Choosing The More Likely Hypothesis

Richard Startz*

revised October 2014

Abstract

Much of economists' statistical work centers on testing hypotheses in which parameter values are partitioned between a null hypothesis and an alternative hypothesis in order to distinguish two views about the world. Our traditional procedures are based on the probabilities of a test statistic under the null but ignore what the statistics say about the probability of the test statistic under the alternative. Traditional procedures are not intended to provide evidence for the relative probabilities of the null versus alternative hypotheses, but are regularly treated as if they do. Unfortunately, when used to distinguish two views of the world, traditional procedures can lead to wildly misleading inference. In order to correctly distinguish between two views of the world, one needs to report the probabilities of the hypotheses given parameter estimates rather than the probability of the parameter estimates given the hypotheses. This monograph shows why failing to consider the alternative hypothesis often leads to incorrect conclusions. I show that for most standard econometric estimators, it is not difficult to compute the proper probabilities using Bayes theorem. Simple formulas that require only information already available in standard estimation reports are provided. I emphasize that frequentist approaches for deciding between the null and alternative hypothesis are not free of priors. Rather, the usual procedures involve an implicit, unstated prior that is likely to be far from scientifically neutral.

* Department of Economics, University of California, Santa Barbara, email: startz@ucsb.edu. Portions of this paper appeared in an earlier working paper titled, "How Should An Economist Do Statistics?" Advice from Jerry Hausman, Shelly Lundberg, Gene Savin, Meredith Startz, Doug Steigerwald, and members of the UCSB Econometrics Working Group is much appreciated.

Contents

1. Introduction

Much of economists' statistical work centers on testing hypotheses in which parameter values are partitioned between a null hypothesis and an alternative hypothesis. In essence, we are trying to distinguish between two views about the world. We then ask where the estimated coefficient (or test statistic) lies in the distribution implied by the null hypothesis. If the estimated coefficient is so far out in the tail of the distribution that it is very unlikely we would have found such an estimate under the null, we reject the null and conclude there is significant evidence in favor of the alternative. But this is a terribly incomplete exercise, omitting any consideration of how unlikely it would be for us to see the estimated coefficient if the alternative were true. Pearson (1938, p. 242) put the argument this way,[1]

> [the] idea which has formed the basis of all the…researches of Neyman and myself…is the simple suggestion that the only valid reason for rejecting a statistical hypothesis is that some alternative hypothesis explains the events with a greater degree of probability.

The principle that the probability of a realized coefficient under the alternative matters is at once well-understood and near-universally ignored by economists. What is less appreciated is the practical point: Our standard procedure of stating whether a coefficient is statistically significant (or equivalently whether the hypothesized value of a coefficient lies outside the confidence interval, or equivalently whether the *p*-value is small) can be a terribly misleading guide as to the odds favoring the null hypothesis relative to the alternative hypothesis. I give examples below to show just how misleading our usual procedures can be. Of course, for

[1] As quoted by Weakliem (1999, p. 363).

practice to change, there needs to be a better way to conduct inference. I present alternative procedures that can be easily implemented in our most common hypothesis testing situations.

My goal here is to offer a perspective on how economists should choose between hypotheses. While some of the points are original, many are not. After all, much of the paper comes down to saying "remember Bayes theorem," which has likely been around since Bayes (1763); or according to the delightful account by McGrayne (2011), at least since Laplace (1774). While it is entirely clear that economists do choose between hypotheses using statistical tests as if Bayes theorem does not exist, it is not because we have not been reminded of the danger of such practice. It seems the advice didn't take. Leamer (1983) laid out much of the argument in the very first volume of the *Handbook of Econometrics*. McCloskey (1992) reports a discussion in which Ken Arrow said, "Statistical significance in its usual form is indefensible." In an influential article in the medical literature, Ioannidis (2005) reminds medical researchers "…the probability that a research finding is indeed true depends on the prior probability of it being true…, the statistical power of the study, and the level of statistical significance." Kass and Raftery (1995) offer some of the theory behind what's said below. The discussion in this monograph is at least foreshadowed in Pearson (1938) and Arrow (1960) and parts are pretty explicit in Leamer (1978, 1983a) and Raftery (1986a, 1995). The hope is that by (a) giving blunt examples of the consequences of ignoring Bayes theorem and (b) offering very easy ways to adjust frequentist statistics to properly account for Bayes theorem, econometric practice may change more than it has in the past.

This monograph is aimed primarily at the classical, frequentist, econometrician who needs to choose between hypotheses. Most results are illustrated in the context of the most simple

econometric situation, one where we have a normally distributed estimator, $\hat{\theta} \sim N(\theta, \sigma_{\hat{\theta}}^2)$, and a null hypothesis $\theta = \theta_0$ versus an alternative $\theta \neq \theta_0$. The canonical example is a test of a regression coefficient. There are five major points.

1. The traditional use of classical hypothesis testing to choose between hypotheses leads to misleading results. As a practical matter, standard practice can be very, *very* misleading. It is entirely possible to strongly reject the null in cases where the null is more likely than the alternative, and vice versa.
2. Choosing between hypotheses requires invoking Bayes theorem. For the most common empirical applications at least, those where the estimated coefficients are approximately normal, applying Bayes theorem is very easy.
3. Once one acknowledges that one wants to compare a null hypothesis to an alternative, something has to be said about the likelihood of particular values of the parameter of interest under the alternative. Use of Bayes theorem does require specifying some prior beliefs. Sometimes this can be done in a way in which the specified priors take a neutral stance between null and alternative; sometimes a completely neutral stance is more difficult.
4. The notion that frequentist procedures specify a null and then take a neutral stance with regard to parameter values under the alternative is wrong. Frequentist decision rules are equivalent to adopting an implicit prior. The implicit prior is often decidedly non-neutral.
5. Economic hypotheses are usually best distinguished by some parameter being small or large, rather than some parameter being exactly zero versus non-zero. Application of

Bayes theorem permits the former, preferred, kind of hypothesis comparison by considering non-sharp nulls. The calculations required for choosing between non-sharp hypotheses are straightforward.

All of this is, obviously, related to frequentist versus Bayesian approaches to econometrics. This paper is addressed to the frequentist econometrician. Nothing addresses any of the philosophical differences between frequentists and Bayesians. Some Bayesian tools are used, although these are really just statements of probability theory and should be uncontroversial. A succinct statement of the goal of the monograph is this:

After running a regression, the empirical economist should be able to draw an inference about the probability of a null versus an alternative hypothesis that is both correct and easy to make.

2. Choosing between hypotheses

In most applications, hypothesis testing begins by stating a null and alternative hypothesis that partition a parameter space. The heart of the frequentist hypothesis testing enterprise is to devise tests where the probability of falsely rejecting the null in favor of the alternative is controlled and is small. There is no problem with the frequentist probability calculations, which are correct of course. Indeed, most of the probabilistic calculations that follow are based on the frequentist results. The issue is that the frequentist approach asks the wrong question. We should be concerned with the relative probabilities of the competing hypotheses, not merely whether the null hypothesis is unlikely. Should you reject an unlikely null in favor of an even more unlikely alternative? Should you "accept" a null yielding an insignificant test statistic if the alternative is more likely than the null?

While formal tests of statistical significance ignore the role of the alternative hypothesis, economists are generally comfortable with the idea that we should pay attention to the power of a test (the probability of rejecting the null under the alternative) as a way to informally bring the probability of the realized coefficient under the alternative into the discussion. Doing so is considered good practice for good reason. As McCloskey (1992) reports from a discussion with Ken Arrow, Arrow's view is that "[statistical significance] is not useless, merely grossly unbalanced if one does not speak also of the power of the test." Or as Arrow (1960, p. 70), writing in honor of Harold Hotelling said,

> It is very remarkable that the rapid development of decision theory has had so little effect on statistical practice. Ever since the classic work of Neyman and Pearson (1933), it has been apparent that, in the choice of a test for a hypothesis, the power of a test should play a role coordinate with the level of significance.

The discussion of power, which continues below, reminds us that in comparing the relative likelihood of a null versus an alternative hypothesis, we need to consider probabilities generated under both hypotheses. The classical test statistics, which are derived under the null hypothesis alone, leave something out.

If our task is to decide whether the data favors the null or the data favors the alternative or the data speaks too quietly to let us decide between the two, then our standard test procedures will often lead us to an incorrect choice for a simple reason. Our standard procedure tells us whether the estimated coefficient is unlikely given the null, not whether the null is unlikely given the estimated coefficient. In other words, our calculations are based on

$\Pr(\hat{\theta}|H_0)$ when what we are interested in is $\Pr(H_0|\hat{\theta})$. The two are related—through Bayes theorem—but they are not the same and can be quite different in practice. The central message here is that rather than offering the traditional statements about statistical significance, economists should remember Bayes theorem and report the probability that the null hypothesis is true given the observed data, $\Pr(H_0|\hat{\theta})$. I show how to do this below.

Using Bayes theorem to compute $\Pr(H_0|\hat{\theta})$ is most often straightforward, but it does introduce one complication. Bayes theorem depends on the unconditional probability of each hypothesis, a value that economists by tradition prefer not to specify. I begin with two points about the unconditional probability. The first point is that it is sometimes easy to specify an unconditional probability that might be considered neutral between the null and the alternative, for example that the unconditional odds of the null and the alternative are 50/50. Second, declining to state an unconditional probability explicitly doesn't mean that you haven't picked one, it just means that your implicit decision follows from the math of Bayes theorem. I give examples below in which I back out the "implicit prior" that comes from using the usual approach to statistical significance and I show that these priors are often not at all neutral. In one example below where the traditional test statistic strongly rejects the null (p-value $= 0.01$), interpreting the evidence as saying that the null is "significantly unlikely" requires that the investigator began with an implicit prior unconditional probability favoring the alternative of 85 percent or more. So our standard procedures can be very far from "letting the data speak."

