experiment B
Method: Instead of predesignating or deciding *in advance of the experiment* that you are going to draw 12 observations, you decide that you are going to keep drawing balls from the urn until you get at least 3 red balls.
Result: You draw the third red ball on the 12^{th} attempt. 9 of the 12 are black and the usual estimate is $\mu = \frac{3}{4}$.

In both experiments, 12 balls were drawn. In both experiments, 9 of the 12 were black. There are several different "loaded" questions one can ask when comparing the two experiments.

1. Are the two "experiments" different?
2. Does the "evidential import" of the two experiments for your beliefs about the true value μ differ when presented with either experiment A or B?
3. Does your evaluation of the experiment depend on the "mental state" of the investigator?

If your instinct is that "the evidential import" of both "experiments" is the same you may be Bayesian. To many Bayesians such an example is a demonstration of a logical flaw in non–Bayesian statistics: in both cases someone has drawn 9 black balls and 3 red balls. Why should I bother to consider which experiment was being performed? If the "mental state" of the experimenter is "locked up" in his/her head and, say,

[15] "la probabilità non esiste. Labbandono di credenze superstiziose sulletere cosmico, su spazio e tempo assoluto . . . su fate e streghe, sono stati fatti essenziali del cammino del pensiero scientifico. Anche la probabilità, se considerata come una cosa dotata di una specie di esistenza obiettiva, è pure un pseudo concetto ingannatore . . . "

inaccessible by someone else analyzing the data, doesn't such a case represent a fundamental problem for the non–Bayesian? I will return to this problem below, but before I do it will be helpful to sketch out generalizations about the differences between Bayesians and non–Bayesian regarding the *role* of statistics.

3 What is Statistics Good For?

First, the Bayesian is typically more ambitious about the goals of statistics: "According to the Bayesian view, scientific and indeed much of everyday *reasoning* is conducted in probabilistic terms" (Howson and Urbach, 1993, page 17).

John Maynard Keynes, for example, an exponent of "logical probability", deployed statistics to a very diverse range of subjects including teleological questions – whether perceived order could be used to provide evidence of the existence of God. He concluded that although such questions were well suited to study by the use of Bayes Law, the problem was that such evidence could only make the existence of God more credible if it were supported by *other* evidence for her existence (Keynes, 1921, page 267).

To understand this point of view it is helpful to think of probability and statistics, for the Bayesian, as tools to bridge the gap between deductive and inductive logic.[16]

Deductive logic is about the validity of "risk free" arguments.

> All men are mortal.
> John is a man
> Conclusion: John is mortal. (*)

Such an argument is deductively valid since *if* the premises are true, then so is the conclusion. A sound argument is a valid argument that has true premises. There are many types of risky arguments. Consider the following example. Imagine you are given the option of randomly selecting an orange from a box known to contain mostly good oranges, and few bad oranges.

> Most of the oranges in the box are good.
> Conclusion: The orange I randomly select will be good. (**)

This argument (**) is risky. Even if the premise is true, the conclusion may be wrong; you may be unlucky and draw one of the few bad oranges.

While probability seems of little value for non–risky arguments such as (*), even a non–Bayesian can easily see how probability might be *helpful* for arguments such as (**). For example, if we know 90% of the oranges in the box are good, the conclusion "there is a 90% chance that the orange I select will be good" seems less risky than the conclusion "there is a 90% chance that the orange will be bad." Probability and statistics for the Bayesian can be viewed as a way to tame risky arguments and make them amenable to the types of reasoning more commonly found in situations requiring merely deductive logic.

As I discuss in section 4.2, a Bayesian is typically more comfortable thinking about the probability of most *propositions* – which can be true, false, or uncertain – than a non–Bayesian. The non–Bayesian is most comfortable thinking about probability as the relative frequency of *events*. In the above example, neither the Bayesian nor the non–Bayesian is that uncomfortable about talking about the event of a randomly chosen orange being good or bad. On the other hand, a non–Bayesian is more likely to feel unclear about a statement like "there is a 90% chance that an asteroid shower is the source of the Chicxulub impactor that produced the Cretaceous/Tertiary (K/T) mass extinction of the dinosaurs 65 million years ago."[17] The *proposition* that "the mass extinction of the dinosaurs was caused by a piece of an asteroid" is either true

[16] A nice and more complete discussion can be found in Hacking (2001). Much of what follows is an abbreviated version of Hacking's discussion.

[17] The Cretaceous/Tertiary Boundary is the boundary between the Cretaceous period and the Tertiary period. The Cretaceous period is the last period of the Mesozoic Era which ended with the sudden extinction of the dinosaurs *inter alia*.

or false.[18]It is not a statement about relative frequency, or the fraction of times that the proposition is true in different "worlds."

The divergence between the two points of view becomes clearest when we begin discussing propositions much more generally. This is because if probability is understood as being useful in induction – one version of the argument goes – it is a small step from this example to considering probability as useful *whenever* one is faced with making a risky decision. By these sorts of notions, most decisions in *life* become subject to the probability calculus because most *propositions* that are risky can and should be should be reasoned about using probability.

Indeed, once you've moved from reasoning about *beliefs* to reasoning about *decisions*, notions of "utility" can often become important. Many (including some Bayesians) have difficulty with this step: the relationship between "beliefs" and "actions" is not always obvious. I, for example, tend to think of them as rather distinct.[19] I think of Voltaire's quip – "I am very fond of truth, but not at all of martyrdom" – as a (perhaps extreme) example of the possible divergence between beliefs and actions. Hacking (1965, page 16) observes "Beliefs do not have consequences in the same way in which actions do. ... [For example] we say that a man did something in consequence of his having certain beliefs, or because he believed he was alone. But I think there is pretty plainly a crucial difference between the way in which his opening the safe is a consequence of his believing he was unobserved, and the way in which the safe's opening is a consequence of his dialing the right numbers on the combination. It might be expressed thus: simply having the belief, and doing nothing further, has in general no consequences, while simply performing the action, and doing nothing further does have consequences."

While the connections between Bayesian probability and Bayesian decision theory are a matter of debate as well, the connections seem tighter.[20] More importantly, an example from "decision theory" will, I think, highlight an important difference between Bayesians and non–Bayesians.

A useful case study comes from L. J. Savage, an important figure in the development of Bayesian ideas, who argued that the role of a mathematical theory of probability "...is to enable the person using it to detect inconsistencies in his own real or envisaged behavior. It is also understood that, having detected an inconsistency, he will remove it"(Savage, 1972, page 57). Indeed, the first seven chapters of Savage (1972) are an introduction to the "personalistic" tradition in probability and utility.

3.1 What's Utility Got to Do With It?

To me, the idea of probability as primarily a tool for detecting inconsistencies sounds strange, nonetheless, it appears to be a view held by many. Savage himself provides an interesting example of "detecting an inconsistency" and then removing it. This case study was the result of a "French" complaint about crazy "American" ideas in Economics. The Frenchman issuing the complaint, Allais (1953), wrote a hotly contested article arguing against the "American" School's view of a "rational man."[21]

[18]Even here, there is a possible non–Bayesian version of characterizing the probability: if we could re-run the world 100,000 times, in about 90% of cases an asteroid of size necessary to lead to mass extinction of the dinosaurs occurs. Perhaps ironically, on this precise question Bottke et al. (2007) seem to arrive at their conclusion this way.

"Using these [estimated] impact rates as input for a Monte Carlo code, we find there is a $\leq 10\%$ chance that the K/T impactor was derived from the background and a $\geq 90\%$ chance it came from the BAF [Baptistina Asteroid family]. Accordingly, we predict that the most likely cause of the K/T mass extinction event was a collision between the Earth and a large fragment from the Baptistina asteroid shower" page 52."

[19]In this regard, it is notable that there is a considerable body of non–Bayesian decision theory, the "Neyman–Pearson" framework being the best known. What is frequently referred to as the "Neyman–Pearson" statistical framework, however, is rarely *explicitly* invoked in most micro–empirical research even though discussions about the "power" and "size" of tests are sometimes themselves the subject of debate. See for example, McCloskey (1985), McCloskey and Ziliak (1996) and Hoover (2007a,b) for one debate on the subject.

[20]There are many subtleties about the distinctions between beliefs and actions that I am ignoring. For instance, in describing Pascal's thesis, Joyce (1999, page 21) – a "Bayesian" – is "careful to formulate [the thesis] as a *norm of rational desire* that governs the fair pricing of risky wagers [those that obey conventional axioms of probability.]" In doing so he is explicit that he is making a statement about *desires* and *not actions* (page 19). "The old guard still insists that the concept of a fair price can only be understood in terms of behavioral dispositions, but it has become clear that the theoretical costs far outweigh benefits."

[21]Ragnar Frisch's remarks, which accompanied Allias' article suggest a fairly heated debate (emphasis added):

> The problem discussed in Professor Allais' paper is of an extremely subtle sort and it seems to be difficult to

Savage, like some Bayesians, argued that maximizing expected utility is good *normative* advice. Although the ideas will probably be familiar as the "Allais Paradox", it may be a good idea to sketch the main idea. If we consider $x_1, x_2, \ldots x_k$ mutually exclusive acts, that occur with probability $p_1, p_2, \ldots p_k$ respectively where $\sum_{i=1}^{k} p_i = 1$, and we can define utility over these acts with a single utility function $U(x)$ with the "usual" properties (increasing in x, etc.), we can define expected utility as

$$E[U] = \sum_{i=1}^{k} U(x_i)p_i$$

If utility is, say, increasing in money, then a "rational" person "should" prefer the gamble that yields the highest expected utility. (Note we postpone a discussion of what probability is until the next section.)

One of the gambles Allais devised to demonstrate that maximization of Expected Utility (what Allais referred to as the "Principle of Bernoulli") wasn't necessarily a good idea, went as follows:

Imagine 100 well shuffled cards, numbered from 1 to 100, and consider the two following pairs of bets and determine which you prefer.

First Gambling Situation

[A.] You win \$500,000 if you draw a card numbered 1–11 (11% chance). If you draw a number from 12–100, you get the status quo (89% chance).

[B.] You win \$2,500,000 if you draw a card numbered 2-11 (10% chance.) Draw a number from 12-100 or 1 and you get the status quo (90% chance).

The Second Pair of Gambles

[C.] You win \$500,000 for certain.

[D.] You win \$2,500,000 if you draw a card numbered 1-10 (10% chance), \$500,000 if you draw a card from 11–99 (89% chance), and the status quo if you draw the card numbered '100'.

As Allais found (and has been found repeatedly in surveys posing such gambles) for most people $B \succcurlyeq A$ (B is preferred to A) and $C \succcurlyeq D$ and, as Savage reports, the same was true for him (Savage, 1972, pages 101–104)!

As most economists will recognize, this is a "paradox" since, from $C \succcurlyeq D$

$$U(500,000) \quad > \quad 0.1U(2,500,000) + 0.89U(500,000) + 0.01U(0)$$

and from $B \succcurlyeq A$

$$0.1U(2,500,000) + 0.9U(0) \quad > \quad 0.11U(500,000) + 0.89U(0)$$

and it is obvious that both inequalities can't be true.[22] There are two ways to handle this "paradox".

reach a general agreement on the main points at issue. I had a vivid impression of these difficulties at the Paris colloquium in May, 1952. One evening when a small number of the prominent contributors to this field of study found themselves gathered around a table under the most pleasant exterior circumstances, it even proved to be quite a bit of a task to clear up in a satisfactory way misunderstandings in the course of the conversation. The version of Professor Allais' paper, which is now published in *ECONOMETRICA* emerged after many informal exchanges of views, including work done by editorial referees. Hardly anything more is now to be gained by a continuation of such procedures. *The paper is therefore now published as it stands on the author's responsibility. The editor is convinced that the paper will be a most valuable means of preventing inbreeding of thoughts in this important field.-R.F*

[22] By simple rearranging of terms, $B \succcurlyeq A$ yields:

$$0.11U(500,000) > 0.10U(2,500,000) + 0.01U(0)$$

and from $C \succcurlyeq D$ we get:

$$0.10U(2,500,000) + 0.01U(0) > 0.11U(500,000)$$

Hence, a contradiction.

1. One possibility, (the one that appeals to me) is that – even after continued reflection – my original preferences are just fine. For me, the fact that at the stated sums of money, etc., the comparison is inconsistent with Expected Utility Theory is merely too bad for the theory, however plausible it sounds. Indeed, as is well–known, it is possible to axiomatize preferences so that Allais Paradox behavior is consistent with "rational" behavior (Chew, 1983).

2. A second possibility is to conclude that something is "wrong" with your "preferences." That was Savage's conclusion; his solution was to "correct himself."

Indeed, as befits a Bayesian, Savage analyzed the situation by rewriting the problem in an equivalent, but different way:

		Ticket Number		
		1	2–11	12–100
First Pair	Gamble A	5	5	0
	Gamble B	0	25	0
Second Pair	Gamble C	5	5	5
	Gamble D	0	25	5

After writing down the problem this way, he then observed that if he were to draw a number from 12-100 he would be indifferent between the outcomes, so he decided to "focus" on what would happen if he should draw between 1 and 11. By doing so, he decided that, in the case of a subsidiary problem –ignoring outcomes higher than 11 – the correct answer depended on whether he would "sell an outright gift of $500,000 for a 10 to 1 chance to win $2,500,000 – a conclusion that I think has a claim to universality, or objectivity." He then concluded that, while it was still true that $C \succcurlyeq D$, upon reflection $A \succcurlyeq B$, not the other way around.

As Savage himself noted: "There is, of course, an important sense in which preferences, being entirely subjective, cannot be in error; but in a different, more subtle sense they can be."

We put aside the frequently knotty subject of "prior beliefs" for the moment, and contrast this Bayesian view with a typical non–Bayesian view about "what statistics is good for".

3.2 What is Statistics Good For? – A Non-Bayesian View

> In its most restricted form [statistical] theory seems to be well adapted to the following type of problem. If two persons disagree about the validity, correctness or adequacy of certain statements about nature they may still be able to agree about conducting an experiment 'to find out'. For this purpose they will have to debate which experiment should be carried out and which rule should be applied to settle the debate. If one of them modifies his requirements after the experiment, if the experiment cannot be carried out, or if another experiment is used instead, or if something occurs that nobody had anticipated, the original contract becomes void. Since the classical theory is essentially mathematical and clearly not normative it is rather unconcerned about how one interprets the probability measures .... The easiest interpretation is probably that certain experiments such as tossing a coin, drawing a ball out of a bag, spinning a roulette wheel, etc., have in common a number of features which are fairly reasonably described by probability measures. To elaborate a theory or a model of a physical phenomenon in the form of probability measures is then simply to argue by analogy with the properties of the standard 'random' experiments.
>
> The classical statistician will argue about whether a certain mechanism of tossing coins or dice is in fact adequately representable by an 'experiment' in the technical stochastic sense and he will do that in much the same manner and with the same misgivings as a physicist asking whether a particular mechanical system is in fact isolated or not. (LeCam, 1977, page 142)

A non–Bayesian doesn't view probability as a singular mechanism for deciding the probability that a proposition is true. Rather, it is a system that is helpful for studying "experimental" situations where it might be reasonable to assume that the experiment is well–described by some chance set–up. Even when attempting to use "non–experimental" data, a non–Bayesian feels more comfortable when she has reason to believe that the non–experimental situation "resembles" a chance set up. Indeed, from a strict Bayesian viewpoint it is hard to understand why in the low sciences there is a great deal of interest in "natural experiments." Put another way, one wants to try to draw a contrast between "experience" and "experiment." In the case of the former, statistical tools may or may not be particularly helpful. Other methods for gaining insight might easily dominate. In the latter case, one generally feels more hopeful that statistical reasoning might help.

4 A Few Points of Agreement, Then ...

Statistics and probability, as we understand them today, got a surprisingly late start in the (European) history of ideas.[23] Before the 17th century a major use of the word "probability" in English was to describe a characteristic of an opinion, and dealt with the authority of the person who issued the opinion. "Thus [it could be said] Livy had more of probability but Polybius had more of truth." Or, "Such a fact is probable but undoubtedly false," relying on the implicit reference of what is 'probable' to authority or consensus. (Barnouw, 1979)

A theme that will recur frequently is the notion that *everything* in metastatistics is a topic of debate. As I discuss in section 4.2 even the definition of probability is the subject of considerable debate. However, it will be helpful to have at least some terminology to work with before enjoining the metaphysics.

4.1 Kolmogorov's Axioms

One place to begin is a review of a few of Kolmogorov's Axioms which Bayesians and non–Bayesians (generally) accept, although they interpret the meaning of "probability" very differently. Though they can be defined with much more care and generality we will define them crudely for the discrete case:

1. Given a sample space Ω of possible events $A_1, A_2, \ldots A_k$ such that:

$$\Omega \equiv \sum_{i=1}^{k} \bigcup A_i \text{ for } \quad i = 1, 2, \ldots k$$

2. The probability of an event A_i is a number which lies between 0 and 1.

$$0 < P(A_i) < 1$$

An event which can not happen has a probability of zero, and a certain outcome has a probability of 1.[24] Two events, A_1 and A_2, are mutually exclusive if $P(A_1 \bigcap A_2) = 0$.

[23] For example, although games of chance greatly antedate anything resembling modern notions of probability – "someone with only modest knowledge of probability mathematics could have won himself the whole of Gaul in a week" Anything like our modern notions of probability did not "emerge permanently in discourse" until 1660. See Hacking (1975, 1990). Perhaps not surprisingly, the history of probability is the subject of much debate as well. For one criticism of Hacking's account see Garber and Zabell (1979).

[24] Even at this point a fuller treatment would include a discussion of the problem of "logical omniscience." All I can do is cite a statment from Savage (1967):

> For example, a person required to risk money on a remote digit of π would have to compute that digit, in order to comply fully with the theory [of personal probability], though this would really be wasteful if the cost of computation were more than the prize invovled. For the postulates of the theory imply that you should behave in accordance with the logical implication of all that you know. Is it possible to improve the theory in this respect,

3. For any two mutually exclusive events probability is additive:

$$P(A_1 \bigcup A_2) = P(A_1) + P(A_2)$$

The same is true for pairwise mutually exclusive events so, for example, we can write:

$$\begin{aligned} P(A_1 \bigcup A_2 \bigcup \ldots \bigcup A_k) &= \sum_{j=1}^{k} P(A_j) \\ &= 1 \end{aligned} \tag{1}$$

If we were intending a proper introduction to probability, even here complications arise. Is k finite? To address that issue properly one would introduce a set of measure theoretic considerations.[25] But there is no need to cavil about such issues presently.