It is worth saying that while there have been deep philosophical disputes between classical and Bayesian statisticians over the nature of probability, these disputes don't have anything to

do with the current discussion. What is being used from Bayesian statistics is Bayes theorem and some mathematical tools. There are many papers advocating for more general advantages of Bayesian analysis; Leamer (1978) is particularly prominent. While Bayes theorem is key, invoking Bayes theorem does not make one a Bayesian. Neyman and Pearson (1933, p. 289) write, "The problem of testing statistical hypotheses is an old one. Its origin is usually connected with the name of Thomas Bayes."

Here is the plan of the remainder of the paper. First I review the use of Bayes theorem to show just where the probability of the estimated coefficient under the alternative enters into the required calculations. Next I work through a "toy" model in which the null and the alternative are each a single point. The simple nature of the model illustrates that the standard approach to statistical significance can be a terribly misleading guide to the probabilities that we actually care about: the probability of the null, $p_{H_0} \equiv \Pr\left(H_0 \middle| \hat{\theta}\right)$, and the probability of the alternative, $p_{H_A} \equiv \Pr\left(H_A \middle| \hat{\theta}\right)$. In this simple model it is easy to show what a "reject if significant at the 5 percent (or whatever) level" rule implies about the investigator's unconditional prior on the hypotheses. Of course, in most real models the null is a point and the alternative is a continuum. I turn to this situation next. Everything learned from the simple example continues to be true, but both new complications and new opportunities arise. Here is one particularly noteworthy complication: We usually think our standard techniques are largely agnostic with regard to prior beliefs about the parameter of interest; this turns out to be not at all true. One pleasant opportunity that comes from invoking Bayes theorem is that it becomes easy to avoid having "sharp" hypotheses that associate an economic theory with an exact value for a

parameter. Instead we can specify a null hypothesis that θ lies "close to" θ_0 rather than being limited to $\theta = \theta_0$, which often gives a test truer to the spirit of the economic theory.

I discuss the relation of the approach offered here to a more general Bayesian approach, and discuss the use of the Bayesian Information Criterion as a useful shortcut for applying Bayes theorem to compute the probability of the null without being explicit about a prior. Finally, I briefly relate the problem of choosing between hypotheses to the more general question of using decision theory to make probabilistic choices.

3. Bayes Theorem

In this section I apply Bayes theorem to the output of standard econometric estimation. In this way we translate from the distribution of the estimator conditional on the null hypothesis, which comes from the usual frequentist analysis, to the distribution of the null hypothesis conditional on the estimator, which is the object of interest. A key point is that the probability of the estimator under the alternative plays a required role in calculating the probability of the null. This is unsurprising, once one thinks about the role played by power in reaching conclusions based on traditional tests. In the last subsection, I begin a discussion of the fact that the traditional frequentist measure of the strength of the null, the *p*-value, is not generally a good measure of the probability of the null. This last is a theme to which I return repeatedly.

3.1 Bayes Theorem Applied to Traditional Estimators

Most traditional hypothesis testing flows from calculations using the probability that we observe some value $\hat{\theta}$ conditional on the null being true, $\Pr\left(\hat{\theta}\middle|H_0\right)$. But to sort out the null

hypothesis from the alternative, what we really want to know is the probability of the null given that we observe $\hat{\theta}$, $\Pr(H_0|\hat{\theta})$, so the conditioning goes the other way. Bayes theorem tells us

$$p_{H_0} \equiv \Pr(H_0|\hat{\theta}) = \Pr\left(\hat{\theta}\middle|H_0\right) \times \frac{\pi(H_0)}{\Pr\left(\hat{\theta}\right)} \tag{1}$$

where $\pi(H_0)$ is an unconditional probability or "prior" on the null and $\Pr\left(\hat{\theta}\right)$ can be thought of as a normalizing constant that will drop out of the problem. Using the analogous equation for the alternative, $p_{H_A} \equiv \Pr\left(\hat{\theta}\middle|H_A\right) \cdot \frac{\pi(H_A)}{\Pr(\hat{\theta})}$, leads to the *posterior odds ratio*

$$PO_{0A} = \frac{p_{H_0}}{p_{H_A}} = \frac{\Pr\left(H_0\middle|\hat{\theta}\right)}{\Pr\left(H_A\middle|\hat{\theta}\right)} = \frac{\Pr\left(\hat{\theta}\middle|H_0\right)}{\Pr\left(\hat{\theta}\middle|H_A\right)} \times \frac{\pi(H_0)}{\pi(H_A)} \tag{2}$$

The posterior odds ratio, which gives the relative probability of the null versus the alternative having seen and analyzed the data, is composed of two factors. The ratio $\Pr\left(\hat{\theta}\middle|H_0\right)/\Pr\left(\hat{\theta}\middle|H_A\right)$ is the *Bayes factor,* sometimes denoted B_{0A}. The Bayes factor tells us how our statistical analysis has changed our mind about the relative odds of the null versus alternative hypothesis being true. In other words, the Bayes factor tells us how we change from the prior odds, $\pi(H_0)/\pi(H_A)$, to the posterior odds. Using equation (2) and the fact that probabilities sum to one, so $p_{H_0} + p_{H_A} = 1$ and $\pi(H_0) + \pi(H_A) = 1$, equation (3) gives the probability of the null being true conditional on the estimated parameter.

$$p_{H_0} = \frac{\Pr\left(\hat{\theta}\middle|H_0\right) \cdot \pi(H_0)}{\Pr\left(\hat{\theta}\middle|H_0\right) \cdot \pi(H_0) + \Pr\left(\hat{\theta}\middle|H_A\right) \cdot \left(1 - \pi(H_0)\right)} \tag{3}$$

Economists should report the value p_{H_0}, or equivalently the posterior odds, because this is the probability that distinguishes the evidence in favor of one hypothesis over the other—

which is almost always the question of interest. This means that the value of $\pi(H_0)$ in equations (2) and (3) cannot be ignored. However, an arguably neutral position on $\pi(H_0)$ is to set it to one-half. In this case the posterior odds ratio is simply the Bayes factor, so one reports how the analysis changes the evidence on the relative merits of the competing hypotheses. Or where reporting p_{H_0} is preferred to reporting the Bayes factor, equation (3) simplifies to

$$p_{H_0} = \frac{\Pr(\hat{\theta}|H_0)}{\Pr(\hat{\theta}|H_0) + \Pr(\hat{\theta}|H_A)} \tag{4}$$

There will be cases of empirical work in which there is a rough consensus value for $\pi(H_0)$ other than one-half. In such cases equation (3) is obviously preferable to the simplified version given in equation (4). When this is not the case, there is a second argument in addition to "arguably neutral" in favor of using equation (4). The Bayes factor in equation (2) tells us how the evidence from observing $\hat{\theta}$ changes the relative probabilities of the hypotheses from the prior odds ratio to the posterior odds ratio. Equation (4) gives a probability of the null based only on the Bayes factor. In this limited sense, the calculation of p_{H_0} given in equation (4) gives an estimate based only on the data.

Loosely speaking, the problem with our standard $\Pr(\hat{\theta}|H_0)$-based approach is that it ignores the probability of observing $\hat{\theta}$ under the alternative; $\Pr(\hat{\theta}|H_A)$ matters too. In what follows, I show that the difference between looking at p_{H_0} versus the standard approach to statistical significance is not simply a philosophical point—the two often lead to quite different conclusions about whether the data tells us that the null or the alternative is the better model.

It is entirely possible to strongly reject the null when the probability of the alternative is not high, indeed even when the alternative is less probable than the null.

3.2 Bayes Theorem and Power

The fact that $\Pr(\hat{\theta}|H_A)$ matters should not come as a surprise, as it is closely linked to the idea that the power of a test should be considered in making decisions based on statistical evidence. As Hausman (1978) put it,

> Unfortunately, power considerations have not been paid much attention in econometrics… Power considerations are important because they give the probability of rejecting the null hypothesis when it is false. In many empirical investigations [two coefficients] … seem to be far apart yet the null hypothesis [that the difference equals zero] … is not rejected. If the probability of rejection is small for a difference … large enough to be important, then not much information has been provided by the test.

In order to build intuition, we can look at what power considerations tell us about the importance of the alternative. Let τ be the outcome of a test coded 1 if the investigator rejects the null hypothesis in favor of the alternative and 0 if the test does not reject. If α is the size of the test, then $\Pr(\tau = 0|H_0) \equiv 1 - \alpha$. If β is the power of the test under the alternative, then $\Pr(\tau = 0|H_A) \equiv 1 - \beta$. Bayes theorem applied to the test outcome tells us

$$\Pr(H_0|\tau = 0) = \frac{\Pr(\tau = 0|H_0)\pi(H_0)}{p(\tau = 0)}, \Pr(H_A|\tau = 0) = \frac{\Pr(\tau = 0|H_A)\pi(H_A)}{p(\tau = 0)} \quad (5)$$

The Bayes factor based on the test outcome is

$$B_{0A} = \frac{\Pr(H_0|\tau = 0)}{\Pr(H_A|\tau = 0)} = \frac{1-\alpha}{1-\beta} \tag{6}$$

and under $\pi(H_0) = \pi(H_A)$ the probability of the null given the test failing to reject is

$$\Pr(H_0|\tau = 0) = \frac{1-\alpha}{(1-\alpha)+(1-\beta)} \tag{7}$$

When a classical test fails to reject the null, a standard, albeit informal, practice is to temper the conclusion that failing to reject implies the truth of the null with commentary on whether the test has good power. If a test is known to have high power, then we typically conclude that a failure to reject the null means the null is likely to be true. But if a test has low power, then the outcome of a test tells us very little. Equations (6) and (7) formalize the intuition behind this practice.

What do we learn from a test with extreme power characteristics in the case that we do not reject the null? Suppose first that a test has very high power, perhaps due to a large number of observations, so $\beta \approx 1$. If the null is false we will almost certainly reject, implying that if we fail to reject the null it must be because the null is true. In terms of equation (7), $\beta \approx 1 \Longrightarrow \Pr(H_0|\tau = 0) \approx 1$. Obversely, if $power \approx size$, $\beta \approx \alpha$, nothing is learned by a failure to reject. We get $B_{0A} = 1$ and $\Pr(H_0|\tau = 0) = 0.50$, in equations (6) and (7).

When a test rejects the null we have

$$\Pr(H_0|\tau = 1) = \frac{\alpha}{\alpha+\beta} \tag{8}$$

If power equals size, then a rejection means nothing since $\Pr(H_0|\tau = 1) = 0.5$. For a very powerful test, $\beta \approx 1$, a rejection is informative since size matters.

More generally, which hypothesis is more likely given a test result depends on both size and power. If we have a test with size equal to the five percent "gold standard" and power equal to

25 percent (just as an example), then the proper conclusion on rejection is that the probability of the alternative is 83 percent—which is strong but well short of "95 percent." And a failure to reject leads only to $p_{H_0} = 0.56$, which is hardly conclusive at all.

All of which is to say that ignoring $\Pr(\hat{\theta}|H_A)$ or ignoring power can lead to considerable error in assessing the relative probability of one hypothesis versus another.

3.3 Does the *p*-Value Approximate the Probability of the Null?

One might reasonably ask why anyone would ever use a standard hypothesis test to choose between hypotheses. Presumably, the reasoning is something like "if the observed test statistic is very unlikely under the null, the null must be unlikely and therefore the alternative is more likely than the null." DeGroot wrote (DeGroot (1973)),

> Because the tail area… is a probability, there seems to be a natural tendency for a scientist, at least one who is not a trained statistician, to interpret the value … as being closely related to, if not identical to, the probability that the hypothesis *H* is true. It is emphasized, however by statisticians applying a procedure of this type that the tail area … calculated from the sample is distinct from the posterior probability that *H* is true. In other words, although a tail area smaller than 0.01 may be obtained from a given sample, this result does not imply that the posterior probability of H is smaller than 0.01, or that the odds against H being true are 100 to 1 (or 99 to 1)….the weight of evidence against the hypothesis H that is implied by a small tail area … depends on the model and the assumptions that the statistician is willing to adopt.

DeGroot did find a set of specific assumptions about the alternative hypothesis under which the traditional *p*-value is a reasonable approximation to the true probability of the null. However, in a more general setting Dickey (1977) asked "Is the Tail Area Useful as an Approximate [Probability of the Null]?", and concluded that the answer is negative. In what follows we shall see that the traditional *p*-value (the tail value) can either under-estimate or over-estimate the true probability of the null. (We shall even see one interesting case in which the *p*-value is correct!) In general, it simply isn't possible to know what a test statistic tells you about the null without considering (a) what the statistic tells you under a given parameter value within the alternative and (b) how we weight the possible parameter values included in the alternative.

4. A Simple Coin-Flipping Example

In this section I work through an example in which an investigator observes a number of coin flips and wishes to find the probability that the coin is fair. This simple problem lets us postpone some complications until later sections. I begin with an example where the alternative is a single point and derive the probability of the null. Next, I show the relationship between the correctly computed probability of the null and the traditional *p*-value. The two are related, but are often very different quantitatively. Once one has computed the probability of the null, one can decide on a decision rule. For example, the investigator might choose to accept the null if its probability is above 95 percent, to accept the alternative if its probability is above 95 percent, or to remain undecided if both probabilities are below 95 percent. I show that, unlike the situation with traditional hypothesis testing, use of Bayes theorem makes it easy to distinguish between accepting the null and having insufficient evidence to reject the null.

The use of priors is sometimes derided as "unscientific," because it injects a non-data based element into inference. In the next subsection I show that the decision rules used in classical hypothesis testing also involve a prior. The prior simply is implicit in the mathematics of Bayes theorem instead of being stated by the investigator. The implicit prior is typically not at all neutral with respect to the null and alternative hypothesis. I use the coin tossing example to illustrate the relation among the implicit prior, the empirical *p*-value, and the investigator's decision rule.

All this is done with a point alternative. In the next to last subsection I consider what difference is made by choosing different values of the probability of the coin landing on heads. This then leads into the final subsection, where I introduce a continuous alternative.

4.1 Choosing Between Point Null and Point Alternative Hypotheses

Hypotheses commonly pick out one single value (the null, $\theta = \theta_0$) from a continuous set of possibilities (the alternative, $\theta \neq \theta_0$). In thinking through the logic of statistical choice it's easier to begin with the null and the alternative each assuming a discrete value ($\theta = \theta_0$ versus $\theta = \theta_A$), holding off for a bit on the more common continuous parameter space. I begin with a simple coin flipping example to see (a) how to properly compute p_{H_0}, (b) just how wrong we might be if we make a choice based on the *p*-value instead,[2] and (c) what assumptions are implied by basing decisions on statistical significance.

Suppose we observe n coin tosses where the probability of a head is θ and wish to choose between the null of a fair coin $H_0: \theta = \theta_0 = ½$ versus the alternative that the probability of a

[2] Checking statistical significance, calculating confidence intervals, and showing *p*-values amount to the same thing here. I frame the discussion in terms of *p*-values since this is the traditional measure of the statistical strength of significance.

head is $\theta = \theta_A$. The number of heads is distributed binomial $B(n, \theta)$. Like most common econometric estimators, for large n the distribution of the sample mean $\hat{\theta}$ is approximately normal, $\hat{\theta} \overset{A}{\sim} N(\theta, \theta(1-\theta)/n)$.

To make the example more concrete suppose the alternative is that the probability of a head is $\theta_A = 4/5$ and that after 32 tosses 21 land heads-up. Since the alternative specifies $\theta_0 < \theta_A$, the classical approach calls for a one-tailed test. The usual *t*-statistic is $(\hat{\theta} - \theta_0)/\sqrt{\frac{\theta_0(1-\theta_0)}{n}} = 1.77$. The normal approximation gives the *p*-value 0.039. (The *p*-values implied by the *t*-distribution and the binomial distribution are 0.043 and 0.025, respectively.) Since the *p*-value is well below the usual five percent level, an investigator might feel comfortable rejecting the null. Indeed, the point estimate $\hat{\theta} = 0.656$ is slightly closer to the alternative than to the null.

In fact, the weight of the evidence modestly favors the null hypothesis over the alternative. Using the exact binomial results, the evidence says that $\Pr(\hat{\theta}|H_0) = 0.030$ compared to $\Pr(\hat{\theta}|H_A) = 0.024$. The Bayes factor in equation (2) gives odds in favor of the null just over 6 to 5. Equivalently, equation (4) gives $p_{H_0} = 0.55$. The evidence against the null hypothesis is quite strong (a small *p*-value), but since the evidence against the alternative is even stronger the data favors the null hypothesis. Thus in this example, recognizing that one needs to account for the behavior of the test statistic under the alternative reverses the conclusion about which hypothesis is more likely.

Unsurprisingly, just as one can construct examples where we see a significant "rejection" of the null even though the data says the null is more likely than the alternative, a reversed

example where the *t*-statistic is insignificant even though the data favors the alternative is also easy to construct. Suppose the alternative were $\theta_A = .55$ and 17 of 32 tosses landed heads-up. The exact *p*-value from the binomial is 0.30 (0.36 from the normal approximation), which is not significant at any of the usual standards. Nonetheless, the evidence slightly favors the alternative, with $p_{H_A} = 0.51$.

4.2 The Relation Between the Probability of the Null and the Traditional *p*-Value

While everyone understands that a *p*-value is not the probability of the null hypothesis, and that 1 minus the *p*-value is not the probability of the alternative, we often act as if the *p*-value is at least a useful guide to the likelihood that the null is true. Figure 1 illustrates just how wrong this can be. To draw Figure 1, I take advantage of the simple nature of the coin tossing example. Because the example has only one parameter, there are one-to-one and onto relations between $\hat{\theta}$ and the *p*-value and between $\hat{\theta}$ and p_{H_0}, which gives the one-to-one and onto relation between the *p*-value and p_{H_0}. (Both *p*-values and p_{H_0} are computed from the binomial distribution.) The former relation is given by

$$\begin{aligned} p_{value} &= 1 - F_B\left(n\hat{\theta}, n, \theta_0\right) \\ \hat{\theta} &= F_B^{-1}(1 - p_v, n, \theta_0)/n \end{aligned} \tag{9}$$

The three curves in Figure 1 show the relation for three different sample sizes with our $n = 32$ example in the middle and the $\hat{\theta} = 21/32$ point marked with the square.