The most important observation to make is that these are, so far, mere *axioms*. At this level, they are mere statements of mathematics. Indeed, we don't even have to consider them "probabilities." They may or may not be readily associated with anything "real" in the world. As Feller (1950, page 1) explains:

> Axiomatically, mathematics is concerned solely with relations among undefined things. *This property* is well illustrated by the game of chess. It is impossible to "define" chess otherwise than by stating a set of rules. ... The essential thing is to know how the pieces move and act. It is meaningless to talk about the "definition" or the "true nature" of a pawn or a king. Similarly, geometry does not care what a point and a straight line "really are." They remain undefined notions, and the axioms of geometry specify the relations among them: two points determine a line, etc. These are rules, and there is nothing sacred about them. We change the axioms to study different forms of geometry, and the logical structure of the several non–Euclidean geometries is independent of their relation to reality. Physicists have studies the motion of bodies under laws of attraction different from Newton's, and such studies are meaningful if Newton's law of attraction is accepted as true in nature.[26]

In sum, these axioms don't commit you to believing anything in particular. One reason you might adopt such axioms (and the reason I do) is because they seem convenient and useful if you are interested in the properties of chance setups or things that resemble chance set ups.

I belabor this obvious point because I think it useful to consider that we *could* begin with different axioms. A nice example comes from Hacking (2001). Consider representing the probability of a certain event, A, as $P(A) = \infty$ and if A were impossible $P(A) = -\infty$. In such a system:

- If the event A and $\tilde{}A$ (the event "not A") have the same probability, then $P(A) = P(\tilde{}A) = 0$
- If the event A is more probable than $\tilde{}A$, then $P(A) > 0$
- If the event $\tilde{}A$ is more probable than A, then $P(A) < 0$

This too could form the basis of a theory of probability, but it is one we choose not to adopt because it seems "inconvenient" to work with and makes it more difficult to study the behavior of chance set ups.

---

making allowance within it for the cost of thinking, or would that entail paradox, as I am inclined to believe but unable to demonstrate? (As cited in Hacking (1967). The published versions seems to have omitted some of the text.)

See Hacking (1967) for a very useful discussion of the issue.

[25] For instance, although it is easy to see how to partition a set of *events* it is not always possible to see how to partition a set of *propositions*.

[26] See Einstein (1920), for a thought–provoking discussion of Euclidean geometry as mathematical statements versus Euclidean geometry as statements about things in the world. As to Feller's observation about alternatives to Newton's law of attractions, see Cartwright (1984) for a provocative discussion of how even "the laws of physics lie" and physicists often fruitfully use different and mutually inconsistent models that makes a related point.

4.2 Definitions of Probability

DeFinetti's declaration in Surprising Idea 6 that "PROBABILITY DOES NOT EXIST" may, at minimum, appear to be a bit intemperate. Indeed, it presupposes that many practically–minded non–Bayesian regression runners are in the grips of some bizarre hallucination. It will help to consider two broad classes of definitions of probability that are sometimes referred to as:

1. "aleatory" or frequency –type probabilities
2. "epistemic" or belief–type probabilities.

Aleatory probabilities are perhaps what is most familiar to the non–Bayesian. For many, the notion of any other type of probability may not have been seriously entertained. It is interesting to observe that criticism of aleatory probability began at the inception of modern statistics and as Hacking (1975, page 15) observes, "Philosophers seem singularly unable to put asunder the aleatory and the epistemological side of probability. This suggests that we are in the grip of darker powers than are admitted into the positivist ontology."

I began my presentation with Kolomogrov's axioms since everyone seems to agree on something like these axioms; disputants disagree on what they are useful for, or what, precisely, they are "about." I won't do a complete survey, but a few moments of reflection may be all that is required to consider how slippery a notion probability could be.[27]

4.3 Aleatory or Frequency–type Probabilities

When we say "the probability that a fair coin will land as heads is $\frac{1}{2}$ " we could take it as a statement of fact, which is either true or not. When we do so, we are generally thinking about probability as something that describes something that results from a mechanism that tosses coins and the geometry of the coin, perhaps. This mechanism can be described as a "chance set-up". We might go on to describe the physics of the place with our coin–toss mechanism. A mechanism that would be perfectly useful in Ann Arbor, MI might not work somewhere in the deep reaches of outer–space.

Nonetheless, most non-Bayesians, it would seem, are content to harbor little doubt that at some fundamental level – whether we know the truth or not – that it is meaningful to talk about the probability of a tossed coin falling heads. When pressed to explain what they mean when they say that the probability is $\frac{1}{2}$ that a fair coin will turn up heads, such a person might say "in the long run, if I were to repeatedly toss the coin in the same way, the relative frequency of heads would be $\frac{1}{2}$." We've yet to worry about "the long run" but even at this level, for example, we would like to exclude the following deterministic but infinite series as being a prototype for what we have in mind:

$$H \quad T \quad H \quad T \quad H \quad T \quad \ldots H \quad T \ldots$$

In such an example, if we know the last coin toss was "H" we are certain that the next coin toss will be "T". An "intuitive" definition that excludes such a possibility was given by Venn (1888), who talked about a probability as a characteristic of a *series* as "one which exhibits individual irregularity along with aggregate regularity." If we denote the number of "trials" by N and the number of times the event "Heads" occurs as $m(N)$, we might go on to define the probability of "heads" as:

$$P(\text{Head}) = \lim_{N \to \infty} \frac{m(N)}{N}$$

An apparent weakness of this definition is, of course, that infinity is rarely observed. The derisive term about such thought exercises is sometimes referred to as "asymptopia" – that suggests something both unrealistic and unattainable.[28]

[27] For a marvelous introductory exposition see Chapter 21 of Hacking (2001). For a nice more complete discussion that includes a bit more mathematical formalism and may be congenial to economists see 2.1 of Poirier (1995).

[28] There is much debate about the utility of *defining* probability as the limiting behavior of a sequence. This debate is

4.4 Objective, Subjective, or "It Depends"

Whether such a concept corresponds to something "real" or "objective" or whether it is "in the mind" is a subject on which much has been written. Sometimes such probabilities have been described as "objective" in order to contrast them with Bayesian "probabilities." However, one Bayesian objection is that there is no such thing as an "objective probability" – any such probability depends on purely subjective beliefs:

> To calculate a frequency, it is necessary to consider a repetitive phenomenon of a standardized variety. The appropriate meaning of "repetitive" and "standardized" is not obvious. To calculate a relative frequency it is necessary to subjectively define (possibly only conceptually) a class of events (known as a *collective*) over which to count the frequency. The relative frequency of the event ["Heads"] should also tend to the same limit for all subsequences that can be picked out in advance.[29]
>
> To the extent that individuals agree on a class of events, they share an objective frequency. The objectivity, however, is in themselves, not in nature. Poirier (1995)

Poirier, a Bayesian, stresses the (implicit) "subjectivity" of the frequentist notion of probability, specifically the notion of a "collective." The non-Bayesian, von Mises (1957, page 12), for example, defines a collective as a "sequence of uniform events or processes which differs by certain observable attributes, say colours, numbers, or anything else. Only when such a collective is defined, then a probability can be defined. If it is impossible to conceive of such a collective, then it is impossible to talk about probability." For von Mises, the notion of collectives with infinite numbers of entities was an *abstraction* to make the mathematical representation of reality "tractable." (Gillies, 2000, page 90) While an extensive discussion of a "collective" is beyond our scope, it is important to acknowledge that there can be "legitimate" disagreements about whether certain probabilities can be said to "exist." von Mises, argues that the reason it is possible to talk about the probability of a tossed coin turning up "heads" is because it is easy to think of the "collective"; it is not possible he says to consider "the probability of winning a battle ... [which] has no place in our theory of probability because we cannot think of a collective to which it belongs."

I personally share von Mises discomfort with defining the "probability of winning a battle", but I imagine others do not. Whether or not it would be "meaningful" to do so, or whether it "has no place in our theory of probability," the ultimate criterion in the non–Bayesian context is: "would doing so, help in understanding." The salient issue is not that different in principle than the qualms a physicist might feel about "whether a particular mechanical system is in fact isolated or not." Whether that is a "defect" of the theory of probability or whether it introduces an undisciplined element of "subjectivity" is a subject upon which there has been much philosophical debate.[30]

4.5 Epistemic Probability

The difficulties that a non–Bayesian might feel about conceiving of the appropriate collective are largely avoided/evaded when we consider a different notion of probability – epistemic. A nice place to start is a

intimately related to the debate about commending an estimator because in repeated applications and in the long run, it would do well. The canonical problem is the "single case" exception (Hacking, 2001, Chapter 22) in which we are asked to consider a situation where, for example, there are two decks of cards: one "the redder pack" has 25 red cards and 1 black card. The other "blacker pack" has 25 black cards and 1 red card." You are presented with two gambles. In the first gamble, you "win" if a red card is drawn randomly from the "redder" pack and lose otherwise ($P(\text{Win} = 25/26)$. In the second gamble, you "win" if a red card is drawn randomly from the "blacker" pack ($P(\text{Win} = 1/26)$. If you "win" you will be transported to "eternal felicity" and if you lose, you will be "consigned to everlasting woe." As Hacking and (C. S. Peirce) most people would choose the first gamble and hope, but *not* because of the long run; if we are wrong, there will be no comfort from the fact that we *would have been right most of the time*! Peirce's "evasion of the problem of induction" is to argue that we should not limit ourselves to merely "individualistic" considerations. "[Our interests] must not stop at our own fate, but must embrace the whole community. This community, again, must not be limited but extend to all races of beings with whom we can come in to immediate or mediate intellectual relation. It must reach, however vaguely, beyond this geological epoch, beyond all bounds. He would not sacrifice his own to save the whole world is, as it seems to me, illogical in all his inferences collectively. Logic is rooted in the social principle." (Peirce, 1878b, page 610-611)

[29] Such a condition rules out, for example, the deterministic series $(H, T, H, T \ldots H, T)$ discussed above.

[30] See Gillies (2000) for a nice discussion.

description from Savage, often called a "radical subjectivist."

> "You may be asking, 'If a probability is not a relative frequency or a hypothetical limiting relative frequency, what is it? If, when I evaluate the probability of getting heads when flipping a certain coin as .5, I do not mean that if the coin were flipped very often the relative frequency of heads to total flips would be arbitrarily close to .5, then what do I mean?' We think you mean something about yourself as well as about the coin. Would you not say, 'Heads on the next flip has probability 0.5' if and only if you would as soon guess heads as not, even if there were some important reward for being right? If so, your sense of 'probability' is ours; even if you would not, you begin to see from this example what we mean by 'probability'."(Savage, 1972)

What is also interesting is that instead of Kolomogrov's axioms reflecting a (possibly) arbitrary set of axioms about unknown concepts which (one hopes) resemble some real world situation, they can also be derived from "betting rules." Again quoting Savage:

> For you, now, the probability $P(A)$ of an event A is the price you would just be willing to pay in exchange for a dollar to be paid to you in case A is true. Thus, rain tomorrow has probability 1/3 for you if you would pay just \$0.33 now in exchange for \$1.00 payable to you in the event of rain tomorrow. (Savage, 1972)

As when we encountered Expected Utility, viewing probability as a device that allows one to make sensible "bets" is not *necessary.* The important distinction between aleatory and epistemic probability is that epistemic probabilities are numbers which obey something like Kolmogorov's axioms but do not refer to anything "real" in the world, but to a (possibly) subjective "degree of belief". Here's one definition from Poirier (1995, page 19):

> Let κ denote the body of knowledge, experience or information that *an individual* has accumulated about the situation of concern, and let A denote an uncertain event (not necessarily repetitive). Then the *probability* afforded by κ is the 'degree of belief" in A held *by the individual* in the face of κ.

Given this definition of probability, stating that the probability that a fair coin lands heads is *not* stating some property of a chance set up – rather it is an expression of belief about what the coin will do.[31] It is important to point out that opinions about this subject vary amongst Bayesians. I. G. Good for instance maintains that "true" probabilities exist but that we can only learn about them by using subjective probabilities. deFinetti, as we saw believes that it is unhelpful to postulate the existence of "true" probabilities.

How does this differ from the aleatory or frequency type probability we discussed above? Again quoting from Poirier (1995, page 19)

> According to the subjective ... interpretation, probability is a property of an individual's perception of reality, whereas according to the ... and frequency interpretations, probability is a property of reality itself.

Among other things, in this view, the probability that a fair coin toss is heads differs across individuals.[32]

[31] It is thus easier to understand Poirier's emphasis in the quotation above on whether the probability is "in nature" or "in themselves": that "To the extent that individuals agree on a class of events, they share an objective frequency. The objectivity, however, is in themselves, not in nature."

[32] Again I have ignored "logical" probabilities which are a class of epistemic probabilities which incorporate notion of evidence. In this view, a probability is a "rational degree–of–belief" about a proposition or a measure of the degree of "credibility" of a proposition.

4.6 Conditional Probability, Bayes' Rule, Theorem, Law?

Is it Bayes' Rule, Law, or Theorem? One of the most powerful ideas of all time, or the source of much mischief? Dennis Lindley (As cited in Simon (1997)) for example observes that "[Bayes'] theorem must stand with Einstein's $E = mc^2$ as one of the great, simple truths." Put aside the intractable issue of what the Reverend Bayes *meant*. This has been the subject of considerable controversy and study.[33]

For the typical non–Bayesian, R. A. Fisher and William Feller for example, Bayes' rule is nothing but a manipulation of the law of conditional probability.

Everyone starts with a *definition* of conditional probability:

$$P(A_i|B) = \frac{P(A_i \bigcap B)}{P(B)} \text{ if } P(B) > 0 \tag{2}$$

Provided the necessary probabilities exist, we can do the same thing in reverse

$$P(B|A_i) = \frac{P(A_i \bigcap B)}{P(A_i)} \text{ if } P(A_i) > 0 \tag{3}$$

Then there is the "conditional" version of the law of total probability: as before, let the A_j be mutually exclusive events from $j = 1 \ldots k$ and $\sum_{j=1}^{k} P(A_j) = 1$ and if $0 < P(B) < 1$:

$$P(B) = \sum_{j=1}^{k} P(B|A_j)P(A_j)$$

What this says in words is that if $P(B)$ is the probability of some event, and it can be accompanied by some of the k mutually exclusive events A_j in some way then the probability that $P(B)$ occurs is merely the sum of the different ways B can occur with A_j times the probability of $P(A_j)$.

Using equations (2), (3) and (1), rearranging and applying this last operation to the denominator yields:

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B)} \tag{4}$$

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_{j=1}^{k} P(B|A_j)P(A_j)} \tag{5}$$

So far, there seems nothing particularly remarkable. However, here the agreement ends. Consider a "Note on Bayes's Rule" by the non–Bayesian Feller (Feller, 1950, page 125):

> In [the above formulas] we have calculated certain conditional probabilities directly from the definition. The beginner is advised always to do so and not to memorize the formula [Bayes' Rule, equation (5)] ... Mathematically, [Bayes' Rule] is a special way of writing [the definition of conditional probability] and nothing more. The formula is useful in many statistical applications of the type described in [the above] examples and we have used it there. Unfortunately, Bayes's rule has been somewhat discredited by metaphysical applications ... In routine practice this kind of argument can be dangerous. A quality control engineer is concerned with one particular machine and not with an infinite population of machines from which one was chosen at random. He has been advised to use Bayes's rule on the grounds that it is logically acceptable and corresponds to our way of thinking. Plato used this type of argument to prove the existence of Atlantis, and philosophers used it to prove the absurdity of Newton's mechanics. But for our engineer the argument overlooks the circumstance that he desires success and that he will do better by estimating and minimizing the sources of various types of errors in prediction and guessing.

[33] For three mutually exclusive analyses of what Bayes' meant – and whether or not he succeeded in proving what he set out to establish – see Hacking (1965), Chapter 12, – or whether Bayes' understanding is consistent with subsequent Bayesian interpreters (beginning with the rediscovery by Laplace (1795)) – see Stigler (1982).

Feller's suggestion that the engineer will do better by minimizing the various types of *errors* is one issue where, at least rhetorically, non–Bayesians differ from Bayesians. For Feller, the focus is on using statistics (or other methods) to put ideas to the test, rejecting those that fail and advancing provisionally with those that survive. Bayes rule is a formula about *revising* one's epistemic probabilities incrementally. This distinction will become apparent when we apply Bayes rule to estimation.

4.7 Reasoning or Estimating with Bayes Rule?

Not surprisingly Bayes' rule is viewed differently by Bayesians: it is a multi (or all) purpose tool of reasoning. Consider first the version given by equation (4). To fix ideas, let us consider one example of "Bayesian Inference." In the above notation, let A_i be a specific hypothesis about the world and let B refer to some "data" that has somehow come in to our possession. For example, A_i might be the hypothesis that a coin is fair and B is the fact that you observed a single toss of the coin and it landed "heads." *Your* job is to ascertain how you should revise your beliefs in light of the data.

1. The "model" or likelihood for the behavior of N tosses of a coin is given by the following likelihood:

$$\mathcal{L}(\theta|N,h) = \left(\begin{array}{c} N \\ h \end{array} \right) \theta^h (1-\theta)^{N-h} \tag{6}$$

   As described by Poirier (1995), for example, $\mathcal{L}$ is a "window" by which to view the world – perhaps an "approximation" to the truth. We might debate what window is appropriate but in the usual context, it isn't something to be "tested" or "evaluated". Moreover the likelihood is a function which tells us "how likely we were to have observed the data we did (N,h)" given the truth of the model and a specific value of θ. (N.B. here the likelihood is a device that tells you – given the parameter θ, what is the probability of observing the occurrence h the number of heads in N tosses of a coin.)

   Instead of using the coin toss mechanism to help you randomize you are going to study the coin (and the mechanism) and learn about it.

2. The next step is to specify a prior distribution – one particularly *convenient* choice is the beta distribution. Priors are subtle things, but let us consider our beliefs about the value of θ to be describable by the following two parameter distribution:

$$\begin{aligned} f(\theta;\alpha,\delta) &= \frac{\Gamma(\alpha+\delta)}{\Gamma(\alpha)\Gamma(\delta)} \theta^{\alpha-1}(1-\theta)^{\delta-1} \\ &= \frac{1}{B(\alpha,\delta)} \theta^{\alpha-1}(1-\theta)^{\delta-1} \end{aligned} \tag{7}$$

   where $\Gamma(\cdot)$ is the gamma function and $B(\cdot)$ is the beta function. This is a very flexible distribution which can put weight on all values between 0 and 1. Figure 1 displays some of the wide variety of shapes the prior distribution can take for different values of α and δ.