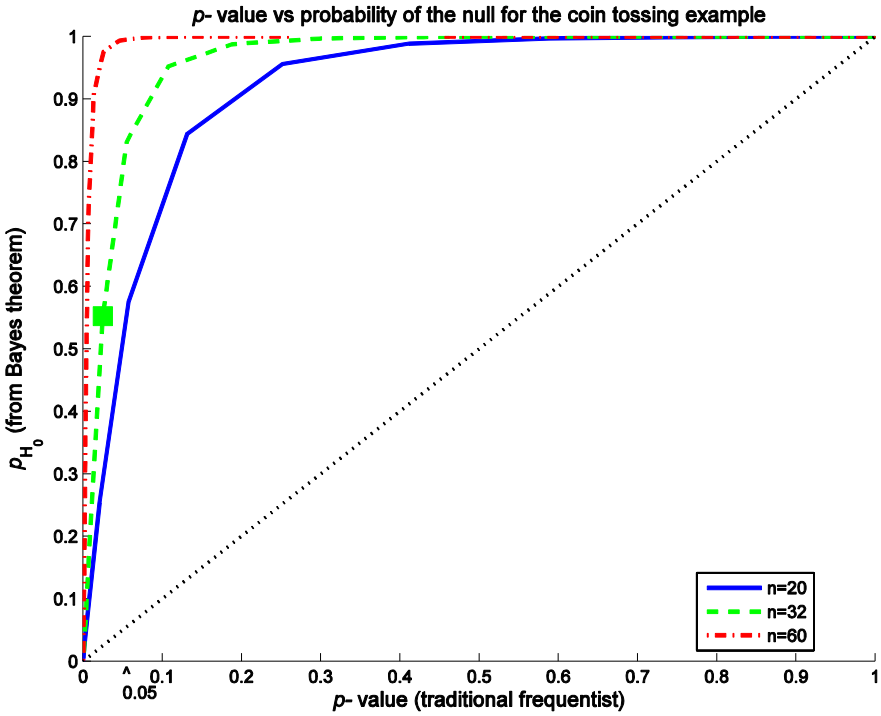

Figure 1

What is true is that the higher the *p*-value the higher the probability of the null. A zero *p*-value implies $p_{H_0} = 0$ and a *p*-value of 1.0 implies $p_{H_0} = 1$. Past that, however, the correct probability for choosing between hypotheses is typically very, very different from the *p*-value, as can be seen by the distance between the curves and the 45° line. In general, the probability of the null is far greater than the *p*-value. The caret on the horizontal axis marks the usual five percent significance level. For test results to the left of the caret, the usual procedures reject the null in favor of the alternative. But results in this region are often associated with the null being more likely than not, $p_{H_0} > 0.5$ and are often associated with very strong evidence in favor of the null against the alternative. For example, for $n = 60$ a *p*-value of 0.046 implies the *null* is almost certainly true, $p_{H_0} = 0.994$, while the probability of the alternative is negligible,

$p_{H_A} = 0.006$. So thinking of the *p*-value as even a very rough approximation for choosing between hypotheses is a very bad idea.

Many approximations in econometrics become more useful as the sample size increases. Using the *p*-value as a guide to the probability of the null is not one of them. Note in Figure 1 that higher values of n yield curves further from the 45° line. This is a general result to which we return below.

In summary the central lesson is: If you start off indifferent between the null and alternative hypothesis, then using statistical significance as a guide as to *which* hypothesis is more likely can be wrong and thinking of the *p*-value as a guide as to *how much* faith you should put in one hypothesis compared to the other can be very wrong.

4.3 Choosing the null, the alternative, or remaining undecided

The classical approach to hypothesis testing instructs that if there is very strong evidence against the null, e.g. the *p*-value is less than 0.05, we should reject the null in favor of the alternative. In principle, lacking strong evidence we fail to reach a conclusion; although in practice investigators quite often act as if failing to reject means we should accept the null. Using p_{H_0} lets us treat the two hypotheses symmetrically and allows for "don't know" to be the result of a specified decision rule. Suppose we wish to choose the null when $p_{H_0} > 0.95$, choose the alternative when $p_{H_A} > 0.95$, and otherwise report results as being indeterminate. Figure 2 shows the relation between the range for p_{H_0} and the corresponding *p*-values for the coin toss example. With the strong standard implied by requiring $p_{H_0} > 0.95$ to choose between hypotheses, the investigator should reject the null in favor of the alternative (in this

example) only for *p*-values under 0.003—a value very different from the usual 0.05 standard. (A *p*-value of 0.05 corresponds roughly to an 80 percent probability of the null.) The investigator should accept the null for *p*-values over 0.11. For *p*-values between 0.003 and 0.11, the evidence is inconclusive when measured against a 95 probability standard. So in this example, a *p*-value that would classically be considered weak evidence against the null is in fact fairly strong evidence in favor of the null.

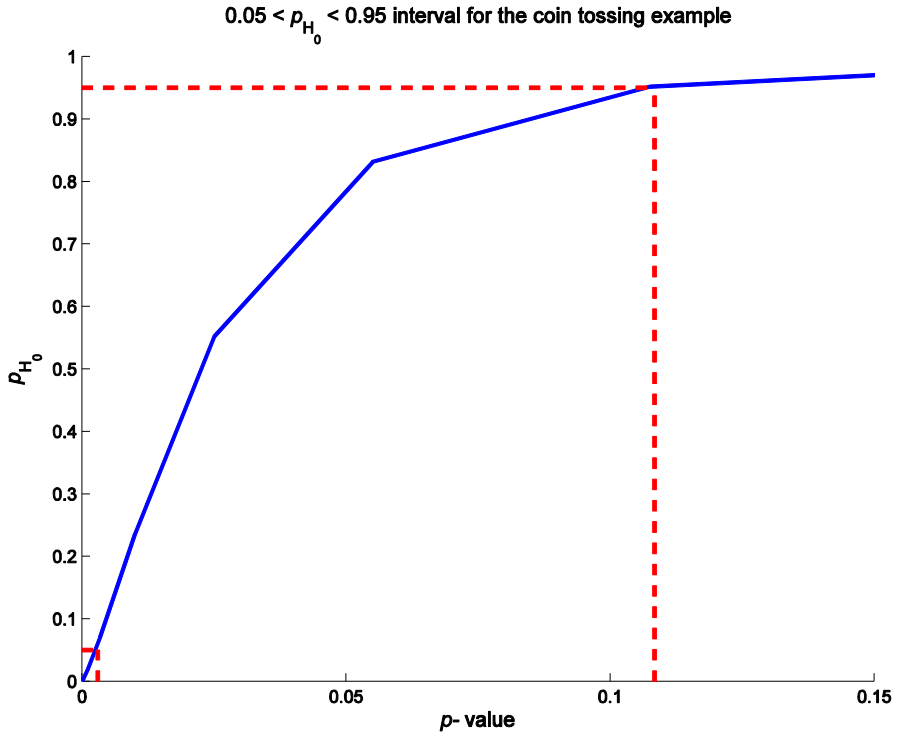

Figure 2

4.4 Implicit priors

One objection to the use of Bayes theorem is that it requires the specification of a prior, $\pi(H_0)$. One response to the objection is that by specifying $\pi(H_0) = 0.5$ we compute Bayes factors that tell us the marginal contribution of the data to what we know about the hypotheses. A somewhat less conciliatory response is that Bayes theorem doesn't go away if the investigator

fails to specify an explicit prior. A decision rule based on the significance level of a test statistic implies a prior through Bayes theorem. The implicit prior is often far from treating the hypotheses with an even hand.

For a given *p*-value in a particular model we can back out the implicit prior that the traditional econometrician is acting *as if* he held. As an example, suppose that the econometrician wants to choose the alternative hypothesis only when the probability of H_A is greater than some specified value q, say $q = 0.95$ or even $q = 0.50$. Suppose the econometrician in the usual way chooses the alternative hypothesis when his test is significant at the preset level. In our coin toss example, we can turn equation (3) on its head to back out an implied level of $\pi(H_A)$. Set the left hand side of equation (3) equal to q, find the value of $\Pr(\hat{\theta}|H_0)$ consistent with the preset *p*-value, and then solve for the minimum value, as in

$$q = 1 - \frac{\Pr(\hat{\theta}|H_0) \cdot (1 - \pi(H_A))}{\Pr(\hat{\theta}|H_0) \cdot (1 - \pi(H_A)) + \Pr(\hat{\theta}|H_A) \cdot \pi(H_A)}$$
$$\pi_{min}(H_A) = \frac{q \cdot \Pr(\hat{\theta}|H_0)}{q \cdot \Pr(\hat{\theta}|H_0) + (1 - q) \cdot \Pr(\hat{\theta}|H_A)} \quad (10)$$

Figure 3 shows the relation between the implicit prior and *p*-values for the coin toss example. The usual econometric procedure is to require significant evidence before rejecting the null in favor of the alternative. We might interpret "significant evidence" as deciding to pick the alternative only when $p_{H_A} > q = 0.95$. Suppose the econometrician requires a "very significant" test result, say the decision rule is to reject the null in favor of the alternative when the *p*-value is 0.01 or lower. For the coin toss example it turns out that this decision rule means the econometrician has an implicit prior weighting in favor of the alternative of 85 percent or more.

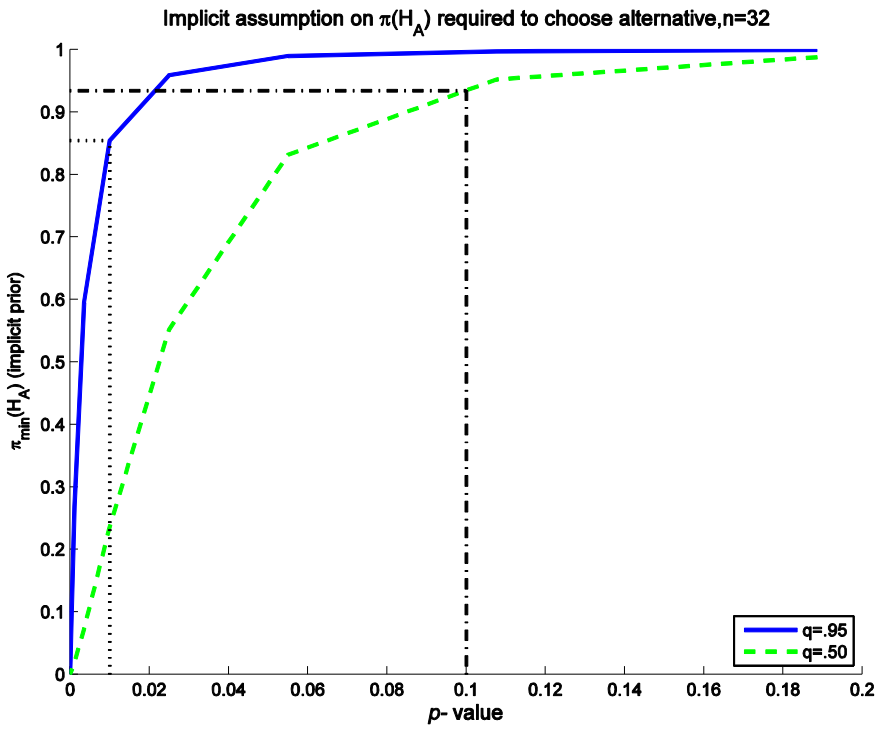

Figure 3

Suppose we were willing to accept much weaker evidence, for example if we were willing to accept the alternative whenever it was more likely, $q = 0.50$, and our decision rule was to reject the null in favor of the alternative when a test was "weakly significant," say a *p*-value below 10 percent. In this case the implicit prior favors the alternative with $\pi_{min}(H_A) > 0.93$.