   Different values of α and δ correspond to different beliefs. One way to get some intuition about what type of beliefs the parameters correspond to is to observe, for example, that the mode of the prior distribution (when it exists), for example, occurs at:

$$\frac{\alpha-1}{\alpha+\delta-2}$$

   It is sometimes helpful to think of $\alpha-1$ as the number of heads 'previously" observed, $\delta-1$ the number of tails, and $\alpha+\delta-2$ as the total number of coin flips previously observed from the experiment. On the other hand, it is not clear how someone could verify that a particular choice of prior was a good or bad description of one's beliefs.

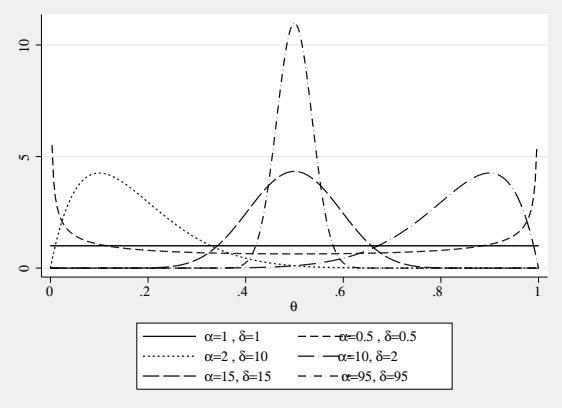

Figure 1: Different Priors Using the Beta Distribution

3. In the third step, we merely plug our prior and our likelihood into Bayes rule and what we come up with is[34]

$$\frac{1}{B((\alpha+h),(\delta+(N-h)))}\theta^{\alpha+h-1}(1-\theta)^{\delta-1+N} \tag{8}$$

Given the usual caveats, equation (8) is a statement about your personal beliefs about the value of θ, modified in light of the the observed coin–toss. The beta distribution is a nice example because it is easier than usual to characterize the resulting "beliefs."

The left panel of Figure 3 shows two different prior distributions – one labeled "less informative" and the other "very informative." The first prior distribution corresponds to Beta(199,1) and the second to Beta(2,2). A convenient fiction to appreciate these prior beliefs is to imagine that in the first case you have previously observed 200 observations, 199 which were heads. In the second case, you have previously observed 4 observations, two of which were heads. The first case corresponds to having "more prior information" than the second.

The mode of the posterior distribution, for example, occurs at:

$$\frac{\alpha+h-1}{\alpha+\delta+N-2}$$

This can be fruitfully compared to the usual non–Bayesian maximum likelihood (or method of moments) estimator which is merely the sample mean:

$$\frac{h}{N}$$

The difference between the posterior mode and the usual non–Bayesian estimator is that the former "adds" $\alpha-1$ heads to the numerator and "adds" $\alpha+\delta-2$ observations to the denominator.[35]

[34] We have omitted one detail in this exposition which is that the expression we are required to evaluate is:

$$\frac{\mathcal{L}(\theta|N,h)f(\theta)}{\int_0^1 \mathcal{L}(\theta|N,h)f(\theta)}$$

In our previous notation, the denominator corresponds to $P(B)$ or $\sum_{j=1}^{k} P(B|A_j)P(A_j)$. One "nice" feature of the beta distribution is that it serves as the "natural conjugate prior" for the binomial distribution – loosely speaking, the functional form of the prior and the likelihood are the same – this means one can treat the denominator as an *integrating constant* and it does not have to be computed directly.

[35] The posterior mode can be rewritten as a weighted average of the sample mean and a prior mean with the role of the prior vanishing as the number of actual sample observations grows large.

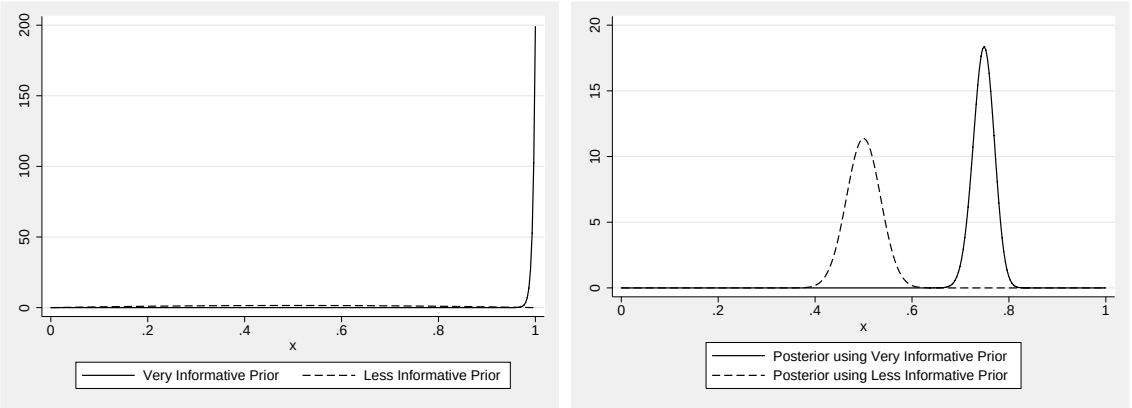

Figure 2: Different Prior and Different Posterior Distributions

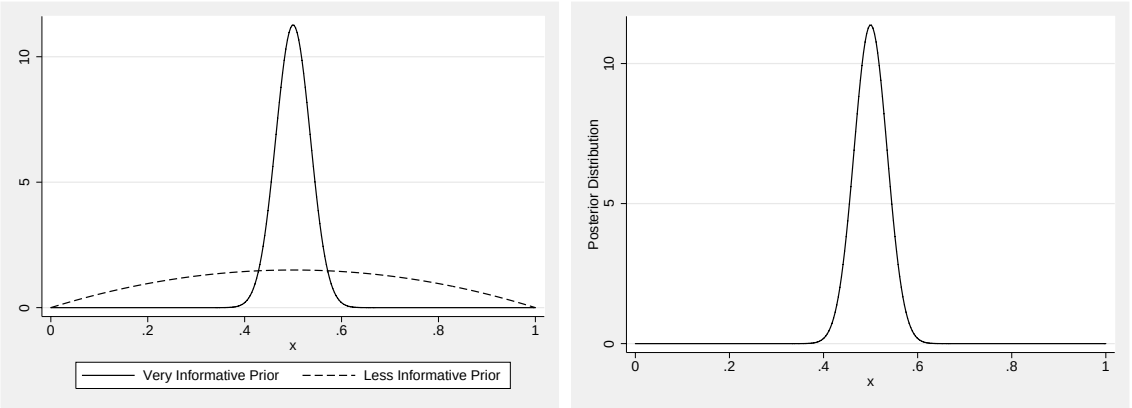

Figure 3: Different Prior Distributions, Same Posterior

To see what effect this has, the right hand panel of figure 3 shows the resulting posterior distributions updating with 200 coin tosses, 100 of which are heads.

For a slightly different type of comparison one can consider two situations:

Prior	Data
Beta(99,99)	2 heads, 2 tails
Beta(2,2)	99 heads, 99 tails

In this case, although the experiments are very different our conclusions are exactly the same.

The role of the prior distribution and the sufficiency of the posterior distribution or likelihood are among the longest standing debates in metastatistics. While a complete review is impossible some of the most frequently enumerated difficulties are:

1. There is no way to verify whether the prior one has chosen adequately characterizes one's beliefs. Also, there is no unique way to translate ignorance or "no information" into a prior distribution.[36] Consider the problem of estimating the length of a square garden which has sides of length between 1 and 5 feet, Based on this information, it seems "natural" to say that there is a 0.5 probability that the garden has sides of *length* between 1 and 3 feet. Equivalently, the information could be cast as saying that the area of the garden is between 1 and 25 square feet. In that case, it would appear just as natural to say that the probability is 0.5 that *area* of the garden is between 1 and 13 square feet. This natural

[36]There are many variants of the following example. This particular variant is slightly adapted from Sober (2002).

assignment of probability, however, implies that the probability is 0.5 that the length of the sides is between 1 and ≈ 3.61 feet ($\sqrt{13}$). However, it would be personally inconsistent to believe both claims and no principled method to reconcile the two different priors.

2. Even if a prior distribution is useful to the person holding it, it is not clear that it is useful to anyone else. LeCam (1977) observes that for the binomial experiment, for arbitrary positive constant C "if we follow the theory and communicate to another person a density $C\theta^{100}(1-0)^{100}$ this person has no way of knowing whether (1) an experiment with 200 trials has taken place or (2) no experiment took place and this is simply an a priori expression of opinion. Since some of us would argue that the case with 200 trials is more 'reliable' than the other, something is missing in the transmission of information."

5 The Importance of the Data Generation Process

5.1 An Idealized Hypothesis Test

Ultimately we would like to return to the "introductory puzzle", but before we do let us introduce some context. The value of hypothesis testing has been frequently debated among non–Bayesians, but it may help to consider an idealized notion of how it is *supposed* to be done – this version is from Kmenta (2000)–wishing to make a statement about a *population* from a "random sample":

Preamble State the maintained hypothesis [E.g. the random variable X is normally distributed with σ^2 equal to $\ldots$]

Step 1 State the null hypothesis and the alternative hypothesis [E.g. $H_0 : \mu = \mu_0$ and $H_A : \mu \neq \mu_0$]

Step 2 Select the test statistics [E.g. $\overline{X}$ based on sample size $n = \ldots$].

Step 3 Determine the distribution of the test statistic under the null hypothesis. [E.g. $\sqrt{n}(\overline{X} - \mu_0)/\sigma$ is distributed $N(0,1)$ – normal, with mean zero and variance 1.]

Step 4 Choose the level of significance and determine the acceptance and the rejection region. [E.g. *"do not reject H_0 if* $-1.96 \leq \sqrt{n}\frac{(\overline{X}-\mu_0)}{\sigma} \leq 1.96$*; otherwise reject it."*]

Step 5 Draw a sample and evaluate the results [E.g. *"the value of $\overline{X}$ is ... which lies inside (outside) the acceptance region."*]

Step 6 Reach a conclusion. [E.g. *"the sample does (does not) provide evidence against the null hypothesis"*] To distinguish between 5% and 1% levels of significance we may add the word *strong* before *evidence* when using the 1% level.

It will be worth noting Kmenta's observations about the procedure: "According to the above scheme, the planning of the test and the decision strategy are set *before* the actual drawing of the sample observations, which does not occur until step 5. This prevents rejudging the verdict to suit the investigator's wishes."

This observation comes up frequently in non–Bayesian discourse, but less frequently among Bayesians: Does the investigator want to ensure him/herself against "rejudging the verdict?" Perhaps they should "rejudge the verdict?" As we will see, this points to a notion of **severity** as being primary, as opposed to merely a concern about the correctness of the various statistical tests (although the two are not unrelated).

5.2 The Introductory Puzzle Revisited

With this in mind, we can now re-introduce the puzzle. Specifically, the puzzle arises because, by using some variant of the above procedure, under one experiment observing 9 black of 12 balls allows one to *reject* the null hypothesis; in the other, observing 9 black of 12 balls would not permit the researcher to reject the null. Some Bayesians point to this example as evidence of a flaw in non–Bayesian reasoning: why should what

is "locked up in the head" of the researcher – her intentions about what she was going to do – matter? In both cases, she has the same "data." This problem appears in many guises: in clinical trials there is a debate about what should be done if, for example, "early" evidence from a trial suggests that a drug is effective. The non–Bayesian response is that the Bayesian view misconstrues the purpose of error probabilities.

First, let's illustrate the problem. In experiment A, the question is: how often would we expect to see 9 black balls out of 12 balls under the null hypothesis.

$$\begin{aligned} P(\widehat{\mu} \geq \frac{3}{4}|\mathcal{H}_0) &\equiv P(X \geq 9|\mathcal{H}_0) \\ &= \sum_{x=9}^{12} \left(\begin{array}{c} 12 \\ x \end{array} \right) \mu^x (1-\mu)^{12-x} \\ &= \left(\begin{array}{c} 12 \\ 9 \end{array} \right) \frac{1}{2}^9 (1-\frac{1}{2})^3 + \left(\begin{array}{c} 12 \\ 10 \end{array} \right) \frac{1}{2}^{10} (1-\frac{1}{2})^2 + \ldots \\ &= \frac{220+66+12+1}{2^{12}} \\ &= \frac{299}{2^{12}} \\ &= 0.073 \end{aligned}$$

In experiment B, the question is: under the null hypothesis, what is the probability of drawing 9 or more black balls before drawing a third red ball. Let $r = 3$ be the pre–specified number of red balls to be drawn before the experiment is to be stopped. Let x index the number of black balls drawn, and let $n = x + r$.

This is a straightforward application of the negative binomial distribution where:

$$\begin{aligned} P(X \geq 9|\mathcal{H}_0) &= \sum_{x=9}^{\infty} \left(\begin{array}{c} r+x-1 \\ r-1 \end{array} \right) \mu^x (1-\mu)^r \\ &= \sum_{x=9}^{\infty} \left(\begin{array}{c} x+2 \\ 2 \end{array} \right) \mu^x (1-\mu)^r \end{aligned}$$

It is very helpful to observe in doing the calculation that

$$\sum_{x=j}^{\infty} \left(\begin{array}{c} x+2 \\ 2 \end{array} \right) (\frac{1}{2})^x = \frac{8+5j+j^2}{2^j}$$

We can then write:

$$\begin{aligned} &= \sum_{x=9}^{\infty} \left(\begin{array}{c} x+2 \\ 2 \end{array} \right) \mu^x (1-\mu)^3 \\ &= \left(\frac{1}{2}\right)^3 \frac{8+5(9)+9^2}{2^9} \\ &= \frac{1}{8}\left(\frac{134}{512}\right) \\ &= 0.0327 \end{aligned}$$

There are several points to make about these "experiments" from a non–Bayesian perspective.

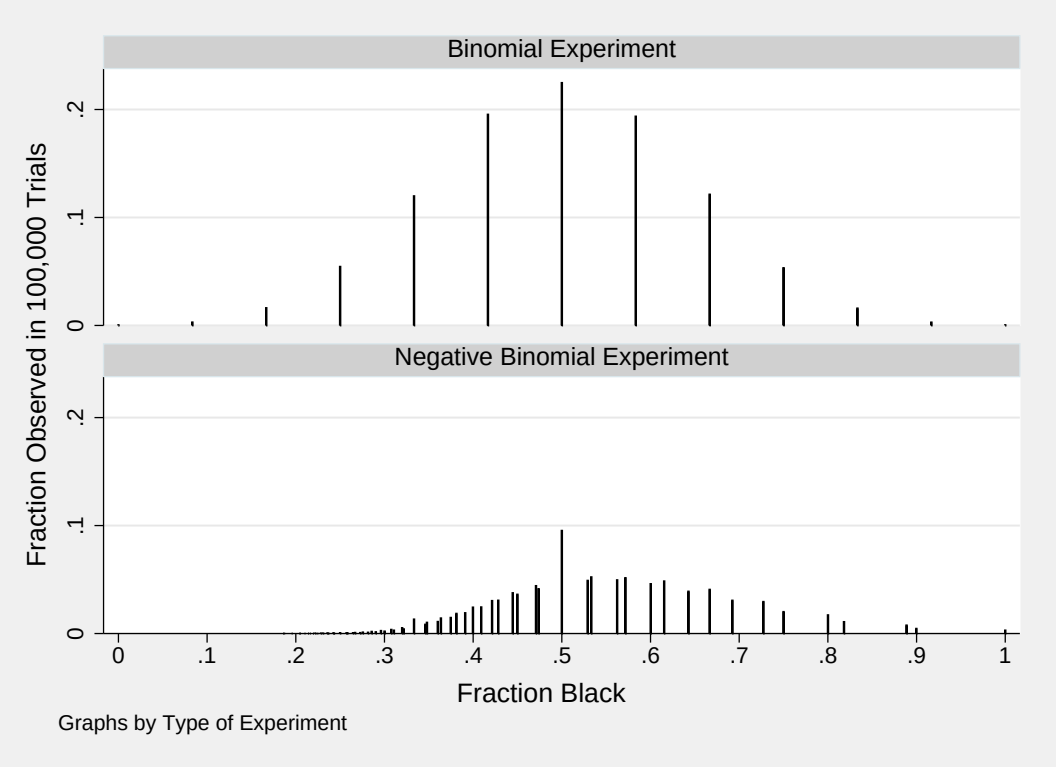

Figure 4: The Introductory Puzzle – Which DGP

1. One point to emphasize is that in experiment A, the sample size is fixed. In experiment B, it was *possible* that the same experimenter would have continued to draw balls from the urn if a third red ball had not been drawn.

2. In neither case is it correct to make a statement such as "given the experimental results (of 9 black and 3 red) there is a 7.3% probability in experiment A (3.3% probability in experiment B) that the null hypothesis is true". The hypothesis is presumably either true or false. The probability statements are statements about one particular "property" of a procedure. Whether it is a "good" procedure depends on a great deal more.

3. For many purposes, neither experiment is particularly "good." It depends on the alternative hypothesis that is the salient rival, but it is easy to come up with cases where Type I and II errors are going to be rather large. Figure 5.2 displays the sampling distribution of the two estimators. Neither experiment is going to be good, for example, at detecting the difference between a true mean of 0.5 and 0.51 for example.

   Indeed, this was the non–Bayesian reaction to our earlier examination of ECMO: these experiments aren't likely to settle a well–meaning debate. Sometimes one is faced with a situation where one is trying to squeeze some inferential blood from an experimental rock. In many such cases, we will not be able to put any proposition to a "severe test."

For the Bayesian the resolution of the problem is quite different – the DGP doesn't (and shouldn't) matter. This if often referred to as "the likelihood principle." To see how this works, recall the statement of Bayes Rule in equation (5)

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_{j=1}^{k} P(B|A_j)P(A_j)}$$

Consider two different likelihoods such that:

$$zP(B|A_i) = P^*(B|A_i) \quad \forall A_i, z > 0$$

Now use Bayes' Rule to show that one's inference is unaffected by use of $P^*(B|A_i)$ instead of $P(B|A_i)$:

$$\begin{aligned} P(A_i|B) &= \frac{P^*(B|A_i)P(A_i)}{\sum_{j=1}^{k} P^*(B|A_j)P(A_j)} \\ &= \frac{zP(B|A_i)P(A_i)}{\sum_{j=1}^{k} zP(B|A_j)P(A_j)} \\ &= \frac{zP(B|A_i)P(A_i)}{z\sum_{j=1}^{k} P(B|A_j)P(A_j)} \\ &= \frac{P(B|A_i)P(A_i)}{\sum_{j=1}^{k} P(B|A_j)P(A_j)} \end{aligned}$$

Indeed, because of this property, Bayes rule is often written as:

$$\underbrace{P(A_i|B)}_{\text{Posterior}} \propto \underbrace{P(B|A_i)}_{\text{Likelihood}} \underbrace{P(A_i)}_{\text{Prior}} \tag{9}$$

Consequently, how the data was generated does not matter for the typical Bayesian analysis. It is also why a Bayesian would view the information in the binomial versus negative binomial experiment as being the "same".

This property of Bayesian inference has been frequently cited as being one of the most significant differences between Bayesians and non–Bayesians.[37]

5.3 If the DGP is Irrelevant Is the Likelihood Really Everything?

A great deal more follows from the Bayesian approach. Unlike the previous example, which might discomfit some non–Bayesians, another implication seems a bit more problematic. One significant difficulty with the simplest versions of Bayesian analysis concerns the distinction between "theorizing after the fact" and "predesignation".

There exist many discussions of this problem. Our discussion follows Sober (2002), who poses a problem involving a deck of cards with 52 different types of cards. Suppose 5 cards are randomly drawn from a typical 52 card deck. Call the configuration of cards that results X. We are now going to use data on X to revise our beliefs about various theories of the world.

Two theories that can "explain" X include:

1. Theory A. The particular 5 cards were randomly drawn from a deck of 52 cards.
2. Theory B. A powerful demon intervened to ensure that the configuration A was drawn.

[37] Poirier (1995) provides a useful parody of the non–Bayesian view as he depicts a statistician's constantly–evolving inference changing as he discovers more about the intent of the investigator.

The essence of Bayesian analysis requires calculating the likelihood of observing X if theory A is true and calculating the likelihood of observing X if theory B is true. Your actual priors aren't particularly important, but assume that $P(A) > 0$ and $P(B) > 0$, although the probability you attach to them can be small.

The problem arises because the likelihood of the second (silly) theory is higher in the second (false) theory than in the first (true) theory. Since there 2,598,960 different 5 card hands that can result:

$$P[X|A] = 1/2,598,960$$

$$P[X|B] = 1$$

Regardless of your prior beliefs about A or B, whatever you believed before, equation (9) instructs you to increase the "weight" you give to the demon hypothesis! (Of course, your posterior density might assign little weight to B, but our interest is merely in the fact that the "experiment" induces you to give more weight than you did before to B). If we continued drawing 5 card hands, and continued to elaborate our demon hypothesis after the fact, we could in principle move you even closer to believing that hypothesis!

If that example strikes you as fanciful, consider a more familiar example, usually called the "optional stopping" problem. To fix ideas, imagine being interested in whether some normally distributed variable (with a known variance of 1) has a mean of zero or otherwise.

1. Take a sample of size 100 and do the usual non–Bayesian hypothesis test in the manner suggested by Kmenta earlier. In this case compute $z = \frac{\sum_{i=1}^{N} X_i}{\frac{1}{\sqrt{N}}}$
2. Continue sampling until $|z| \geq k_{0.05}$ or $N = 1000$, whichever comes first, where $k_{0.05}$ is the appropriate critical value for a five percent confidence interval.

As the non–Bayesian knows, the first procedure provides a far more reliable indicator that the mean is zero than the second test. With the sampling size fixed in advance, if $|z|$ turns out to be greater than the appropriate critical value, the usual conclusion is that either the null is false or "something surprising happened." Under the second DGP, the probability of type I error is 53 percent (see Mayo and Kruse, 2002). What the results of the experiment would do to a non–Bayesian's *beliefs* is a separate matter, but it is clear that s/he would find a "rejection" much more informative in the first case.

By contrast, for the Bayesian who adheres to the Likelihood Principle, both experiments provide the **same** information: the posterior probability assigned to the null hypothesis should be the same, regardless of which experiment is performed. The Bayesian is free to ignore the "intentions" of the experimenter (i.e. the DGP), presumably "locked up" in the mind of the experimenter. Confronted with the evidence consistent with the usual rejection of the null for the non–Bayesian, the change in the posterior beliefs of the Bayesian would be the same under both experiments.

An interesting debate on "optional stopping" can be found in the famous Savage forum (1962, pages 70 forward), where precisely this example is discussed. For Armitage, a non–Bayesian, the DGP is very important and a flaw of Bayesian reasoning:

> I think it is quite clear that likelihood ratios, and therefore posterior probabilities, do not depend on a stopping rule. Professor Savage, Dr Cox and Mr Lindley [participants in the forum] take this necessarily as a point in favour of the use of Bayesian methods. My own feeling goes the other way. I feel that if a man deliberately stopped an investigation when he had departed sufficiently far from his particular hypothesis, then 'Thou shalt be misled if thou dost not know that'. If so, prior probability methods seem to appear in a less attractive light than frequency methods, where one can take into account the method of sampling (Savage et al., 1962, page 72).

G. A. Barnard, another forum participant – who originally proposed that the two experiments should be the same (Barnard, 1947a,b) and introduced the notion to Savage – expressed the view that the appropriate

mode of inference would depend on whether the problem was really a matter of choosing among a finite set of well–defined alternatives (in which case ignoring the DGP was appropriate) or whether the alternatives could not be so clearly spelled out (in which case ignoring the DGP was not appropriate.)[38]

5.4 What Probabilities Aren't – the Non–Bayesian View

In a phrase, a Bayesian is more congenial to the notion that probabilities generated in the course of hypothesis testing represent the "personal probability that some claim is true or not", while such probabilities are merely devices that help "guide inductive behavior by assessing the usefulness of an experiment in revealing an 'error'"[39]. One problem sometimes cited by Bayesians is that non–Bayesians don't understand what "probability" means. To put it succinctly, a "p–value" is *not*:

- "The probability of the null hypothesis.
- The probability that you will make a Type I error if you reject the null hypothesis.
- The probability that the observed data occurred by chance."(Goodman, 2004)

The usual set up begins with a "null hypothesis" and an "alternative hypothesis." Hypotheses can be simple or composite: an example of a simple hypothesis is "the population mean of a binomially distributed random variable is 0.5". That is, we can completely characterize the distribution of the random variable under the hypothesis. A "composite" hypothesis is a hypothesis that does not completely characterize the distribution of the random variable. An example of such a hypothesis is "the population mean of a binomially distributed variable is greater than 0.5". In addition to a set of "maintained hypotheses" ("the experimental apparatus is working correctly") the next step is specifying a *test statistic*. In the usual hypothesis testing procedure, the distribution of this test statistic under the null hypothesis is known.

There are many ways to demonstrate that the probabilities that are used in hypothesis testing do not represent the probability that some hypothesis is true. A distinction that is sometimes made is "before trial" and "after trial" views of power and size. The following example comes from Hacking (1965).

Consider two hypotheses, a null (H0) and an alternative (H1), which are the only two possible states of the world. Let E_1, E_2, E_3, E_4 be the four possible outcomes and let the following be true about the world:

	$P(E_1)$	$P(E_2)$	$P(E_3)$	$P(E_4)$
H0:	0	0.01	0.01	0.98
H1:	0.01	0.01	0.97	0.01

We are interested in two tests, R and S, and specifically the power and size of the tests. Let the size of a test be the probability of incorrectly rejecting the null when it is true, and let the power of the test be 1 less the probability of type II error (not rejecting H0 when it is false). For tests of a given size, more powerful tests are "better". The caveat about "a given size" is necessary since we can always minimize size by deciding on a rule that always rejects.

		Before Trial	
		Size	Power
Test R	Reject H0 if and only if E_3 occurs	0.01	0.97
Test S	Reject H0 if and only if E_1 or E_2 occurs	0.01	0.02

If one takes a naive view of "power" and "size" of tests, the example is problematic. The size of both tests are the same, but test R is much more powerful – much less likely to fail to reject the null when it is false. *Before the trial*, we would surely pick test R.

What about *after* the trial? Consider the case when E_1 occurs. In that case test R instructs us to "accept" the null" when *after* the trial we *know* with complete certainty that the null is false. The standard

[38]Savage, before becoming familiar with the arguments in Barnard (1947a,b), viewed the DGP as relevant ("I then thought it was a scandal that anyone in the profession could advance an idea – [that the DGP was irrelevant] – so patently wrong"). By the time of the forum he had come around to exactly the opposite point of view – "I [can] scarcely believe that people resist the idea [that DGP was irrelevant] that is so patently right."

[39]It should be noted than many Bayesians would argue that hypothesis testing *per se* itself is not a terribly sensible framework. They would also probably argue, nonetheless, that hypothesis tests are best interpreted in a Bayesian way.

"evasion" of the problem for non–Bayesians is to observe (as Hacking (1965) and Mayo (1979) observe), that this is not a test that would usually be countenanced since there exist uniformly more powerful tests than R. This evasion, however, does not get to the heart of the problem.

5.5 What Should "Tests" Do?

The previous discussion has attempted to be clear why the "probabilities" of the usual hypothesis testing procedures should **not** be conflated with the "probability that the hypothesis is true."

What, then, is the "heart of the problem"? One argument, now associated with Mayo (1996), is that hypothesis tests should be used to put propositions to "severe" tests. The purpose of the probabilities for the non–Bayesian is to ascertain, as much as one can, how reliable specific procedures are at detecting errors in one's beliefs.

What is a severe test? In C. S. Peirce's words:

> [After posing a question or theory], the next business in order is to commence deducing from it whatever experimental predictions are extremest and most unlikely ...in order to subject them to the test of experiment. The process of testing it will consist, not in examining the facts, in order to see how well they accord with the hypothesis, but on the contrary in examining such of the probable consequences of the hypothesis as would be capable of direct verification, especially those consequences which would be very unlikely or surprising in case the hypothesis were not true. When the hypothesis has sustained a testing as severe as the present state of our knowledge ... renders imperative, it will be admitted provisionally ... subject of course to reconsideration. (Peirce (1958, 7.182 and 7.231) as cited in Mayo (1996))

Perhaps no better account can be given than Peirce's quotation. A nice quick gloss of a slightly more formal version of this idea is given in Mayo (2003):

Hypothesis H passes a severe test T with $\mathbf{x}$ if:

(i) $\mathbf{x}$ agrees or "fits" H (for a suitable notion of fit).

(ii) with very high probability, test T would have produced a result that fits H less well than $\mathbf{x}$, if H were false or incorrect.

Mayo (1982) gives a nice example of why error probabilities *of themselves* are not enough, and why specification of an "appropriate" test statistic is a key ingredient. Mayo's example involves testing whether the probability of heads is 0.35 (H_0) against the alternative that it is 0.10 (H_1). It is an "artificial" example, but doesn't suffer the defect of the previous example – namely that the test is not the best in its class.

Suppose it is agreed that four coins will be tossed and that the most powerful test of size 0.1935 will be chosen. The following table shows the likelihood of observing various outcomes in advance of the experiment:

# Heads	0	1	2	3	4
$P(H_0\|\cdot)$	0.1785	0.3845	0.3105	0.1115	0.0150
$P(H_1\|\cdot)$	0.6561	0.2916	0.0486	0.0036	0.0001

Consider the following two tests:

Test 1 Reject $H_0 \iff h = 0, 4$ Size = 0.1935 Power = 1 - 0.3438
Test 2 Reject $H_0 \iff h = 0$ Size = 0.1785 Power = 1 - 0.3439

Given the set up, the most powerful test of size 0.1935 is Test 1 – it is (slightly) more powerful than Test 2. But preferring test 1 clearly doesn't make sense: if one sees all heads, it is surely more likely that H_0 is true, yet Test 1 instructs you to reject. Mayo's solution is to observe that Test 1 fails to use an *appropriate* test statistic – one that measures how well the data "fits" the hypothesis. Even though if one is searching for tests of size 0.1935 or better with the most "power", one chooses Test 1 at the cost of a non-sensical test statistic. The *usual* sort of test statistic might be the fraction of heads (F) less 0.35. Such a statistic has the property of punishing the hypothesis in a sensible way.

In this case, the test statistic takes on the following values:

# Heads	$F - 0.35$
0	-0.35
1	-0.10
2	0.15
3	0.40
4	0.65

In this account, Test 2 corresponds to the decision rule "Reject if $F - 0.35 < -0.1$" and the outcomes are now ordered by their departure from the null (in the direction of the alternative). The use of an appropriate sense of "fit" serves to show that the probabilities *per se* are not important – they don't directly correspond to a measure of belief. Rather, they are (one step) in assessing how good the test is at revealing an "error" (Mayo, 2003). The theory doesn't tell you in most non-trivial cases, however, how to generate a sensible test statistic – that depends on context.

While this example is admittedly superficial, it helps explain why, in constructing a good experiment, the importance of other (possibly not well–defined) alternatives cannot be ignored. How *severe* a test is is always relative to some other possible alternatives. Suppose we collect data on unaided eyesight and the use of corrective glasses or contact lenses. If one proposed to "test" the theory that eye glass wearing **caused** unaided eyesight to get worse and found a "significant" rejection of the null of no correlation in favor of the alternative that the correlation was negative the "p value" might be small but it would fail to be a **severe** test against the hypothesis that people with poor uncorrected vision are more likely to wear eye glasses.

5.6 Randomization and Severity

One place where Bayesians and non–Bayesians differ is on the usefulness of randomization. Here, we can only introduce the problem.

Consider a case where the true state of the world can be characterized simply by the following:

$$y = \beta_0 + \beta_1 T + \beta_2 X + \epsilon \tag{10}$$

where, for simplicity, the β are unknown parameters, the X are things that "cause" y and are observed, ϵ are things that "cause" y but are not observed, and $T = \mathbb{1}$(received treatment).

For the non–Bayesian, one of the benefits of randomization is that the X variables available are usually very inadequate. Also ϵ is some convolution of omitted variables and functional form misspecification: it is not generally plausible to make a statement like "ϵ follows the Normal distribution", although statements like that are often found in the literature. Hence, though one could write down a "likelihood", it isn't necessary for the non–Bayesian.

A caricature might make this clear: it is **not** the case that "on the first day, God created y and made it a linear deterministic function of T and X; on the second day, in order to make work for econometricians, God appended a normally distributed error term with mean 0."

Indeed, in a RCT, when the experimenter can intervene and assign T randomly, the "model" the experimenter estimates is often much less complicated:

$$y = \beta_0 + \beta_1 T + \epsilon \tag{11}$$

For purposes of estimation one *could* write down a normal likelihood for this model:

$$y = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{y_i - \beta_0 - \beta_1 T_i}{2\sigma^2}\right) \tag{12}$$

With this likelihood, one could then specify prior beliefs about the fixed parameters β_0 and β_1, stipulate the homoskedasticity (i.e., that the variance of ϵ was a constant for all observations, or model the heteroskedasticity), etc. After seeing the data a Bayesian could update his/her beliefs about the values of these

Table 1: Hypothetical RCT on the Efficacy of ECMO Pretreatment Values of Key Variables (Standard Errors in Parentheses)

Pre-treatment variable	Treatment	Control
Birth weight (grams)	3.26	3.21
	(0.22)	(0.23)
Age (days)	52	54
	(13)	(14)

two parameters. Note that, in this formulation, there appears to be nothing special about the likelihood to distinguish it from any other comparison of means – nothing tells us, for example, that T was assigned randomly.

Nonetheless, writing down the likelihood seems a bit bizarre for the non–Bayesian. For example, if the treatment, T, was the nicotine patch and y was some outcome like "quit smoking successfully", no one thinks that only the patch matters and nothing else that can be observed matters – clearly the price of cigarettes, social norms, etc. play a role. Indeed, available covariates are usually not used except to "test" the validity of the design. Specifically, in repeated samples:

$$E[\overline{y}_1] - E[\overline{y}_0] = \beta_1 \tag{13}$$
$$E[\overline{X}_1] - E[\overline{X}_0] = 0 \tag{14}$$
$$E[\epsilon_1] - E[\epsilon_0] = 0 \tag{15}$$

where $\overline{y}_1 \equiv \frac{1}{N_1}\sum_{i;T=1} y_i$, $\overline{y}_0 \equiv \frac{1}{N_0}\sum_{i;T=0} y_i$, the subscript 1 refers to the treatment group, the subscript 0 refers to the control group, and so on.

Taken literally, (14) suggests that, on average, in repeated samples, the mean for *any* pre–treatment variable should be the same. An auxiliary implication is that if one ran the regression but also included pre–treatment variables, the estimate of the effect of the treatment should not change. If it does change substantially, this is evidence against the design, and a cause for concern.

To fix ideas, suppose the treatment under consideration is ECMO (which we considered in section 1.1) and suppose a standard randomization scheme was employed on a large sample of children. A standard procedure is to report the averages for several variables. Table 1 is a hypothetical table.

Usually researchers report whether there are any "significant" differences between the treatment and control group means. The intended purpose is to ensure that the two groups satisfy a *ceteris paribus* condition: in ways we can observe, are the two groups roughly the same – this is sometimes referred to as "balance." If sample sizes are large enough, more frequently than not the values in the two columns will not be "significantly" different. It serves as a "check" that the randomization achieved its intended purpose.

What variables should be included in this "check?" Presumably such a list does not include hair color, although this, in principle, should be balanced as well. The usual rule is to consider "pretreatment variables which are predictive of the outcome." These may or may not be part of a proper "model" of infant death, but are there to assure one's self, that if there is a large difference in the groups after treatment that the researcher will not mistakenly attribute to the treatment what was really a failure of the *ceteris paribus* condition.

Bayesians frequently point to a flaw in this argument:

> My doubts were first crystallized in the summer of 1952 by Sir Ronald Fisher. 'What would you do,' I had asked, 'if, drawing a Latin Square at random for an experiment, you happened to draw a Knut Vik square?' Sir Ronald said he thought we would draw again and that, ideally, a theory explicitly excluding regular squares should be developed ...
>
> The possibility of accidentally drawing a Knut Vik square or accidentally putting just the junior rabbits into the control group and the senior ones into the experimental group illustrates a

Table 2: Bad Luck in a Hypothetical RCT on the Efficacy of ECMO Pretreatment Values of Key Variables (Standard Errors in Parentheses)

Pre-treatment variable	Treatment	Control
Birth weight (grams)	3.26	2.1
	(0.22)	(0.23)
Age (days)	52	140
	(13)	(14)

> flaw in the usual . . . argument that sees randomization as injecting 'objective' or gambling–device probabilities into the problem of inference (Savage et al., 1962, page 88).