One of the arguments for using classical testing rather than considering Bayes theorem is that by avoiding specifying a prior we take a neutral approach, a more "scientific" approach, to the data. By considering Bayes theorem we see that all the usual approach does is use an implicit prior, rather than make the prior explicit. We usually think that our standards for significance are chosen precisely to point in the direction of the null unless we have strong evidence to the contrary.[3] But as this example illustrates, our usual standards do not

[3] Thanks to Gene Savin (private communication) for pointing this out.

accomplish that goal. In other words, in this example the *p*-values we usually regard as providing strong evidence against the null and in favor of the alternative do not in fact provide such evidence unless the econometrician already leaned strongly toward the alternative.

4.5 The importance of the alternative

Classical hypothesis testing only explicitly considers the performance of a test under the null, although informal considerations of power obviously depend on what values make up the alternative. Equations (2) through (8) make clear that the value of the alternative is also critical.

I illustrate the importance of the value of the alternative by looking at how our conclusions about the null are affected by the choice of the point alternative. This then sets up the following discussion of a continuous alternative. Figure 4 graphs p_{H_0} for $\theta_A \in \{0.5,1.0\}$ for the coin toss example. The figure also shows the power of a one-tailed, five percent test with the critical value given by $F_B^{-1}(.95, n, \theta_0)/n$. (I ignore the small error in computing size and power due to the fact that the binomial is a discrete distribution, since it doesn't matter for the illustration.) At the left side of Figure 4, where the power of the classical test is very low, we see that the proper conclusion is that we cannot distinguish the null from the alternative. If the relevant alternative is indistinguishable from the null, $\theta_A = \theta_0 = 0.5$, then *unlike in classical tests,* the only reasonable conclusion is that the null and alternative are of equal likelihood. Classical tests fail to take into account that when the null and the alternative are essentially the same, the power to reject the alternative equals the size of the test. Equally, if the alternative were $\theta_A = 1$, where the power is very high, we ought to conclude that the null hypothesis must be true if a single flip fell tails up. In other words, what we conclude about the null depends critically on what we think the relevant alternative is.

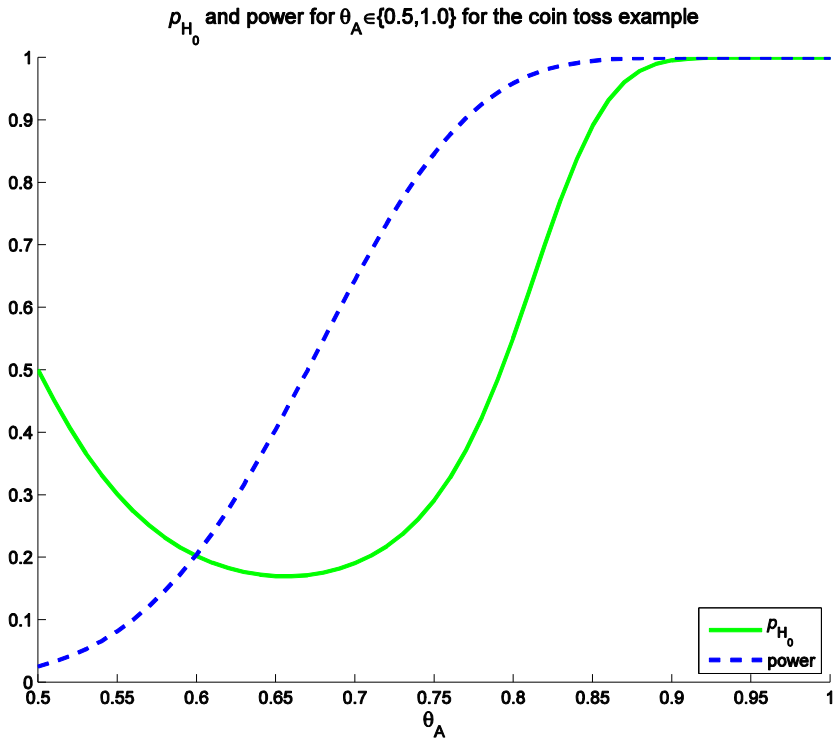

Figure 4

In terms of intuition, this is pretty much the lesson. However In most empirical applications the parameter of interest is allowed to be continuous, as opposed to being limited to two values as in the example above. The remainder of the paper takes up this practical issue; first, to build intuition, in a continuation of the coin toss example and then in the more relevant case of a normally distributed estimator.

4.6 A Continuous Alternative for the Coin-Toss Example

Evaluating probabilities was easy in the coin toss example because the null and the alternative are each a single point. Most often, the interesting choice is between $\theta = \theta_0$ and θ equals something else, $\theta \neq \theta_0$. While the essentials don't change from the two-value example, the details get a little more complicated. In the next section, I consider the details for the most common situation: estimators which are approximately normally distributed. I continue here

with the coin toss example because it is easier in some ways. The central issue is how to think of the alternative hypothesis when the alternative is "θ equals something else," rather than just a single point.

To see the importance of considering the alternative, consider Figure 4 again. When the alternative consists of a specified point computing p_{H_0} is straightforward, but when the value of θ_A ranges between 0.5 and 1., the value of p_{H_0} we find for the example data ranges between 0.17 and 1.0. If the alternative is to include more than one value for θ_A, we need (loosely speaking) to weight the outcomes in Figure 4 according to the weights we assign to each value of θ_A. In the earlier toy model, the alternative hypothesis specified a single point for θ_A. Now under the alternative we need to consider a range of values for θ_A.

More formally, we replace conditioning on H_A with a probability distribution over θ_A, as in $\pi(\theta_A, H_A) = \pi(\theta_A|H_A)\pi(H_A)$. Thus the full expression for the probability of the alternative becomes

$$\begin{aligned} p_{H_A} &= \Pr\left(\hat{\theta}\middle|\theta_A, H_A\right) \cdot \frac{\pi(\theta_A, H_A)}{\Pr\left(\hat{\theta}\right)} \\ &= \int_{-\infty}^{\infty} \frac{\Pr\left(\hat{\theta}\middle|\theta_A\right)\pi(\theta_A|H_A)\pi(H_A)}{\Pr\left(\hat{\theta}\right)} d\theta_A \\ &= \int_{-\infty}^{\infty} \Pr\left(\hat{\theta}\middle|\theta_A\right)\pi(\theta_A|H_A) d\theta_A \times \frac{\pi(H_A)}{\Pr\left(\hat{\theta}\right)} \end{aligned} \tag{11}$$

Note that we break the prior for the alternative into a conditional density and a discrete weight, $\pi(H_A)$. This lets us proceed as before with computation of the posterior odds ratio, Bayes factor and p_{H_0},

$$PO_{0A} = \frac{p_{H_0}}{p_{H_A}} = \frac{\Pr(\hat{\theta}|H_0)}{\int_{-\infty}^{\infty} \Pr(\hat{\theta}|\theta_A)\pi(\theta_A|H_A)d\theta_A} \times \frac{\pi(H_0)}{\pi(H_A)}$$

$$p_{H_0} = \frac{\Pr(\hat{\theta}|H_0) \cdot \pi(H_0)}{\Pr(\hat{\theta}|H_0) \cdot \pi(H_0) + \int_{-\infty}^{\infty} \Pr(\hat{\theta}|\theta_A)\pi(\theta_A|H_A)d\theta_A \cdot (1 - \pi(H_0))} \quad (12)$$

$$p_{H_0} = \frac{\Pr(\hat{\theta}|H_0)}{\Pr(\hat{\theta}|H_0) + \int_{-\infty}^{\infty} \Pr(\hat{\theta}|\theta_A)\pi(\theta_A|H_A)d\theta_A}, \text{ if } \pi(H_0) = .5$$

For our toy model, the parameter of interest is the probability of a coin landing heads-up, so we might assume a prior for that probability that is uniform between zero and one. That gives a conditional density $\pi(\theta_A|H_A) = 1, \theta_A \in [0,1]$ and zero elsewhere. The required integral in equations (11) and (12) becomes $\int_0^1 f_B(n\hat{\theta}, n, \theta_A)d\theta_A$, which is easily computed numerically.