Savage's example of having all the young rabbits in the control group and all the older rabbits in the treatment group is perhaps more recognizable than the distinction between Latin Squares and Knut Vik squares, which come from classical agricultural experimentation. It is also related to the problem of random or pseudo random number generation. If generating a sequence of random 0s and 1s, for instance, by chance (if only infrequently), some of these sequences will be undesirable – for example, if drawing a sequence of 1000 numbers, it is possible that one draws all zeroes or all ones.

In the context of the RCT, the analogous problem, loosely based on our ECMO example, is described in Table 2. Assignment to the two groups was randomized but "bad luck" happened and the control group was comprised of the lightest birth weight babies (and viewed by the doctors as usually the least healthy) who were, on average, potentially in need of ECMO at much older ages (again, viewed by the doctors as an indicator of general frailty).

In this example, the problem is that the treatment and control groups are not "balanced". The treatment babies are (before treatment) healthier on average than the control babies. The typical non–Bayesian would generally find the numbers in Table 2 evidence against the validity of the design. [40]

For Bayesians, this suggests that the logic of randomization is flawed. If "balance" is the primary reason for randomization, why not deliberately divide into groups which look similar (and would "pass" a balancing test) without randomization *per se*. How does introducing uncertainty into treatment assignment help? Indeed, to some Bayesians, all it can do is lower the value of the experiment. From Berry and Kadane (1997): "Suppose a decision maker has two decisions available, d_1 and d_2. These two decisions have current (perhaps posterior to certain data collection) expected utilities $U(d_1)$ and $U(d_2)$ respectively. Then a randomized decision, taking d_1 with probability λ and d_2 with probability $1-\lambda$ would have expected utility $\lambda U(d_1) + (1-\lambda)U(d_2)$ If the randomization is non-trivial, i.e. if $0 < \lambda < 1$, then randomization could be optimal only when $U(d_1) = U(d_2)$, and even then a non-randomized decision would be as good."

Savage goes further: "It has been puzzling to understand, why, if random choices can be advantageous in *setting up* an experiment, they cannot also be advantageous in its analysis."

There is much more to say, for instance, it may be useful to think about this class of problems in terms of the severity concept that we introduced earlier. However, it may be more instructive to consider two examples from real research. In one, I identify the problem as lack of severe testing. In the second, I identify the problem that the world is a "complicated place": assertions that were felt to be well–grounded by numerous studies seem less so in the face of a well designed experiment.

[40] There are many solutions to the problem of "inadmissible" samples in practice (unbalanced samples, see Jones (1958) for example.) One could merely conduct two experiments with more homogeneous samples. That is, one could conduct an experiment on low birth weight babies and a separate experiment on high birth weight babies. Sometimes block randomization is employed: the children might be subdivided into groups according to their "healthiness" and the randomization might be performed separately within blocks.

6 Case Study 1: "Medication Overuse Headache"

In Section 1.1 I briefly mentioned the case of ECMO – a treatment for infants with persistent pulmonary hypertension whose success was initially uncertain, but retrospectively seems of great benefit. Here I would like to consider a potentially "mirror–image" case: a treatment is being administered that, in my view, is potentially quite harmful. Also, I would argue, the literature is of unbelievably low quality. I locate the problem with the theory in the fact that, instead of behaving like Mayo's "error statistician" or engaging in "Peircean severe testing", the researchers began with a prior belief and then set about "updating" it. It should be noted that none of the studies involving this topic used "Bayesian statistics".[41] Rather, the question is "is there enough evidence to proceed with the expert consensus" or is more "severe" testing necessary?

This case is particularly useful because, as with many problems in medicine and social sciences (and elsewhere), it involves a problem of dubious ontology (is there "really" such a thing as MOH?) as well as the problem of "new hypotheses" that "accommodate" the evidence instead of having a theory held in advance that "predicted" the evidence (much like our "demon" example in section 5.3).

A road map for what follows is:

1. During a period of time when the field was considered a "backwater" a diagnosis of MOH was developed. This theory argued that people with chronic severe headache pain caused their pain by taking pain medication "too frequently" and that, if they merely stopped taking the pain medication, their pain condition would improve.

2. The evidence for this theory was that patients who agreed to stop their pain medication had higher rates of improvement than those who didn't. These studies typically ignore serious selection bias due to non–random attrition and regression to the mean.

3. In one of the few published critiques of the theory, it was noted that millions of users of analgesics for reasons other than headache do not develop migraine. In response, the theory evolved to state that only those individuals "predisposed" to migraine get MOH.

4. When a definition of the diagnosis that required improvement in pain after cessation of the offending medications was proposed it was strongly criticized. The definition was revised to empty it of potentially refutable content.

Every challenge to the theory that pain medication causes pain has been met by "accommodating" the evidence. Rather than reject the theory, at every turn the theory has accommodated the new evidence by making it more difficult to test. Furthermore, there is a complete absence of "severe testing."

6.1 What *is* Medication Overuse Headache? Nosology and Dubious Ontology

The essence of "medication overuse headache" as a term for a certain class of chronic headache pain[42] is the idea that the patient *causes* their pain by taking pain (or other headache) medication in excess of

[41] See Zed et al. (1999) and Headache Classification Committee of the International Headache Society (2004) for extensive bibliographies.

[42] The nosology of headache is elaborate and I can only coarsely define two types of headache here. According to the National Headache Foundation (`http://www.headaches.org/consumer/tension_type.html`, accessed 10 December 2007), "Tension-type headache is a nonspecific headache, which is not vascular or migrainous, and is not related to organic disease. The most common form of headache, it may be related to muscle tightening in the back of the neck and/or scalp [and is] characterized as dull, aching and non-pulsating pain [that] affect[s] both sides of the head. Symptoms . . . may include:

- Muscles between head and neck contract
- A tightening band-like sensation around the neck and/or head which is a "vice-like" ache
- Pain primarily occurs in the forehead, temples or the back on head and/or neck"

Migraine headaches are most commonly associated with severe unilateral head pain often accompanied by nausea and vomiting, photophobia (fear of light) and phonophobia (fear of sound) that can last from a few hours to several days. In some fraction of migraine patients the head pain is preceded or accompanied by visual disturbances called auras.

arbitrary norms (set by researchers in the area) of appropriate use. The "offending" medication, as it is often referred to, can be any of a very diverse set with very different effects and mechanisms of action. These include ergotamine, caffeine, morphine, sumatriptan, and many other drugs. Opioids (morphine and related medications) are generally thought to be more of a problem than the other medications.(Saper and Lake, 2006a)[43] Obermann and Katsarava (2007) cite a global prevalence rate of 1% and describe it as the "third most frequent headache type after tension–type headaches and migraine."[44]

> There are two ways to account for this phenomenon. The obvious one is that these people take chronic daily analgesics because they have chronic daily headaches. This is the explanation embraced by our patients and, until recently, by most physicians [who are not headache specialists](Edmeads, 1990). [45]

I believe this case study is illuminating because it suggests that the problem is *not* one of failing to view probability as epistemic, but is because researchers in the area have systematically *not* confronted their long–held views with severe testing.

6.2 Some Salient Background

6.2.1 Early History

In a recent review of the subject, Obermann and Katsarava (2007) date the first clear identification of MOH to a 1951 study, without a control group or randomization, which described 52 patients who took daily amounts of ergotamine and improved after "the ergotamine was withdrawn. A recommendation of the first withdrawal program followed and was introduced in 1963." The view that medication overuse was a *cause* of migraine pain "became well–established" in the early 1980s (Capobianco et al., 2001). It was first officially defined by the International Headache Society (1988) – the international association of neurologists with a specialty in headache – as "drug induced headache" in the International Classification of Headache Disorders (ICHD-1)(Obermann and Katsarava, 2007).

The view that medication use *caused* head pain was developed during a period of time when it was widely held that "migraine was a disorder of neurotic women" (Silberstein, 2004).[46]

6.2.2 The Evidence

While a complete review of the evidence is not possible, let me take one representative example: Mathew et al. (1990).[47] Figure 6.2.2 is a modified version of a table from Mathew et al. (1990). The title of the table

[43] In the U.S. opioids were the standard of care as late as the 19th century until it was supplanted by aspirin in the early 20th century at roughly the same time over the counter use of such medications was made illegal.(Meldrum, 2003). Moreover, outside of the MOH literature it is generally viewed that opioids are under prescribed because of (sometimes irrational) fear of promoting addiction, censure by police, etc.Lipman (2004). "The history of opioid use (or nonuse) in neuropathic pain is instructive. The natural reluctance to prescribe opioids to patients with neuropathic pain of benign cause was, for many years, reinforced by the received wisdom that opioids were ineffective in neuropathic pain [such as headache], based on weak evidence. It took many years before this 'truth' was questioned. Reexamination in the later 1980s was followed by controlled studies that clearly substantiated an important analgesic action of morphine and fentanyl and, later, other opioids in neuropathic pain."(Scadding, 2004).

[44] Medication overuse headache has gone by several different names including *analgesic rebound*, *ergotamine rebound*, *medication induced headache*, *transformed migraine chronic migraine*, *daily headache*, *drug-induced headache*, *painkiller headache*, *medication–misuse headache*, *analgesic-dependent headache* (Obermann and Katsarava, 2007, Silberstein et al., 1994), etc.

[45] As is widely recognized, severe chronic daily migraine occurred before the use of offending medications was common or even possible. The most widely cited example comes from the important neurologist Thomas Willis (1675) who recorded his treatment of Viscountess Anne (Finche) Conway in the 17th century.

[46] Even current scholars in the field are negative about early developments in the field of headache. "Prior to the 1980s, the field of headache was rarely influenced by what would be generally accepted as scholarly, credible research." (Saper, 2005)

[47] I describing the work as "representative" however, the view among experts in the field is considerably more favorable. Ward (2008) describes it as "his favorite article" in a recent review. Mathew (2008) responds by noting "The impact of this article on the American and European headache communities was substantial. Until then, the Europeans had not appeared to appreciate the clinical significance of medication overuse or the existence of chronic daily headache ... [and that] One enduring

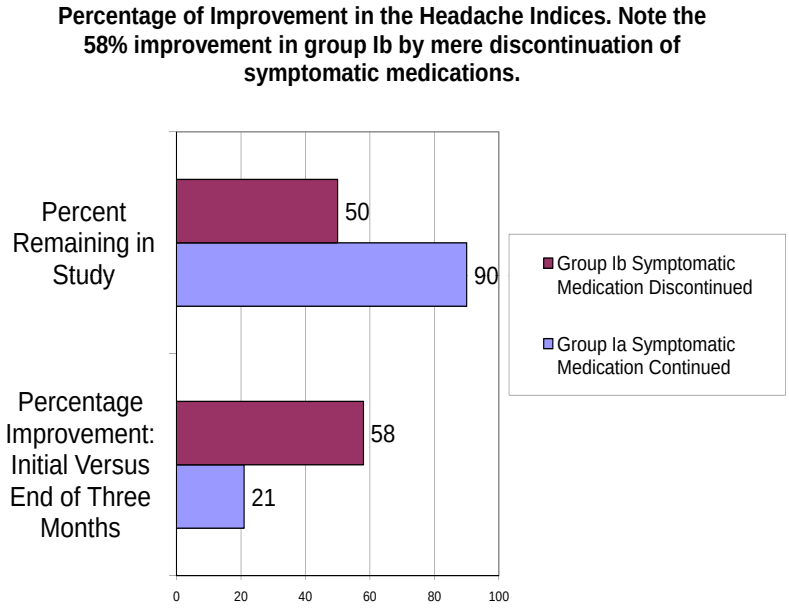

Figure 5: The Evidence In a Nutshell

is also from the original article. Patients were assigned[48] to different treatment groups and their progress was observed.[49]

As Mathew et al. (1990) report, the data in the figure were based only on those patients who remained in the study – 90% in the group which continued to receive medication and only 50% in the group which had the medication withdrawn. With slight variations Mathew et. al's conclusions have become standard.[50] As far as I have been able to determine, the literature has not concerned itself with the problem of non–random attrition. One might expect that a patient who fails to improve after stopping the offending medication might be more likely to drop out than one who has improved (for some reason possibly unrelated to treatment). One possible reason for improvement might be mere "regression to the mean". For evidence that strongly suggests this is a problem see Whitney and Von Korff (1992).

fact continues to disappoint me. In spite of the extensive effort made to emphasize the importance of medication overuse in managing the headache population, many practitioners – including neurologists – continue to overprescribe symptomatic medications, thereby condemning their patients to treatment failure." For a more recent systematic review, see Zed et al. (1999).

[48]No randomization appears to be involved. The patients were merely "grouped."

[49]The measurement of "improvement" is not clear, but it appears to have been asymmetric. Improvement was measured as a percentage change in a headache index if the patient improved, and was given a value of zero if the patient did not improve.

[50]One definition of Medication Overuse Headache (MOH) is:

1. Occurs in a patient with a primary headache disorder who uses symptomatic or immediate relief medications very frequently (daily), often in excessive quantities.
2. Tolerance to symptomatic medications develop and headaches become worse on continuing the treatment.
3. The patient may show symptoms of withdrawal on discontinuing the medication, with increased headache lasting for a variable period of time, as long as 3-4 weeks.
4. Headache ultimately improves after stopping the offending medications even though the primary headache disorder needs continuing prophylactic treatment.

6.2.3 First Criticism

In one of the first (and extremely rare[51] criticisms of research in this area) Fisher (1988) observed:

> As I understand it, analgesics beget more headache by making more brain serotonin available, which paradoxically increases the pain. My question is whether it holds only for headache or for other pains as well? In our arthritic clinic aspirin was used in doses of 8 to 14 tablets a day for about 15 years. I asked physicians who attended that clinic in those years whether they had ever noticed [the development of headache pain] ... in any of the patients and they never had. Also 3 to 5 million people in the United States are taking aspirin daily to prevent arterial thrombosis. Should we expect headache ... under these circumstances?[52]

As Fisher understood, the answer to his questions was "no" and advocated a randomized trial on the effect of withdrawal from headache medications where the control group would be subject to sham double–blind withdrawal. One can think of this as a proposal for a "severe test."

The response in the literature was to maintain that the theory was, in the main, correct and to merely amend the theory to accommodate the troubling fact highlighted by Fisher.[53] A typical amendment stated that "'analgesic abuse headache' may be restricted to those patients who are already headache sufferers" [and that] individuals with ... migraine, are predisposed to developing chronic daily headache in association with regular use of analgesic"(Bahra et al., 2003).

6.3 Redefining MOH to Avoid a Severe Test

The process of defining MOH provides a clear example of researchers avoiding a severe test. A useful place to start is the *initial* International Classification of Headache Disorders–2 (ICHD-2).[54] The initial ICHD-2 definition of "medication overuse headache" in displayed in Table 3 (Headache Classification Committee of the International Headache Society, 2006).[55]

A key aspect of the definition is criterion C: the patient's *decision* to continue using analgesics at more than the approved rate[56] This was immediately recognized to be a problem: existing standards of treatment for other forms of migraine, such as "menstrual migraine" *required* the use of analgesics at a rate which could then (inappropriately) be described as MOH.[57]

Another important aspect of the definition of MOH that was the subject of great dispute was item D – the requirement that after removing the patient from the offending medication the patient would improve. As reported in the literature, the problem was that such a requirement vitiated using MOH as a "diagnosis" in the traditional sense:

> "the problem is that medication overuse headache cannot be diagnosed until the overuse has been discontinued and the patient has been shown to improve. This means that when patients

[51]The only other criticism I have been able to locate are letters to the editors (Gupta 2004a,b,c).

[52]Using language evocative of severe testing Fisher remarked that "Claude Bernard, in speaking of an hypothesis, said that it is not sufficient to merely gather all the facts that support it but even more importantly, one must go out of one's way to find every means of refuting it."

[53] While Fisher's challenge has never been even approximately met, the question of whether headaches arise *de novo* from analgesic headache has been investigated and substantiates Fisher's claim. It has been conceded that Fisher was correct. (Bahra et al., 2003, Lance et al., 1988)

[54]For a brief history of the ICHD, see Gladstone and Dodick (2004).

[55]Boes and Capobianco (2005) and Ferrari et al. (2007) describe some of the tangled history as well. It should be noted that the history is a matter of some dispute.

[56]The rate of analgesic use is *usually* defined in terms of treatment days per month, such that treatment occurs at least two or three days each week, with intake of the drug on at least 10 days per month for at least three months.

[57]From Schuster (2004): "The definition does not apply to headache in women who take medications for five or six consecutive days for menstrually associated migraines but are treatment-free the rest of the month, acknowledged Fred D. Sheftell, MD, who participated in updating the classification. He said that if a woman took medications just four other days of the month, she would inappropriately meet the 10-days-a-month rule. For that reason, she must also be taking the drugs at least two to three days each week to meet the criteria."

Table 3: *Initial* 2004 International Headache Society classification criteria for analgesic-overuse headache

A	Headache present on $\geq$ 15 days/month with at least one of the following characteristics and fulfilling criteria C and D:
	1 bilateral
	2 pressing/tightening (non-pulsating) quality
	3 mild or moderate intensity
B	Intake of simple analgesics on $\geq$ 15 days/month for $>$ 3 months
C	Headache has developed or markedly worsened during analgesic overuse
D	Headache resolves or reverts to its previous pattern within 2 months after discontinuation of analgesics

have it, it cannot be diagnosed. It can be diagnosed only after the patient does not have it any more." (Olesen et al., 2006)

After a meeting of experts in Copenhagen, this offending section – requiring improvement after going off the "causal" medications – was quickly removed. Whatever their intent, however, this redefinition seemed to make a MOH diagnosis impossible to refute.[58] Indeed, it was immediately noted that "the revision [to the definition of MOH] has eliminated the need to prove that the disorder is caused by drugs, that is, the headache improves after cessation of medication overuse."(Ferrari et al., 2007) Although they suggested that "probable MOH" be introduced their main focus was that sub-forms of MOH be defined for different types of medications, with opioids singled out as particularly problematic.[59]

The case of opioids is especially interesting since it is generally believed that opioid–related MOH is more worrisome and it has been argued that "sustained opioid therapy should rarely be administered to headache patients."(Saper and Lake, 2006b) This case is also useful since it might be falsely assumed that individuals doing research in this area (and supporting the idea of MOH) are incapable or not disposed to putting hypotheses to severe testing. As noted previously, researchers in this area routinely make no adjustment of any sort for the high rates of attrition in studies looking at chronic headache pain.