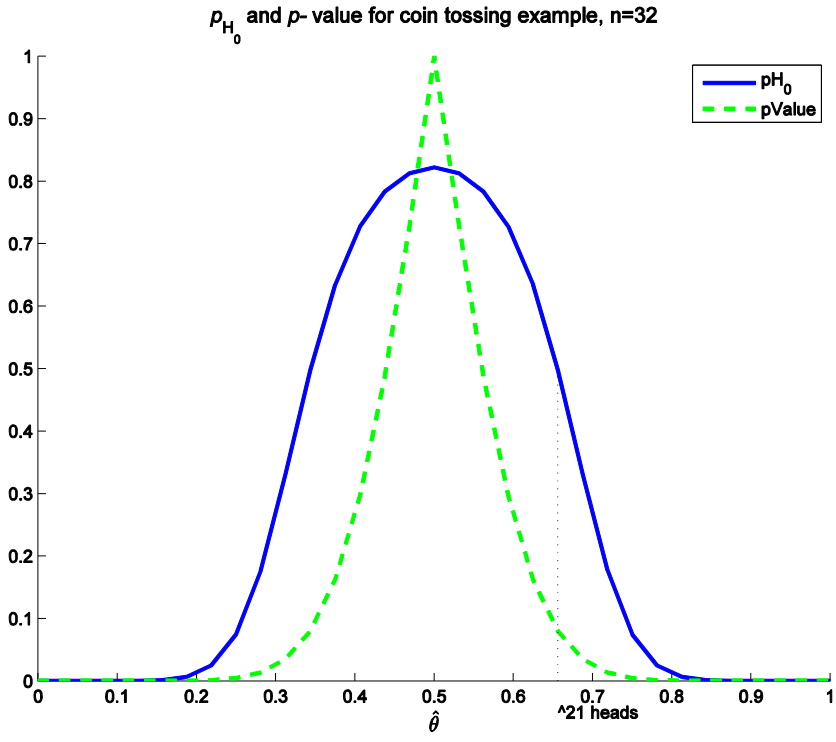

Figure 5

Figure 5 shows the values of p_{H_0} as well as two-sided p-values for the null $\theta = 0.5$ for the coin tossing example for all possible outcomes (values of $\hat{\theta}$) for 32 tosses. Note that p_{H_0} may be

either higher or lower than the traditional *p*-value. And the difference may be substantial. For the earlier example of 21 heads the *p*-value, given by $F_B\left(n\theta_0 - (n\hat{\theta} - n\theta_0), n, \theta_0\right) + \left(1 - F_B\left(n\theta_0 + \left(n\hat{\theta} - n\theta_0\right), n, \theta_0\right)\right)$, is 0.08, suggesting that the null of a fair coin is unlikely. But the correct calculation is that the probability of a fair coin is 50 percent. Thus we see that the basic point that p_{H_0} and the *p*-value may be quite far apart continues to be true when the alternative is continuous rather than a single point.

5. Regression estimates

Testing whether a coin is fair does not have a large share of the market for econometric applications. I turn now to hypotheses about a regression parameter, or in fact about any parameter which is approximately normally distributed. All the basic lessons in the toy example continue to be true, and some new issues arise.

The standard, classical approach to hypothesis testing nests a point null within a more general alternative. I focus in this section on nested tests and, for the moment, on point nulls. (Non-sharp nulls are discussed below.) As a practical matter this means considering two-sided alternative hypotheses centered around the null, which I think best represents the implicit alternative most used by classical econometricians. In what follows I consider two ways to specify an explicit alternative, one based on the uniform distribution and one based on the normal distribution. In each case, I give an easy-to-apply formula for p_{H_0} that adjusts the standard *t*-statistic taking into account the width of the interval around the null that the investigator wishes to consider. For a uniform alternative, I show below that the probability of the null can be approximated by

$$p_{H_0} \approx \frac{\phi(t)}{\phi(t) + \left[c/\sigma_{\hat{\theta}}\right]^{-1}} \tag{13}$$

where t is the usual *t*-statistic , c is the width of the interval around θ_0, $\sigma_{\hat{\theta}}$ is the standard error of the estimated coefficient, and $\phi(\cdot)$ is the standard normal density.

For a normal alternative where the standard deviation of the alternative is σ_A, I show below that the probability of the null can be approximated by

$$p_{H_0} \approx \frac{\phi(t)}{\phi(t) + \left[\sigma_A/\sigma_{\hat{\theta}}\right]^{-1}\phi(0)} \tag{14}$$

In both cases, the approximation is excellent for a very diffuse alternative (c or σ_A large, respectively). Exact expressions follow.

A key consideration is how to compute the integral $\int_{-\infty}^{\infty} \Pr(\hat{\theta}|\theta_A)\pi(\theta_A|H_A)d\theta_A$ in equation (11). Because this is a central issue in Bayesian econometrics, Bayesians have introduced a number of numerical techniques for this purpose. Here, I limit attention to uniform and normal distributions for $\pi(\theta_A|H_A)$, as these are both familiar and lead to analytic solutions. The discussions of the uniform and normal priors include sufficient repetition that the reader interested in only one can skip the other discussion. Although the details differ, for each case I derive the solution for p_{H_0}, consider the effects of the width of the prior on the solution for p_{H_0}, and derive the width of the prior that is implicit in the usual use of hypothesis tests. Finally, I examine the interesting case of one-tailed tests, in which the traditional *p*-values turn out to lead to correct inference about the probability of the null.

5.1 Uniform Prior for the Alternative Hypothesis

Consider the classical regression model and the hypothesis that a particular regression coefficient equals zero versus the alternative that it doesn't. With enough assumptions, $\hat{\theta}$ is distributed normally around the true parameter value with standard deviation $\sigma_{\hat{\theta}}$. That leaves the question of setting up the alternative for the prior in a hierarchical fashion. As before, we might view $\pi(H_0) = \frac{1}{2}$ as arguably neutral. One attractive choice for the prior under the alternative is uniform, $\theta|H_A \sim U\left[\theta_0 - \frac{c}{2}, \theta_0 + \frac{c}{2}\right]$, (so $\pi(\theta|H_A)$ is a constant $1/c$ over the specified interval and 0 elsewhere), with the range chosen so that most everyone is comfortable that all reasonable values of θ are included, while within that range no one value of θ is favored over any other. (For reasons discussed below, consistency requires that the range includes the true value of θ.) Using the uniform makes it easy to evaluate the required integral since $\pi(\theta|H_A)$ is $1/c$ inside the limits and zero elsewhere. For example, $t = 1.96$ corresponds to the p-value $= 0.05$. However if we specified a prior that ex post turned out to be three standard deviations on either side of the null, we should report $p_{H_0} = 0.26$ or equivalently $p_{H_A} = 0.74$. This suggests that a "significant t-" is evidence in favor of the alternative, but rather weaker evidence than is usually thought. Even this conclusion is quite sensitive to the choice of c.

Turn now to the derivation of equation (13). Continuing with the uniform distribution and normal $\hat{\theta}$, we can write the integral we need as

$$\int \Pr(\hat{\theta}|\theta)\pi(\theta|H_A)d\theta = \frac{1}{c}\int_{\theta_0-\frac{c}{2}}^{\theta_0+\frac{c}{2}} \Pr(\hat{\theta}|\theta)d\theta \tag{15}$$

Note that the integral is taken with respect to the conditioning variable θ rather than the random outcome $\hat{\theta}$.

Conveniently, the mean and random variable arguments are interchangeable in the normal density for $\hat{\theta}$. If, as is approximately true in the regression case, $\hat{\theta}$ is distributed $N(\theta_0, \sigma_{\hat{\theta}}^2)$, then equation (15) can be re-written in terms of the standard normal CDF, $\Phi(\cdot)$.

$$\begin{aligned}\int_l^u \Pr(\hat{\theta}|\theta)d\theta &= \int_l^u \left(2\pi\sigma_{\hat{\theta}}^2\right)^{-\frac{1}{2}} \exp\left\{\frac{1}{2\sigma_{\hat{\theta}}^2}\left(\hat{\theta}-\theta\right)^2\right\}d\theta \\ &= \int_l^u \left(2\pi\sigma_{\hat{\theta}}^2\right)^{-\frac{1}{2}} \exp\left\{\frac{1}{2\sigma_{\hat{\theta}}^2}\left(\hat{\theta}-\theta\right)^2\right\}d\hat{\theta}\end{aligned} \tag{16}$$

In other words, the required integral is the same as if we had $\theta \sim N(\hat{\theta}, \sigma_{\hat{\theta}}^2)$. Thus the integral in equation (15) becomes

$$\int \Pr(\hat{\theta}|\theta)\pi(\theta|H_A)d\theta = \frac{1}{c}\left[\Phi\left(\frac{\theta_0 + {}^{c}/_{2} - \hat{\theta}}{\sigma_{\hat{\theta}}}\right) - \Phi\left(\frac{\theta_0 - {}^{c}/_{2} - \hat{\theta}}{\sigma_{\hat{\theta}}}\right)\right] \tag{17}$$

Define $t = \frac{\hat{\theta}-\theta_0}{\sigma_{\hat{\theta}}}$, as usual. Note the cdfs in equation (17) are $\Phi(-t+\kappa)$ and $\Phi(-t-\kappa)$ for $\kappa \equiv c/2\sigma_{\hat{\theta}}$. Since $\Phi(-t+\kappa) - \Phi(-t-\kappa) = \Phi(t+\kappa) - \Phi(t-\kappa)$ for any constant κ, and since $\Pr(\hat{\theta}|H_0) = \sigma_{\hat{\theta}}^{-1}\phi\left(\frac{\hat{\theta}-\theta_0}{\sigma_{\hat{\theta}}}\right)$, equation (12) becomes

$$p_{H_0} = \frac{\sigma_{\hat{\theta}}^{-1}\phi(t)\cdot\pi(H_0)}{\sigma_{\hat{\theta}}^{-1}\phi(t)\cdot\pi(H_0) + \frac{1}{c}\left[\Phi\left(t+\frac{c}{2\sigma_{\hat{\theta}}}\right) - \Phi\left(t-\frac{c}{2\sigma_{\hat{\theta}}}\right)\right]\times\left(1-\pi(H_0)\right)} \tag{18}$$

or if we continue with $\pi_{H_0} = \pi_{H_A} = 1/2$,

$$p_{H_0} = \frac{\phi(t)}{\phi(t) + \frac{\sigma_{\hat{\theta}}}{c}\left[\Phi\left(t + \frac{c}{2\sigma_{\hat{\theta}}}\right) - \Phi\left(t - \frac{c}{2\sigma_{\hat{\theta}}}\right)\right]} \tag{19}$$

Note that if the range chosen for the alternative is large relative to the standard error, $c \gg \sigma_{\hat{\theta}}$, then the term in square brackets in equation (18) is approximately 1.0 and the equation simplifies to

$$p_{H_0} = \frac{\phi(t) \cdot \pi(H_0)}{\phi(t) \cdot \pi(H_0) + \frac{\sigma_{\hat{\theta}}}{c} \times \left(1 - \pi(H_0)\right)} \tag{20}$$

which, with the addition of $\pi(H_0) = 1/2$, gives equation (13) above. In the example above with $t = 1.96$ and $\frac{c}{2\sigma_{\hat{\theta}}} = 3$, the exact (equation (19)) value is $p_{H_0} = .29$ compared to the $p_{H_0} = .26$ approximation in equation (13). For $t = 0$, the approximation is correct to three decimal places. The approximation improves for large c and for small t.

In an applied problem, the investigator presumably specifies c as the ex ante region of interest under the alternative. However, it is interesting to understand the generic effects of the width of c. The probability of the null is non-monotonic in a somewhat non-intuitive way. Figure 6 gives the probability of the null according to equation (19) for various values of t (with associated two-tailed *p*-values) and $c/2\sigma_{\hat{\theta}}$. Note that the proper calculation of p_{H_O} is typically very different from the *p*-value. In general, the *p*-value overstates the evidence against the null by a great deal. The non-monotonicity arises through the second term in the denominator of equation (19), where there is a tradeoff between the area covered by the alternative and the height of the density of the alternative prior.

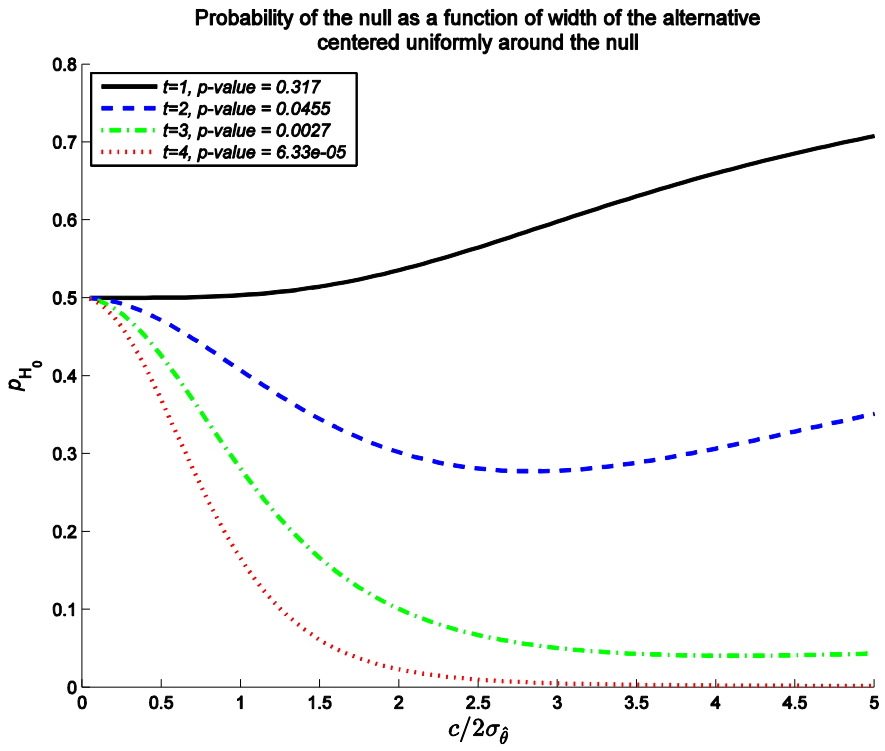

Figure 6

The choice of width of the alternative matters. Unsurprisingly, if the only relevant alternative values are very close to the null, then the data cannot distinguish between the two hypotheses as power is close to size. Formally, $\lim_{c\to 0} p_{H_0} = 1/2$ in equation (19) (by l'Hôpital's rule).

Perhaps more interestingly, $\lim_{c\to\infty} p_{H_0} = 1$. In other words, if the investigator wishes to take a neutral attitude towards all possible alternative values, he needn't even look at the data because the null hypothesis will always be chosen. This is a version of the Bayesian's "Lindley paradox," discussed further below. The critical point is that if one is to choose between hypotheses, one *can't* ignore the specification of the alternative.

Why is this important?

A traditional argument against using Bayes theorem has been that it is difficult to specify priors, or that it is unscientific for the result to depend on priors. In the coin tossing example with a point alternative, the 50/50 prior can be argued to be neutral. In the more common continuous case, there is not a similarly neutral statement. What is important is that the frequentist procedure is also non-neutral with respect to the area of the alternative. For a given frequentist decision rule we can invert equation (19) to find the implicit value of $c/\sigma_{\hat{\theta}}$. Suppose that the frequentist decision rule is to prefer the alternative whenever p_{H_A} is greater than some value q. Figure 7 graphs the maximum width of the uniform prior against *p*-values for both $q = 0.5$ and $q = 0.95$.

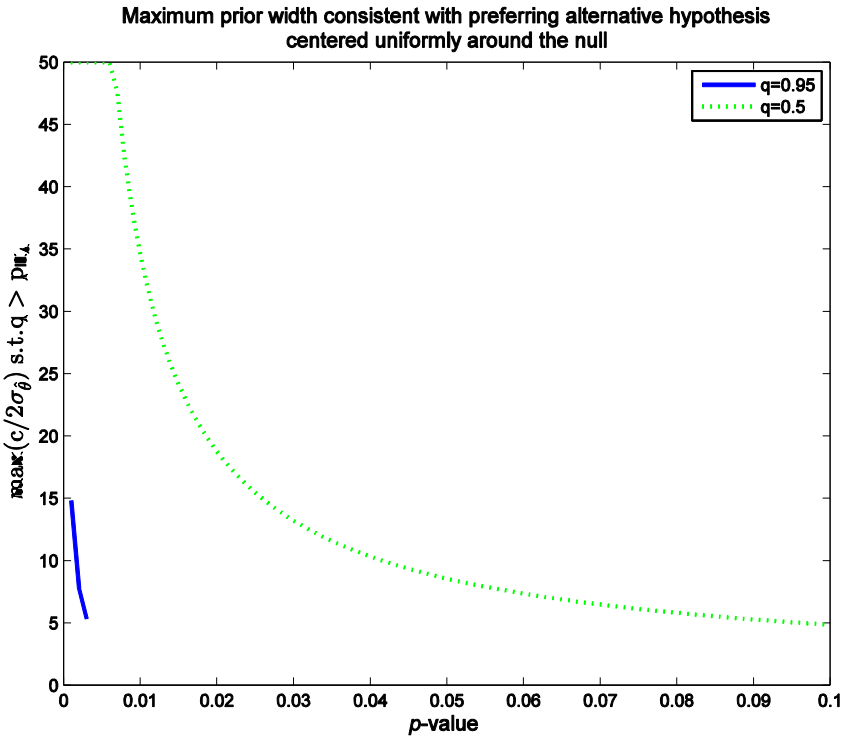

Figure 7

Suppose an econometrician wants to choose the alternative whenever it is the more likely outcome ($q = 0.5$), starting from a neutral position ($\pi_{H_0} = \pi_{H_A}$). Figure 7 tells us that the usual

frequentist criteria likely leads to the right decision, in the following sense. Strong evidence against the null, a 1 percent *p*-value, points toward the alternative so long as the implicit alternative is no more than 35 standard errors wide. Even weak evidence, a 10 percent *p*-value, points toward the alternative for implicit priors as much as five standard errors wide.

Note, however, that while in this case the usual criteria leads to the right decision if the econometrician's standard is preponderance of the evidence, the same conclusion is not true if a stronger standard of evidence is required. Suppose the decision rule is to pick the alternative only when the gold standard $p_{H_A} > 0.95$ is met. What Figure 7 shows is that even what is usually thought to be very strong evidence against the null, 1 percent *p*-value, is inconsistent with choosing the alternative. There is *no* uniform alternative for which a "strongly significant" frequentist rejection of the null is correctly interpreted as strong evidence for the alternative. In fact, one can show numerically that the largest *p*-value consistent with $p_{H_A} > 0.95$ is 0.003.

5.2 Normal Prior For the Alternative Hypothesis

As an alternative to the uniform we might specify a normal distribution for the prior for θ, centering the prior at the null, $\theta \sim N(\theta_0, \sigma_A^2)$. For example, if $t = 1.96$ and we specify a prior where σ_A turns out to be three times the standard deviation of $\hat{\theta}$ we should report $p_{H_0} = 0.31$ or equivalently $p_{H_A} = 0.69$, rather than the p-value $= 0.05$. (It is worth remembering when comparing a normal prior to a uniform prior that the normal necessarily has fatter tails than the uniform.) This suggests that a "significant *t*-" is evidence in favor of the alternative, but rather weaker evidence than is usually thought. Even this conclusion is sensitive to the choice of σ_A.