An illustrative exception to non–severe testing involves, not a test of MOH, but rather a study of the efficacy of sustained opioid therapy – opioids being considered a particularly pernicious cause of MOH.(Saper and Lake 2006a,b) Although Saper et al. (2004) had no control group, the researchers treated individuals who dropped out for *any* reason, died for non–opioid related reasons, were suspected of "cheating" (using more opioids than allowed by the doctors), etc. as *treatment failures*. This also included some patients who reported a substantial improvement but were considered to have "failed" to satisfy the *researchers'* definition of significant reduction in functional impairment. In defining treatment failure more broadly, the researchers were essentially using a "worst case" bound.[60] While the use of "worst case bounds" is infrequent (or non–existent) in the MOH literature, the argument for doing so has validity: it is entirely consistent with the notion of "severe testing." Indeed, leading researchers in MOH are aware of the potential value of such bounds. Saper and Lake (2006b), for example, harshly criticize a meta–analysis on RCTs on the efficacy of

[58] Among the reasons given was the fact that "patients could become chronic due to medication overuse, but this effect might be permanent. In other words, it may not be reversible after discontinuation of medication overuse. Finally, a system whereby medication overuse headache became a default diagnosis in all patients with medication overuse would encourage doctors all over the world to do the right thing, namely, to take patients off medication overuse as the first step in a treatment plan."

[59] See, for example Saper and Lake (2006a), for a proposal to distinguish opioid using MOH patients from the remaining "less complicated' cases.

[60] See Horowitz and Manski (1998) for a detailed discussion of such bounds. It should be noted that where such bounds are used, common practice is to report both "best case" *and* "worst case" bounds.

Table 4: *Revised* ICHD-II criteria for Medication-Overuse Headache

A. Headache present on ≥ 15 days per month

B. Regular overuse for $>$ 3 months of one or more acute/symptomatic treatment drugs as defined under sub forms of 8.2

   1. Ergotamine, triptans, opioids, or combination analgesic medication on $\geq$ 10 days/month on a regular basis for ≥ 3 months
   2. Simple analgesics or any combination of ergotamine, triptans, analgesic, opioids on ≥ 15 days/month on a regular basis for ≥ 3 months without overuse of any single class alone

C. Headache has developed or markedly worsened during medication overuse

opioids for non–cancer pain for failure to adhere to "intent–to–treat" principles. In this instance, this meant treating as failures those individuals who began opioid treatment but then stopped for *any* reason.[61]

The severity of the tests to which opioid efficacy has been confronted is in sharp contrast to extant studies of MOH (sometimes by the same researchers) where a failure of a patient to reduce his medications is not treated as a failure of MOH–therapy. Indeed, where attrition rates of 40% or higher are common, were the literature to treat those who were unwilling or unable to abstain from the offending medication as failures of "MOH–therapy", it would appear that few, if any, of the studies in Zed et al. (1999) for example that purport to provide evidence favorable to the existence of MOH would continue to do so. Indeed, although plagued by non–random attrition and written by advocates of MOH, it has been observed that patients with MOH who "lapse" and reestablish medication overuse have higher measured "quality of life" on average than those with MOH who do not lapse(Pini et al., 2001).

It might fairly be argued that an intelligent Bayesian might not have moved his posterior much in light of the foregoing. Moreover, it is certainly the case that no formal Bayesian analysis has been employed. At least superficially, the "usual" statistical analysis was employed. What this literature *doesn't* do, however, is:

1. Test the theory in such a way that the observed result would be unlikely if the obvious alternative (the one "favored by patients") was true – that it is chronic pain that causes use of pain relieving medication, not the other way around.

2. Employ the "usual" techniques to make tests more "severe" – the failure to use worst case bounds, for instance, to deal with the problem of non–random attrition.

3. React to each threat to the theory as potential reason to abandon the theory. Instead, the reaction of researchers has been continued modification of the theory until it is no longer capable of being refuted by evidence.

[61] Indeed, the notion of "intent–to–treat" can be seen as part of an attempt to test a hypothesis **severely** and not a notion that is an inevitable consequence of adopting "frequentist" probability notions. See Hollis and Campbell (1999) for a discussion.

7 Case Study 2: "Union Wage Premium"

I would now like to consider an econometrically sophisticated literature – the literature on union wage effects. Two comprehensive and influential surveys are Lewis (1963, 1986). In these works, Lewis literally cites hundreds of studies attempting to estimate the causal effect of union status on wages.[62] See also the useful discussion in Heckman (1990).

7.1 Early History

The idea that labor unions might raise wages is one of the oldest debates in economics and was one of the earliest motivating examples for the famous "supply and demand" cross in a study by Jenkin (1870) (see Humphrey (1992) for a short history). Ironically, although Jenkin (1868) concluded that the supply and demand analysis wasn't particularly relevant for explaining the consequences of union wage setting, subsequent neo-classical theorizing in the main focused on the simple model depicted in Figure 7.1, where W is the real wage, L is the quantity of labor, $\mathcal{D}$ is the employer demand curve, S_c is the supply curve without unionization and S_u is the supply curve with unionization. Until Lewis' influential survey, opinions diverged between those who believed that unions could rarely control the supply of labor, such as Milton Friedman, and those that thought they could and therefore acted to create unemployment, such as Paul Samuelson (see Friedman (1950).)

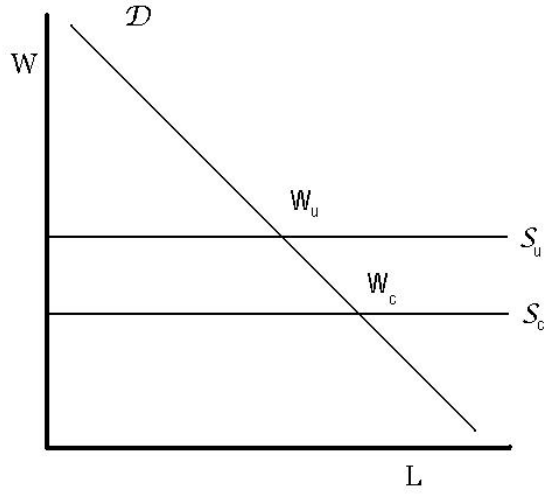

Figure 6: Union Wage Effects in the Neo-Classical Model

7.2 A Battery of Severe Tests

The analysis of union wage effects has become more sophisticated with the advent of large micro data sets, but let me highlight some of the comparisons researchers have employed to analyze the question:

- Ashenfelter (1978) constructs control groups based on industry, race, and worker type (i.e.craftsmen, operatives, laborers)
- Freeman (1984) compares wage rates for the same individual at *different points in time.* At one point in time the worker is in a unionized job at a different point in time the worker is in a non–unionized job.

[62] He referred to the union premium in wages for between otherwise identical workers as the wage "gap" to distinguish it from what might obtain in a world without unions.

- Lemieux (1998) compares wage rates for the same individual who holds *two jobs*, one of which is unionized, the other which is not.
- Krashinsky (2004) compares wage rates of *identical twins*, one who is unionized and one who is not.
- Card (1992) constructs control groups, based on observable characteristics, which tend to receive the same wage in the non–union sector as well as controlling for differences in permanent characteristics (i.e. person–specific fixed effects).
- DiNardo and Lemieux (1997) and Card et al. (2003) compare US and Canadian workers, exploiting differential timing in the decline of unionization in the two countries.

Depending on the precise context, the union wage effect as measured in these studies ranges from positive 5-45%, with the vast majority of studies being at the higher end. All of the aforementioned studies adopt a distinctly non–Bayesian approach to the econometric analysis. The variety of research designs was not motivated by an attempt to "refine" posterior beliefs, but to put the hypothesis that unions raise wages to the most severe test possible with existing data. Each of the papers described above tried to "rule out" other explanations for the difference in union and non–union wages. (Perhaps this is why the posterior distribution of estimates of the union wage effect are as tight as they are–a survey of labor economists found remarkable unanimity on the average size of the effect (Fuchs et al., 1998). The posterior mode of the economists surveyed was that unions raised wages 15% relative to similar non–union workers.)

What is also useful about this example is that there exists at least one Bayesian analysis – Chib and Hamilton (2002), which helpfully contrasts some unsophisticated non–Bayesian estimates from a small sample of workers. These non–Bayesian estimates vary from about 16% to 25%. If one treats these estimates as "average treatment effects for the treated (ATOT)", these estimates are similar to their Bayesian posterior distributions.[63]

To put it a bit too simply, the basic empirical model has long been some variant of the following:

$$\begin{aligned} \log w_i &= X\beta_0 + \alpha_i + \epsilon_0 \quad \text{if } U_i = 0 \\ &= X\beta_1 + \psi\alpha_i + \epsilon_1 \quad \text{if } U_i = 1 \\ P(U_i = 1) &= F(Z\gamma) \end{aligned}$$

where U_i is the union status indicator for worker i, w_i is the wage of worker i, β and γ are parameters –the first two differ depending on whether the worker is unionized or not – X and Z are observed covariates, the ϵ are unobserved terms, and ψ is the ratio of the return to unobserved time invariant individual specific characteristics in the union to the non–union sector. The function $F(\cdot)$ is some type of cumulative density function and the elements of Z may overlap with X.

The Bayesian analysis of Chib and Hamilton (2002) takes a variant of the above model and is focused on how the effect of unions on wages varies across *individuals.* As has long been recognized, however, to a great extent unionization in the U.S. occurs at the *establishment* level (this is in contrast to unionization in Europe which frequently adheres to most workers in an industry). As Krashinsky (2004) observes, this has meant that the aforementioned empirical work has been unable to rule out the possibility of a "firm or enterprise specific fixed effect": a worker's union status could merely be a marker, for instance, for the profitability

[63]The distinction between ATOT and other estimands is important since it isn't particularly meaningful to consider the effect of union status on say, the CEO of a large multinational, to take a stark example. (U.S. law, for example, prohibits this possibility.) The paper, unfortunately takes a naive approach to characterizing the treatment heterogeneity: considering the variation in the effect of union status, characterizing it by the estimated probability of being unionized. That is, the effect of unionization is allowed to vary across workers whose demographic characteristics put them at the same "risk" of being unionized. This conflates the treatment effect for workers with extremely low levels of observed human capital (who typically have very low probabilities of being unionized) with the treatment effects for those who can't be unionized (bosses) or those with high levels of education who are generally hostile to unionization. For an arguably much more sensible characterization of the heterogeneity in treatment effects see Card (1992) for example. Card (1992) also deals with the problem of measurement error in union status which is ignored in the empirical example in Chib and Hamilton (2002) but has long been an important issue in non–Bayesian analyses, see Freeman (1984), Jakubson (1991), or Card (1996) for three examples.

or generosity of the employer.[64] Put in other words, the list of "ceteris paribus" conditions considered in previous research did not include "working at the same firm".

As Freeman and Kleiner (1990) observe about the estimates of union wage effects with individual data, the "treatment effect" of most interest comes from an experiment on "firms" and not on "individuals" *per se*

> While it is common to think of selectivity bias in estimating the union wage effect in terms of the difference between the union premium conditional on the observed union (and nonunion) sample and the differential that would result from random organization of a set of workers or establishments, we do not believe that this is the most useful way to express the problem. What is relevant is not what unionization would do to a randomly chosen establishment but rather what it would do to establishments with a reasonable chance of being unionized-to firms close to the margin of being organized rather than to the average nonunion establishment.

DiNardo and Lee (2004) use a regression discontinuity design, which, in their context provides a very good approximation to a randomized controlled trial (RCT) of the sort discussed in the quote from Freeman and Kleiner (1990). Like previous work in this area, one of the motivating ideas was to put the hypothesis "do unions raise wages" to a more **severe** test, one that would allow for, among other things, a firm–specific effect.

This was possible in this research design since it used data on "firms." The research design essentially focused on comparing firms where the union "barely won" to those who "barely lost".

We can only be brief here, but the experiment is a "regression discontinuity design" based on an aspect of (US) labor law. Workers most often become unionized as the result of a highly regulated secret ballot. If more than 50% of the workers vote for the union, the workers win collective bargaining rights. If 50% or fewer do so, the workers do not win the right to collective bargaining. By comparing outcomes for employers at *firms* where unions barely won the election (e.g., by one vote) with those where the unions barely lost, one comes close to the idealized RCT. The test is severe against the hypothesis that unobserved differences in the *firms* that are unionized versus those that are not unionized can explain the different wages, etc., of unionized workers.

It is rather easy to display the data from regression discontinuity designs. Figure 7 plots an idealized version of the key displays: in each, the average value of some outcome where these averages are computed for different values of the vote share. The figure on the left corresponds to the case where unionization has an effect on the outcome in question. In the figure on the right (and one that resembles the figures in DiNardo and Lee (2004)), there is no detectable effect of unionization.

Figure 7.2 plots an idealized version of the key displays that correspond to ensuring the validity of the research design or "balance": in each the average value of some pretreatment outcome (in the study by DiNardo and Lee (2004) this included firm size and measures of the health of the firm) are plotted for different values of the vote share. The graph on the left corresponds to the good case: firms in establishments where the union barely lost the election look the same as those where the union barely won. This corresponds to what was found in DiNardo and Lee (2004). The figure on the right corresponds to a situation which is evidence against the design: firms in establishments where the union barely lost look much different than firms where the union barely won. In this case, the *ceteris paribus* conditions would seem to be violated.

To summarize the results of the study, the authors find (perhaps surprisingly given the huge literature documenting significant union wage effects) **no** effect of unionization on the myriad of outcomes they examine such as wages, enterprise solvency, productivity, etc. Limitations of scope prevent elaborating in more detail, but the study by DiNardo and Lee (2004) points to an important problem with any "non–severe" test of a hypothesis – Bayesian or non–Bayesian. If one takes the results from DiNardo and Lee (2004) seriously, it is hard to see how the problem could even be *addressed* with individual data – irrespective of whether Bayesian or non–Bayesian statistics were employed.

[64] This possibility wasn't ignored, however. The problem was the lack of data. See for example Abowd and Farber (1982), Freeman and Kleiner (1999), and Freeman and Kleiner (1990) for example, which take the possibility quite seriously.

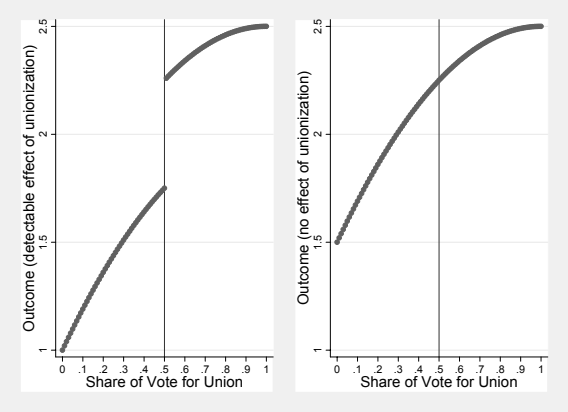

Figure 7: Two Types of Findings in an Regression Discontinuity Design

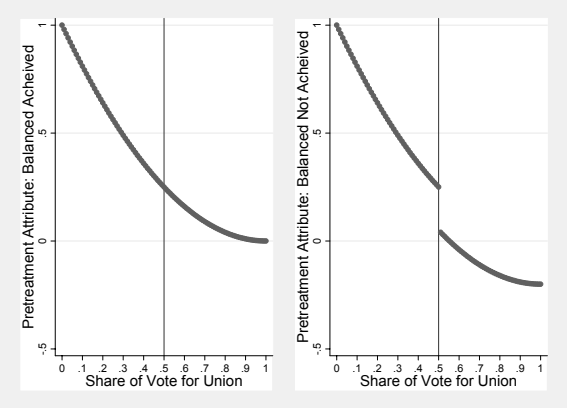

Figure 8: Evidence for or against "balance" in an Regression Discontinuity Design

8 Concluding Remarks

What I have written has only scratched the surface of longstanding disagreements. For any suggestion of dissent with any "Bayesian" views discussed in this essay, there exists volumes of counterarguments. Likewise, the debates among those who do not employ Bayesian methods are no less voluminous. I, myself, don't have a single "theory of inference" to which I adhere.

I have sought, for reasons of clarity, to highlight the *differences* between Bayesian and non–Bayesian perspectives, I risk overstating them. It seems fitting therefore, to conclude by illustrating that one can often find"non–Bayesian features" in Bayesian work and "Bayesian features" in non–Bayesian work.

8.1 Bayesian Doesn't Have to Mean "Not Severe"

The idea that only non–Bayesians look for "severe tests", or "try to learn from errors" is not correct. One nice example comes from a recent careful study by Kline and Tobias (2008), which is interested in estimating the the "effect" of Body Mass Index (BMI) on earnings.

Their point of departure is a two equation model:

$$y_i = f(s_i) + x_i\beta + \epsilon_i \tag{16}$$
$$s_i = z_i\theta + u_i \tag{17}$$

where y_i is log average hourly wages, x is a vector of demographic characteristics (schooling, experience, etc.) thought to have an effect on wages, s_i is the BMI of an individual, and $f(\cdot)$ is some continuous function of s which the authors introduce to allow for the reasonable possibility that if BMI has an effect on wages, it is not necessarily linear.[65]

The most obvious possible problem is "confounding" – the relationship we observe between BMI and wages merely might merely represent the influence of other omitted factors that are correlated with BMI: in their model this is represented by a correlation between ϵ and u.[66] One such confounder they consider is "preferences for long-term investments, which we mean to represent characteristics that simultaneously impact decisions affecting both health and human capital accumulation."

One solution to this confounding problem is identification of an "instrumental variable" that provides "exogenous" variation in BMI (i.e a variable which is correlated with BMI but not correlated with the unobserved determinants of wages). The authors discuss two possible instrumental variables for BMI – mother's BMI and father's BMI and argue for their validity in several ways, including references to other literature.

In the case where $f(s)$ is linear in s, usual non–Bayesian practice is two stage least squares or the method of instrumental variables. One test which sometimes seems to capture the notion of a"severe" test of the hypothesis that the instrumental variables are valid is an "over–identification test." Specifically, if both instrumental variables are valid, the estimated effect of BMI should be similar whether mother's BMI is used alone as an instrumental variable, father's BMI is used alone, or both are used (Newey, 1985). If the test rejects, it is unclear how to proceed, but as the authors note:

[65] Although this is an important focus of their paper, it is one that I will not focus on since my interests in using this example lie elsewhere. To quote from their paper "Again, it is important to recognize that many applied studies in the treatment-response literature, and to our knowledge all of those that have been conducted on this specific topic, assume the relationship between the treatment variable and the outcome variable is linear (i.e., $f(s) = \alpha_0 + \alpha_1 s$), and define the slope of this function as the causal effect of interest. The assumption of linearity is likely made on computational considerations, as IV [instrumental variables] is simple to use in this context."

[66] Again because my interests lie elsewhere, one might not wish to stipulate that it is possible to talk clearly about a "causal effect" of BMI on wages, because such an effect seems to presuppose that in whatever manner we "manipulate" an individual's BMI, we would expect that the "effect" of BMI on wages would be the same. However, it is possible to imagine that the causal arrow *from* BMI *to* wages because 1) high BMI is equivalent to "bad health" and "bad health" lowers wages or 2) high BMI is equivalent to "unattractive" and employers discriminate against those who are "unattractive" for reasons possibly unrelated to "productivity." If the latter were true, a "successful" but "unhealthy diet" that lowered BMI would *raise* wages; if the former were true, such a diet would *lower* wages. See DiNardo (2007) for a discussion.

> From a theoretical perspective, however, it seems reasonable that the BMIs of the parents are either jointly valid as instruments, or jointly invalid, thus potentially calling into question what is actually learned from this procedure. On the empirical side, however, the correlation between parental BMI was found to be reasonably small (around .16), suggesting that something can be learned from this exercise, and that its implementation is not obviously redundant or 'circular.'"

The authors' observation that the correlation between parental BMI is small seems, to me, to suggest the importance of "severity". Had the correlation been much higher, one might have been tempted to conclude that the proposed test was "obviously redundant" or "circular."

Consequently, Kline and Tobias (2008) provide an "informal test" of the validity of the "exclusion restriction" by asking whether after using *one* instrumental variable, a model which excludes the other instrumental variable from the "structural equation" (equation 16 above) is well supported. Indeed they calculate the Bayes factor associated with the hypothesis that the effect of father's BMI on wages is zero while maintaining the validity of mother's BMI as an instrumental variable and find strong support for that hypothesis. They also find strong support for the reverse.

While this does not exhaust the "specification testing" performed in the study, it does indicate an attempt to put a hypothesis to the severest test possible. Interestingly, although the study is clearly a "Bayesian" analysis, the authors found it useful to conduct such a test in an "informal" way – without attempting to "shoehorn" the specification testing into a complete Bayesian analysis.[67]

8.2 Non–Bayesian Doesn't Have to Mean "Severe"

My personal view is that statistical theory is often useful for situations in which we are attempting to describe something that looks like a "chance set up." How one might go from information gleaned in such situations to draw inferences about other *different* situations, however, is not at all obvious. Some Bayesians might argue that one merely needs to formulate a prior, impose a "window on the world" (a.k.a. a likelihood) and then use Bayes' rule to revise our posterior probability. I am obviously uncomfortable with such a view and find Lecam's summary to the point:

> "The only precept or theory which seems relevant is the following: 'Do the best you can'. This may be taxing for the old noodle, but even the authority of Aristotle is not an acceptable substitute" (LeCam, 1977).

This view also comports well with C. S. Peirce's classic description of a "severe test" I discussed earlier.

However, even at this level of vagueness and generality, it is worth observing that such views are not shared by non–Bayesians, or if they are, there is no common vision of what is meant by severe testing. Except for the most committed Bayesians, nothing in statistical theory tells you how to "infer the truth of various propositions." As I have argued elsewhere (DiNardo, 2007), often the types of theories economists seem interested in are so vague that it is often impossible to know what, in principle, would constitute "evidence" even in an "ideal" situation.

Certainly it is the case that non–Bayesian researchers are frequently unwilling to use statistical tools to change their views about *some* assessments. For one clear example, compare Glaeser and Luttmer (1997) and Glaeser and Luttmer (2003). The latter paper is a revised version of the former paper. The papers "develops a framework to empirically test for misallocation. The methodology compares consumption patterns for demographic subgroups in rent-controlled and free-market places. [They] find that in New York

[67] It is also interesting to observe that this Bayesian "over identification test" is arguably better–suited to "severity" than recent non–Bayesian interpretations of such over identification tests. In the linear instrumental variables model, for example, the failure of the over–identification test has been recently re–interpreted *not* as a rejection of the premises of the estimated model but as evidence of "treatment effect" heterogeneity (Angrist, 2004). In this context, one possible cause (though not the only possible cause) of "treatment effect heterogeneity" is that the true relationship between BMI and wages is, say, quadratic but the investigator specifies a linear relationship. In such a case, one could no longer ensure that the estimated relationship would be invariant to the choice of instrumental variables even if the instrumental variables were "valid." The informal test proposed by Kline and Tobias (2008), however, easily accommodates such a situation since it allows $f(s)$ to be non–linear without exhausting any overidentification.

City, which is rent-controlled, an economically and statistically significant fraction of apartments appears to be misallocated across demographic subgroups"(Glaeser and Luttmer, 2003, page 1027).

A significant difference between the two papers is that the latter, Glaeser and Luttmer (2003), includes an interesting falsification test (not included in Glaeser and Luttmer (1997)). If the methodology works as intended, then when performing the same analysis on data from cities without rent–control – they consider Chicago, IL and Hartford CT – they should consistently estimate no welfare loss due to rent–control. Contrary to such a presumption, however, for both cities, they estimate large amounts of apartment misallocation (although these estimates are smaller than their estimate for NYC) that are statistically quite precise. Indeed, they acknowledge that "strictly interpreted, the results reject the identifying assumptions. In both cities, the procedure finds statistically significant misallocation"(Glaeser and Luttmer, 2003, page 1044). Nonetheless, while there are differences between the two versions of the papers, it is not clear whether such a rejection played any role in changing the inferences they draw. Indeed, they argue "While this is disturbing, the large difference between our New York results and the results for these placebo cities suggests that even though our identifying assumptions may not exactly be true, the failure of the assumptions is unlikely to fully account for the observed misallocation in New York"(Glaeser and Luttmer, 2003, page 1044). What does it mean to say a set of assumptions fails to to "fully account for the observed misallocation?" and why should the results be viewed as "disturbing?" (if indeed they should be?) In such a case, I think it is fair to say that we should have little confidence that their proposed *methodology* has a "truth preserving virtue." Note, however, this is only weakly related to one's views about the merits of rent–control in New York City.

Much of the variation among non–Bayesians in their reaction to such statistical informations seems to involve the "primacy" of certain types of (economic) *models*. Very roughly speaking, one can point to "a design–based approach" which focuses on creating or finding situations which resemble "chance set ups" and where an analysis of the DGP proceeds separately from a single specific highly articulated theoretical economic model. Historically, this approach has been associated with an emphasis on such issues as pre–specified analysis, "serious" specification testing, replicability, avoiding "confounding" , identification, etc.[68]

By contrast, one can also identify at least one strand of so–called "structural approaches" where there is little or no distinction between a DGP and a highly articulated "theoretical economic model"[69]. An archetypal example of this approach, perhaps, is the multinomial logit of McFadden (1974) in which the consumer choice model – utility function, specification of heterogeneity in tastes,etc – delivers a complete DGP in the form of a likelihood function. A feature of such an approach is that, in principle, once the model has been estimated, one can study "counterfactual policy simulations" or "experiments" which may have never been performed but can be described within the model.

This line of research gave birth to further developments which have yielded a wide variety of attitudes toward what might be called "severe testing." At one extreme, some researchers such as Edward Prescott apparently "completely reject econometrics as a useful scientific tool. Instead [Prescott] promotes *calibration* as the preferred method for "uncovering" the unknown parameters of structural models and for evaluating and comparing their ability to fit the data" (Rust, 2007, page 4).

While not rejecting the usefulness of statistics outright, Keane (2007) argues

> that determinations of the usefulness of ... 'well-executed' structural models – both as interpreters of existing data and vehicles for predicting the impact of policy interventions or changes in the forcing variables – should rest primarily on how well the model performs in validation exercises. By this I mean: (1) Does the model do a reasonable job of fitting important dimensions of the historical data on which it was fit? (2) Does the model do a reasonable job at out-of-sample prediction – especially when used to predict the impact of policy changes that alter the economic environment in some fundamental way from that which generated the data used in estimation?

[68] Some of this discussion draws from a brief discussion in unpublished lecture notes by David Card although for reasons of focus and brevity my account is not the same. (Card, 2007).

[69] It should not be surprising that term "structural model" encompasses a wide array of activities which have very different emphasis, including – to take just one example –classic studies of demand systems, etc. (See, for example, Deaton and Muellbauer (1980).) Moreover, some work involving "structural estimation" occurs in studies that also involve a design– based approach. For a simple illustration see DiNardo and Lemieux (1992, 2001). Consequently, I use the phrase "single strand" advisedly.

> My use of the word 'reasonable' here is unapologetically vague. *I believe that whether a model passes these criteria is an intrinsically subjective judgment, for which formal statistical testing provides little guidance. This perspective is consistent with how other sciences treat validation*"(Keane, 2007, page 32) (*emphasis added*).

Indeed Keane provides an illuminating illustration of this view by discussing an example of estimating the parameters of a life-cycle human capital investment model. After describing how the simplest version of the model fails to fit the data, he goes on to explain:

> "by adding a number of extra features that are not essential to the model, but that seem reasonable (like costs of returning to school, age effects in tastes for schooling, measurement error in wages, and so on), we were able to achieve what we regard as an excellent fit to the key quantitative features of the data – although formal statistical tests still rejected the hypothesis that the model is the 'true' data generating process (DGP). Despite these problems, there is nothing to indicate that the profession might be ready to drop the human capital investment model as a framework for explaining school and work choices over the life-cycle (Keane, 2007, page 33)

As one might expect, Keane does not set forth a specific context in which one might find estimates of such a model "useful" or to what extent, if any, the inferences drawn from such an approach should influence the choices we make or what we advocate to others. Surely the model with or without ammendments can't be "reasonable" for *all* contexts.

The "metastatistcal" question is "how much confidence should one have in a judgment supported by such an approach?" The answer, to say the least, is not obvious.

References

Abowd, J. and H. Farber (1982) Job queues and the union status of workers. *Industrial and Labor Relations Review* **35**.

Allais, M. (1953) Le comportement de l'homme rationnel devant le risque: Critique des postulats et axiomes de l'ecole americaine. *Econometrica* **21**, 503–546.

Allan, C. K., R. R. Thiagarajan, P. J. del Nido, S. J. Roth, M. C. Almodovar and P. C. Laussen (2007) Indication for initiation of mechanical circulatory support impacts survival of infants with shunted single-ventricle circulation supported with extracorporeal membrane oxygenation. *Journal of Thoracic and Cardiovascular Surgery* **133**, 660–667.

Angrist, J. (2004) Treatment effect heterogeneity in theory and practice. *Economic Journal* **114**, 52–83.

Ashenfelter, O. (1978) Union relative wage effects, new evidence, and a survey of their implications for wage inflation. In R. Stone and W. Peterson, eds., *Economic Contributions to Public Policy* MacMillan Press.

Askie, L. M. and T. Win (2003) The use of oxygen in neonatal medicine half a century of uncertainty. *Neoreviews* **4**, e340–348.

Bahra, A., M. Walsh, S. Menon and P. J. Goadsby (2003) Does chronic daily headache arise de novo in association with regular use of analgesics? *Headache* **43**, 179–90.

Barnard, G. A. (1947a) The meaning of a significance test. *Biometrika* **34**, 179–182.

Barnard, G. A. (1947b) A review of *Sequential Analysis* by Abraham Wald.*Journal of the American Statistical Association* **42**, 658–669.

Barnouw, J. (1979) A review of *the emergence of probability: a philosophical study of early ideas about probability, induction and statistical inference. Eighteenth Century Studies* **12**, 438–443.

Bartlett, R. H. (2005) Extracorporeal life support: history and new directions. *ASAIO J* **51**, 487–489.

Bartlett, R. H., D. W. Roloff, R. G. Cornell, A. F. Andrews, P. W. Dillon and J. B. Zwischenberger (1985) Extracorporeal circulation in neonatal respiratory failure: a prospective randomized study. *Pediatrics* **76**, 479–487.

Bayes, T. (1958) An essay towards solving a problem in the doctrine of chances. *Biometrika* **45**, 296–315 by the late Reverend Mr. Bayes, communicate by Mr. Price, in a letter to John Canton, M.A. and F.R.S.

Berger, J. O. and R. L. Wolpert (1988) *The Likelihood Principle* Hayward, CA: Institute of Mathematical Statistics 2nd edition.

Berry, D. A. (1989) [investigating therapies of potentially great benefit: Ecmo]: Comment: Ethics and ECMO. *Statistical Science* **4**, 337–340.

Berry, D. A. (2006) Bayesian clinical trials. *Nature Reviews Drug Discovery* **5**, 27–36.

Berry, S. M. and J. B. Kadane (1997) Optimal bayesian randomization. *Journal of the Royal Statistical Society, Part B* **59**, 813–189.

Bickel, P. J. and E. L. Lehmann (2001) Frequentist inference. In N. J. Smelser and P. B. Baltes, eds., *International Encyclopedia of the Social and Behavioral Sciences* 5789–5796 Oxford: Pergamon.

Boes, C. J. and D. J. Capobianco (2005) Chronic migraine and medication-overuse headache through the ages. *Cephalalgia* **25**, 378–90.

Bottke, W. F., D. Vokrouhlický and D. Nesvorný (2007) An asteroid breakup 160 million years ago as the probable source of the K/T impactor. *Nature* **449**, 48–53.

British Journal of Opthamology (1974) Editorial: Retrolental fibroplasia (rlf) unrelated to oxygen therapy. *British Journal of Ophthalmology* **58**, 487–489.

Capobianco, D. J., J. W. Swanson and D. W. Dodick (2001) Medication-induced (analgesic rebound) headache: historical aspects and initial descriptions of the North American experience. *Headache* **41**, 500–502.

Card, D. (1992) The effect of unions on the distribution of wages: Redistribution or relabelling? NBER Working Paper 4195 National Bureau of Economic Research Cambridge, MA.

Card, D. (1996) The effect of unions on the structure of wages: A longitudinal analysis. *Econometrica* **64**, 957–979.

Card, D. (2007) Lecture notes for Topics in Labor Economics, Department of Economics, Harvard University.

Card, D., T. Lemieux and C. W. Riddell (2003) Unionization and wage inequality: A comparative study of the U.S., U.K. and Canada. Working Paper 9473, National Bureau of Economic Research Cambridge, MA.

Cartwright, N. (1984) *How the Laws of Physics Lie*. New York: Oxford University Press.

Chew, S. H. (1983) A generalization of the quasilinear mean with application to the measurement of income inequality and decision–theory resolving the Allais paradox. *Econometrica* **51**, 1065–1092.

Chib, S. and B. H. Hamilton (2002) Semiparametric Bayes analysis of longitudinal data treatment models. *Journal of Econometrics* **110**, 67–89.

Couzin, J. (2004) The new math of clinical trials. *Science* **303**, 784 – 786.

Davis, P. G., A. Tan, C. O'Donnell and A. Schulze (2004) Resuscitation of newborn infants with 100% oxygen or air: a systematic review and meta-analysis. *Lancet* **364**, 1329–1333.

Deaton, A. and J. Muellbauer (1980) *Economics and Consumer Behavior* Cambridge: Cambridge University Press.

deFinetti, B. (1974) *The Theory of Probability*. New York: John Wiley two volumes.

DiNardo, J. (2007) Interesting questions in freakonomics. *Journal of Economic Literature* **45**, 973–1000.

DiNardo, J. and D. S. Lee (2004) Economic impacts of new unionization on private sector employers: 1984-2001. *Quarterly Journal of Economics* **119**, 1383 – 1441.

DiNardo, J. and T. Lemieux (1992) Alcohol, marijuana, and American youth: The unintended consequences of government regulation. NBER Working Paper 4212 National Bureau of Economic Research Cambridge, MA.

DiNardo, J. and T. Lemieux (1997) Diverging Male Wage Inequality in the United States and Canada, 1981–1988: Do Institutions Explain the Difference? *Industrial and Labor Relations Review* .

DiNardo, J. and T. Lemieux (2001) Alcohol, marijuana, and American youth: the unintended consequences of government regulation. *Journal of Health Economics* **20**, 991–1010.

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

Edmeads, J. (1990) Analgesic-induced headaches: an unrecognized epidemic. *Headache* **30**, 614–5.

Einstein, A. (1920) *Relativity: The Special and General Theory* New York: Henry Holt and Company translated by Robert W. Lawson.

Feller, W. (1950) *An Introduction to Probability Theory and Its Applications* Volume 1 Wiley.

Ferrari, A., C. Coccia and E. Sternieri (2007) Past, present, and future prospects of medication-overuse headache classification. *Headache* **Epub ahead of print.**

Fisher, C. M. (1988) Analgesic rebound headache refuted. *Headache* **28**, 666.

Food and Drug Administration, U.S. Department of Health and Human Services (2006) Guidance for the use of Bayesian statistics in medical device clinical trials - draft guidance for industry and FDA staff. Accessed Center for Devices and Radiological Health, Division of Biostatistics, Office of Surveillance and Biometrics.

Freedman, D. A. (1995) Some issues in the foundation of statistics. *Foundations of Science* **1**, 19–39.

Freeman, R. (1984) Longitudinal analysis of the effects of trade unions. *Journal of Labor Economics* **2**, 1–26.

Freeman, R. B. and M. Kleiner (1990) The impact of new unionization on wages and working conditions. *Journal of Labor Economics* **8**, S8–S25.

Freeman, R. B. and M. M. Kleiner (1999) Do unions make enterprises insolvent? *Industrial and Labor Relations Review* **52**, 510–527.

Friedman, M. (1950) Some comments on the significance of labor unions for economic policy. In D. M. Wright, ed., *The Impact of the Union: Eight Economic Theorists Evaluate the Labor Union Movement* New York: Harcourt, Brace and Company institute on the Structure of the Labor Market, American University, Washington D.C.

Fuchs, V. R., A. B. Krueger and J. M. Poterba (1998) Economists' views about parameters, values, and policies: Survey results in labor and public economics. *Journal of Economic Literature* **36**, 1387–1425.

Garber, D. and S. Zabell (1979) On the emergence of probability. *Archive for the History of the Exact Sciences* **21**, 33–53 communicated by C. Truesdell.

Gillies, D. (2000) *Philosophical Theories of Probability.* Philosophical Issues in Science London: Routledge.

Gladstone, J. P. and D. W. Dodick (2004) From hemicrania lunaris to hemicrania continua: an overview of the revised international classification of headache disorders. *Headache* **44**, 692–705.

Glaeser, E. L. and E. F. P. Luttmer (1997) The misallocation of housing under rent control. Working Paper 6220 National Bureau of Economic Research.

Glaeser, E. L. and E. F. P. Luttmer (2003) The misallocation of housing under rent control. *American Economic Review* **93**, 1027–1046.

Good, I. J. (1971) 46656 varities of Bayesians. *American Statistician* **25**, 62–63.

Good, I. J. (1983) *Good Thinking: The Foundations of Probability and Its Applications* Minneapolis: University of Minnesota Press.

Good, I. J. and R. A. Gaskins (1971) Nonparametric roughness penalties for probability densities. *Biometrika* **58**, 255–277.

Goodman, S. (2004) Basic Bayes – I. Technical report Johns Hopkins University National Institutes of Health Bethesda, Maryland from a Workshop sponsored by the Food and Drug Administration and Johns Hopkins University on "Can Bayesian Approaches to Studying New Treatments Improve Regulatory Decision–Making?".

Goodman, S. N. (1999) Toward evidence-based medical statistics. 1: The p value fallacy. *Ann Intern Med* **130**, 995–1004.

Groopman, J. (2007) *How Doctors Think*. Houghton Mifflin Company.

Gupta, V. K. (2004a) Chronic daily headache with analgesic overuse: epidemiology and impact on quality of life. *Neurology* **63**, 1341.

Gupta, V. K. (2004b) Classification of primary headaches: Pathophysiology versus nosology? *British Medical Journal* Letter to the Editor.

Gupta, V. K. (2004c) De novo headache and analgesic consumption: pathophysiological insights from nosologic complexity? *Headache* **44**, 375 375 375.

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

Hacking, I. (1967) Slightly more realistic personal probability. *Philosophy of Science* **34**, 311–325.

Hacking, I. (1975) *The Emergence of Probability: A Philisophical Study of Early Ideas About Probability, Induction and Statistical Inference* Cambridge: Cambridge University Press.

Hacking, I. (1983) *Representing and intervening: Introductory topics in the philosophy of natural science.* Cambridge, England: Cambridge University Press.

Hacking, I. (1990) *The Taming of Chance*. Number 17 in Ideas in Context Cambridge, England: Cambridge University Press.

Hacking, I. (2001) *An Intoduction to Probability and Inductive Logic*. Cambridge, UK: Cambridge University Press.

Hansmann, G. (2004) Neonatal resuscitation on air: it is time to turn down the oxygen tanks [corrected]. *Lancet* **364**, 1293–1294.

Headache Classification Committee of the International Headache Society (2004) The international classification of headache disorders: Second edition *Cephalalgia* **24**, 9–160.

Headache Classification Committee of the International Headache Society (2006) New appendix criteria open for a broader concept of chronic migraine. *Cephalalgia* **26**, 742.

Heckman, J. J. (1990) Varieties of selection bias. *American Economic Review* **80**, 313–318.

Hollis, S. and F. Campbell (1999) What is meant by intention to treat analysis? survey of published randomized controlled trials. *British Medical Journal* **319**, 670–674.

Hoover, K. (2007a) The rhetoric of signifying nothing: A rejoinder to Ziliak and McCloskey *Journal of Economic Methodology* **Forthcoming**.

Hoover, K. (2007b) Sound and fury: Mccloskey and significance testing in economics. *Journal of Economic Methodolog* **Forthcoming**.

Horowitz, J. L. and C. F. Manski (1998) Censoring of outcomes and regressors due to survey nonresponse: Identification and estimation using weights and imputations. *Journal of Econometrics* **84**, 37–58.

Howson, C. (1997) A logic of induction. *Philosophy of Science* **64**, 268–290.

Howson, C. and P. Urbach (1993) *Scientific Reasoning: The Bayesian Approach*. Chicago and La Salle, Illinois: Open Court, second edition.

Humphrey, T. (1992) Marshallian cross diagrams and their uses before alfred marshall: The origins of supply and demand geometry. *Economic Review* 3–23.

Jakubson, G. (1991) Estimation and testing of the union wage effect using panel data. *Review of Economic Studies* **58**, 971–991.

Jaynes, E. T. (1976) Confidence intervals vs Bayesian intervals. In W. L. Harper and C. A. Hooker, eds., *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science* volume II Holland: Reidel Publishing Co.

Jenkin, F. (1868) Trade–unions: how far legitimate? in S. C. Colvin and J. A. Ewing, eds., *Papers, Literary, Scientific, &c by the late Fleeming Jenkin* volume 2 London: Longmans, Green & Co. originally published in the *North British Review* March, 1868.

Jenkin, F. (1870) The graphic representation of the laws of supply and demand, and their application to labour in S. C. Colvin and J. A. Ewing, eds., *Papers, Literary, Scientific, &c by the late Fleeming Jenkin* volume 2 London: Longmans, Green & Co.

Jones, H. L. (1958) Inadmissible samples and confidence limits. *Journal of the American Statistical Association* **53**, 482–490.

Joyce, J. M. (1999) *The Foundations of Causal Decision Theory*. Cambridge, MA: MIT Press.

Keane, M. P. (2007) Structural vs. atheoretic approaches to econometrics. *Journal of Econometrics* `http://gemini.econ.umd.edu/jrust/research/JE_Keynote_7.pdf`.

Keynes, J. M. (1921) *A Treatise on Probability*. London, England: Macmillan and Co., Limited.

Kline, B. and J. Tobias (2008) The wages of BMI: Bayesian analysis of a skewed treatment-response model with nonparametric endogeneity. *Journal of Applied Econometrics* .

Kmenta, J. (2000) *Elements of Econometrics*. Ann Arbor, MI: The University of Michigan Press second edition.

Koenker, R. (2007) Personal communication.

Krashinsky, H. A. (2004) Do marital status and computer usage really change the wage structure? *Journal of Human Resources* **3**, 774–791.

Kyburg Jr., H. E. and M. Thalos, eds. (2003) *Probability Is the Very Guide to Life: The Philosophical Uses of Chance*. Chicago and La Salle, Illinois: Open Court.

Lance, F., C. Parkes and M. Wilkinson (1988) Does analgesic abuse cause headaches de novo? *Headache* **28**, 61–2.

Laplace, P. S. (1795) *Essai Philosophique sur les Probabilités*. New York: John Wiley & Sons sixth edition translated as "Philosophical Essay on Probabilities" from the Sixth French Edition by Frederick Wilson Truscott and Frederick Lincoln Emory and First U.S. edition published in 1902.

LeCam, L. (1977) A note on metastatistics or 'an essay toward stating a problem in the doctrine of chances'. *Synthese* **36**, 133–160.

LeCam, L. (1990) Maximum likelihood: An introduction. *International Statistical Review* **58**, 153–171.

Lemieux, T. (1998) Estimating the effects of unions on wage inequality in a panel data model with comparative advantage and non–random selection. *Journal of Labor Economics* **16**, 261–291.

Lequier, L. (2004) Extracorporeal Life Support in Pediatric and Neonatal Critical Care: A Review. *Journal of Intensive Care Medicine* **19**, 243–258.

Lewis, H. G. (1963) *Unionism and relative wages in the United States*. Chicago: University of Chicago Press.

Lewis, H. G. (1986) Union relative wage effects. In O. Ashenfelter and R. Layard, eds., *Handbook of Labor Economics* volume 2 chapter 20, 1139–1182 Amsterdam: North Holland.

Lindley, D. V. (1971) The estimation of many parameters. In V. P. Godambe and D. A. Sprott, eds., *Foundations of Statistical Inference* 435–447 Toronto, Canada: Holt, Rinehart and Winston.

Lindley, D. V. (2000) The philosophy of statistics. *The Statistician* **49**, 293–337.

Lipman, A. G. (2004) Does opiophobia exist among pain specialists? *Journal of Pain and Palliative Care Pharmacotherapy* **18**, 1–5.

Mathew, N. T. (2008) Response: Drug–induced refractory headache. *Headache* **48**, 729.

Mathew, N. T., R. Kurman and F. Perez (1990) Drug induced refractory headache – clinical features and management. *Headache* **30**, 634–538.

Mayo, D. G. (1979) Testing statistical testing. In J. C. Pitt, ed., *Philosophy in Economics* volume 16 of *The University of Western Ontario Series in Philosophy of Science* 175–203 Dordrecht, Holland: D. Reidel Publishing Company papers Deriving from and Related to a Workshop on Testability and Explanation in Economics held at Virginia Polytechnic Institute and State University, 1979.

Mayo, D. G. (1982) On after–trial criticisms of neyman–pearson theory of statistics. *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association* **1**, 145–158.

Mayo, D. G. (1996) *Error and the Growth of Experimental Knowledge* Science and Its Conceptual Foundations. Chicago: University of Chicago Press.

Mayo, D. G. (2003) Severe testing as a guide for inductive reasoning. In H. E. Kyburg, Jr and M. Thalos, eds., *Probability Is the Very Guide of Life: The Philosophical Uses of Chance* 89–117 Chicago and LaSalle, Illinois.

Mayo, D. G. and M. Kruse (2002) Principles of inference and their consequences. in D. Corfield and J. Williamson, eds., *Foundations of Bayesianism* volume 24 of *Applied Logic* Kluwer Academic Publishers.

Mayo, D. G. and A. Spanos (2006) Severe testing as a basic concept in a neymanpearson philosophy of induction. *British Journal for the Philosophy of Science* **57**, 323–357.

McCloskey, D. (1985) The loss function has been mislaid: The rhetoric of significance tests. *American Economic Review* **Supplement 75**, 201–205.

McCloskey, D. and S. Ziliak (1996) The standard error of regressions. *Journal of Economic Literature* **34**, 97–114.

McFadden, D. (1974) Conditional logit analysis of qualitative choice behavior. In P. Zarembka, ed., *Frontiers of Econometrics* volume 3 chapter 4, 105–142 New York and London: Academic Press.

Meldrum, M. L. (2003) A capsule history of pain management. *Journal of the American Medical Association* **290**, 2470–2475.

Newey, W. (1985) Generalized method of moments specification tests. *Journal of Econometrics* **29**, 229–256.

Neyman, J. (1957) Inductive behavior as a basic concept of philosophy of science. *Revue dInstitute International de Statistique* **25**, 7–22.

Obermann, M. and Z. Katsarava (2007) Management of medication-overuse headache. *Expert Review of Neurotherapeutics* **7**, 1145–1155.

Olesen, J., M.-G. Bousser, H.-C. Diener, D. Dodick, M. First, P. Goadsby, H. Gbel, M. Lainez, J. Lance, R. Lipton, G. Nappi, F. Sakai, J. Schoenen, S. Silberstein and T. Steiner (2006) New appendix criteria open for a broader concept of chronic migraine. *Cephalalgia* **26**, 742.

Paneth, N. and S. Wallenstein (1985) Extracorporeal membrane oxygenation and the play the winner rule. *Pediatrics* **76**, 622–623.

Peirce, C. S. (1878a) The doctrine of chances *Popular Science Monthly* **12**, 604–615.

Peirce, C. S. (1878b) The probability of induction *Popular Science Monthly* **12**, 705–718.

Peirce, C. S. (1958) Collected papers in A. Burks, ed., *Collected Papers* volume 7–8 Cambridge, MA: Harvard University Press.

Pini, L.-A., A. Cicero and M. Sandrini (2001) Long–term follow–up of patients treated for chronic headache with analgesic overuse *Cephalalgia* **21**, 878–883.

Poirier, D. J. (1995) *Intermediate Statistics and Econometrics: A Comparative Approach* Cambridge, MA: MIT Press.

Rust, J. (2007) Comments on: "structural vs. atheoretic approaches to econometrics" *Journal of Econometrics* .

Saper, J. R. (2005) Editorial to the guidlines for trials of behavioral treatments for recurrent headache *Headache* **45 (Supplement 2)**, S90–S91.

Saper, J. R. and A. E. Lake (2006a) Medication overuse headache: type i and type ii. *Cephalalgia* **26**, 1262.

Saper, J. R. and A. E. Lake (2006b) Sustained opioid therapy should rarely be administered to headache patients: Clinical observations, literature review and proposed guidelines *Headache Currents* **3**, 67–70.

Saper, J. R., A. E. Lake, R. L. Hamel, T. E. Lutz, B. Branca, D. B. Sims and M. M. Kroll (2004) Daily scheduled opioids for intractable head pain: long-term observations of a treatment program. *Neurology* **62**, 1687–94.

Savage, L. J. (1967) Difficulties in the theory of personal probability *Philosophy of Science* **34**, 305–310.

Savage, L. J. (1972) *The Foundations of Statistics* New York: Dover Publications revised and expanded version of the original 1954 work.

Savage, L. J., M. Bartlett, G. A. Barnard, D. R. Cox, E. S. Pearson, C. A. B. Smith et al. (1962) The foundations of statistical inference: A discussion in G. A. Barnard and D. R. Cox, eds., *The Foundations of Statistical Inference: A Discussion* Meuthen's Monographs on Applied Probability and Statistics London & Colchester: Spottiswoode Ballantyne & Co. Ltd. a Discussion Opened by L. J. Savage at the Joint Statistics Seminar of Birbeck and Imperial Colleges. Discussants also include H. Ruben, I. J. Good, D.V. Lindley, P. Armitage, C.B. Winsten, R. Syski, E. D. Van Rest and G. M. Jenkins.

Scadding, J. W. (2004) Treatment of neuropathic pain: historical aspects. *Pain Medicine* **5 Suppl 1**, S3–S8.

Schuster, L. (2004) Revised guidelines for medication overuse headache *Neurology Reviews* **12**.

Shah, P. S. (2005) Resuscitation of newborn infants. *Lancet* **365**, 651–2; author reply 652–3.

Silberstein, S. (2004) Introduction: Aching heads in J. Kempner, ed., *Aching Heads, Making Medicine: Gender and Legitimacy in Headache* Philadelphia, PA.

Silberstein, S. D., R. B. Lipton, S. Solomon and N. T. Mathew (1994) Classification of daily and near-daily headaches: proposed revisions to the IHS criteria. *Headache* **34**, 1–7.

Silverman, W. A. (1980) *Retrolental Fibroplasia. A Modern Parable* London: Grune and Stratton.

Silverman, W. A. (2004) A cautionary tale about supplemental oxygen: the albatross of neonatal medicine. *Pediatrics* **113**, 394–396.

Simon, J. (1997) The philosophy and practice of resampling statistics accessed May 1, 2008.

Sober, E. (2002) Bayesianism its scope and limits in R. Swinburne, ed., *Proceedings of the British Academy* volume 113 21–38 The British Academy Oxford: Oxford University Press.

Society, I. H. (1988) Classification and diagnostic criteria for headache disorders, cranial neuralgias and facial pain *Cephalagia* **8**, 1–96.

Stigler, S. M. (1982) Thomas bayes's bayesian inference *Journal of the Royal Statistical Society.* **145**, 250–258 series A (General).

Thourani, V. H., P. M. Kirshbom, K. R. Kanter, J. Simsic, B. E. Kogon, S. Wagoner, F. Dykes, J. Fortenberry and J. M. Forbess (2006) Venoarterial Extracorporeal Membrane Oxygenation (VA-ECMO) in Pediatric Cardiac Support *Annals of Thoracic Surgery* **82**, 138–145.

Urbach, P. (1985) Randomization and the design of experiments *Philosophy of Science* **52**, 256–273.

Vanderveen, D. K., T. A. Mansfield and E. C. Eichenwald (2006) Lower oxygen saturation alarm limits decrease the severity of retinopathy of prematurity. *J AAPOS* **10**, 445–448.

Venn, J. (1888) *The Logic of Chance: An Essay on the Foundations and Province of the Theory of Probability, With Especial Reference to its Logical Bearings and its Application to Moral and Social Science* London and New York: Macmillan third edition first Edition, 1866. Second Edition, greatly expanded 1876. Third Edition 1888.

von Mises, R. (1957) *Probability, Statistics, and Truth* London and New York: Allen and Unwin Ltd. and the MacMillan Company.

Ward, T. N. (2008) My favorite article: Drug–induced refractory headache *Headache* **48**, 728–729.

Ware, J. H. (1989) Investigating therapies of potentially great benefit: Ecmo *Statistical Science* **4**, 298–306.

Wei, L. J. and S. Durham (1978) The randomized play-the-winner rule in medical trials *Journal of the American Statistical Association* **73**, 840–843.

Whitney, C. W. and M. Von Korff (1992) Regression to the mean in treated versus untreated chronic pain *Pain* **50**, 281–285.

Willis, T. (1675) *Two discourses concerning the soul of brutes which is that of the vital and sensitive of man : the first is physiological shewing the nature, parts, powers, and affections of the same : the other is pathological, which unfolds the diseases which affect it and its primary seat; to wit, the brain and nervous stock, and treats of their cures : with copper cuts ; by Thomas Willis ; Englished by S. Pordage, student in physick.* translation of De anima brutorum. Englished by S. Pordage, student in physick. Lodon: printed for Thomas Dring at the Harrow near Chancery-Lane End in Fleetstreet Ch. Harper at the Flower-de-Luce against St. Dunstan's Church in Fleet-street, and John Leigh at Stationers-Hall.

Wilson, E. B. (1952) *An Introduction to Scientific Research* New York: McGraw–Hill.

Zed, P. J., P. S. Loewen and G. Robinson (1999) Medication-induced headache: overview and systematic review of therapeutic approaches. *Ann Pharmacother* **33**, 61–72.

Zelen, M. (1969) Play the winner rule and the controlled clinical trial *Journal of the American Statistical Association* **64**, 131–146.

Zellner, A. (1984) Causality and econometrics in *Basic Issues in Econometrics* chapter 1, 35–74 Chicago and London: University of Chicago Press.