The first step in analyzing the normal alternative is to multiply the probability of $\hat{\theta}$ under the alternative by the prior. Let the prior under the alternative, $\pi(\theta|H_A)$, be distributed $N(\theta_0, \sigma_A^2)$. Bayes theorem tells us that

$$p\left(\theta\middle|\hat{\theta}, H_A\right) = \frac{p\left(\hat{\theta}\middle|\theta\right) \cdot \pi(\theta|H_A)}{\int_{-\infty}^{\infty} p\left(\hat{\theta}\middle|\theta\right) \cdot \pi(\theta|H_A) d\theta} \tag{21}$$

Applying enough algebra to the product of two normal densities in the numerator,[4] one can show that $p\left(\theta\middle|\hat{\theta}, H_A\right)$ is a normal density given by

$$\begin{aligned} \theta|\hat{\theta}, H_A &\sim N\left(\tilde{\theta}, \tilde{\sigma}^2\right) \\ \tilde{\theta} &= \frac{[\sigma_A^2/\sigma_{\hat{\theta}}^2]\hat{\theta} + \theta_0}{1 + [\sigma_A^2/\sigma_{\hat{\theta}}^2]} \\ \tilde{\sigma}^2 &= \frac{\sigma_A^2}{1 + [\sigma_A^2/\sigma_{\hat{\theta}}^2]} \end{aligned} \tag{22}$$

Intuitively, as $\sigma_A \to 0$ the distribution of θ collapses around the null and as $\sigma_A \to \infty$ the distribution of θ is entirely determined by the data, i.e. the distribution equals $N\left(\hat{\theta}, \sigma_{\hat{\theta}}^2\right)$.

Bayesian procedures offer a short cut for deriving the Bayes factor when a point null is nested inside a more general alternative, as is the case here. The Bayes factor is given by the Savage-Dickey ratio[5] which is the ratio of $p\left(\theta\middle|\hat{\theta}, H_A\right)$ to $\pi(\theta|H_A)$, both evaluated at $\theta = \theta_0$. With a reminder that the density of a nonstandardized normal, $\hat{\mu} \sim N\left(\mu, \sigma_{\mu}^2\right)$ can be expressed in terms of the standard normal density, $\frac{1}{\sigma_\mu}\phi\left(\frac{\hat{\mu}-\mu}{\sigma_\mu}\right)$, the density of $\pi(\theta = \theta_0|H_0) = \left(\frac{1}{\sigma_A}\right)\phi\left(\frac{\theta_0-\theta_0}{\sigma_A}\right) = \phi(0)/\sigma_A$. Similarly, $p\left(\theta = \theta_0\middle|\hat{\theta}, H_A\right) = \frac{1}{\tilde{\sigma}}\phi\left(\frac{\theta_0-\tilde{\theta}}{\tilde{\sigma}}\right)$. The numerator in $\phi(\cdot)$ is usefully re-written using

[4] Poirier (1995), pp. 536ff.
[5] See Dickey (1971) or Koop (2003) 69ff.

$$\theta_0 - \tilde{\theta} = \theta_0 - \frac{\left[\sigma_A^2/\sigma_{\hat{\theta}}^2\right]\hat{\theta} + \theta_0}{1 + \left[\sigma_A^2/\sigma_{\hat{\theta}}^2\right]} = \frac{\left[\sigma_A^2/\sigma_{\hat{\theta}}^2\right]}{1 + \left[\sigma_A^2/\sigma_{\hat{\theta}}^2\right]}\left(\theta_0 - \hat{\theta}\right) = \frac{\tilde{\sigma}^2}{\sigma_{\hat{\theta}}^2}\left(\sigma_{\hat{\theta}} \times \frac{\theta_0 - \hat{\theta}}{\sigma_{\hat{\theta}}}\right)$$
$$\frac{\theta_0 - \tilde{\theta}}{\tilde{\sigma}} = -\frac{\tilde{\sigma}}{\sigma_{\hat{\theta}}} t \tag{23}$$

where, as before, t is the usual *t*-statistic. Thus using the Savage-Dickey ratio for the Bayes factor gives

$$BF_{0A} = \frac{\frac{1}{\tilde{\sigma}}\phi\left(\frac{\tilde{\sigma}}{\sigma_{\hat{\theta}}} t\right)}{\frac{1}{\sigma_A}\phi(0)} = \frac{\sqrt{1 + (\sigma_A/\sigma_{\hat{\theta}})^2}\,\phi\left(\sqrt{\frac{(\sigma_A/\sigma_{\hat{\theta}})^2}{1 + (\sigma_A/\sigma_{\hat{\theta}})^2}}\, t\right)}{\phi(0)} \tag{24}$$

which gives the probability of the null

$$p_{H_0} = \frac{\sqrt{1 + (\sigma_A/\sigma_{\hat{\theta}})^2}\,\phi\left(\sqrt{\frac{(\sigma_A/\sigma_{\hat{\theta}})^2}{1 + (\sigma_A/\sigma_{\hat{\theta}})^2}}\, t\right)}{\phi(0) + \sqrt{1 + (\sigma_A/\sigma_{\hat{\theta}})^2}\,\phi\left(\sqrt{\frac{(\sigma_A/\sigma_{\hat{\theta}})^2}{1 + (\sigma_A/\sigma_{\hat{\theta}})^2}}\, t\right)} \tag{25}$$

Note that if the range chosen for the alternative is large relative to the standard error, $\sigma_A \gg \sigma_{\hat{\theta}}$, then $\sqrt{1 + (\sigma_A/\sigma_{\hat{\theta}})^2} \approx \sigma_A/\sigma_{\hat{\theta}}$ and $\sqrt{\frac{(\sigma_A/\sigma_{\hat{\theta}})^2}{1+(\sigma_A/\sigma_{\hat{\theta}})^2}} \approx 1$, which gives equation (14) above. In the example above with $t = 1.96$ and $\sigma_A/\sigma_{\hat{\theta}} = 3$, the exact value is $p_{H_0} = .36$ compared to the $p_{H_0} = .31$ approximation. For $t = 0$, the approximation is correct to two decimal places. The approximation improves for large $\sigma_A/\sigma_{\hat{\theta}}$ and for small t.

In an applied problem, the investigator presumably specifies σ_A depending on the ex ante region of interest under the alternative. However, as was true for the width of the uniform prior, it is interesting to understand the generic effects of the width of σ_A. The probability of the null is non-monotonic in a somewhat non-intuitive way, as shown in Figure 8 which computes

the probability of the null according to equation (25) for various values of t (with associated two-tailed *p*-values) and $\sigma_A/\sigma_{\hat{\theta}}$. Note that the proper calculation of p_{H_O} is typically very different from the *p*-value. In general, the *p*-value overstates the evidence against the null by a great deal. The non-monotonicity arises through the numerator in the Bayes factor in equation (24), where higher $\sigma_A/\sigma_{\hat{\theta}}$ increases $\sqrt{1+(\sigma_A/\sigma_{\hat{\theta}})^2}$ and decreases $\phi(\cdot)$.

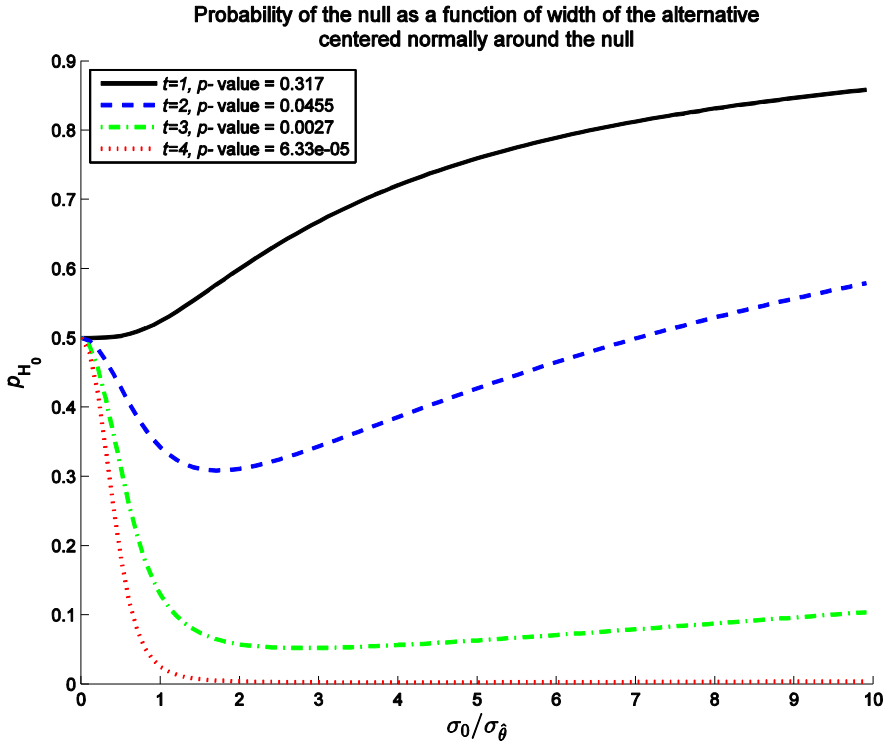

Figure 8

As before, the choice of width of the alternative matters. If the only relevant alternative values are very close to the null, then the data cannot distinguish between the two hypotheses as power is close to size. Formally, $\lim_{\sigma_A \to 0} p_{H_0} = 1/2$. Also, analogous to the result for the uniform prior, $\lim_{\sigma_A \to \infty} p_{H_0} = 1$.

Here too, the frequentist procedure is non-neutral with respect to the width of the alternative. For a given frequentist decision rule we can invert equation (25) to find the implicit

value of $\sigma_A/\sigma_{\hat{\theta}}$. Suppose that the frequentist decision rule is to prefer the alternative whenever p_{H_A} is greater than some value q. Figure 9 graphs the maximum width of the uniform prior against *p*-values for both $q = 0.5$ and $q = 0.95$. The vertical axis is scaled in units of the ratio of the prior standard deviation to the standard deviation of the estimated coefficient. The horizontal axis gives *p*-values.

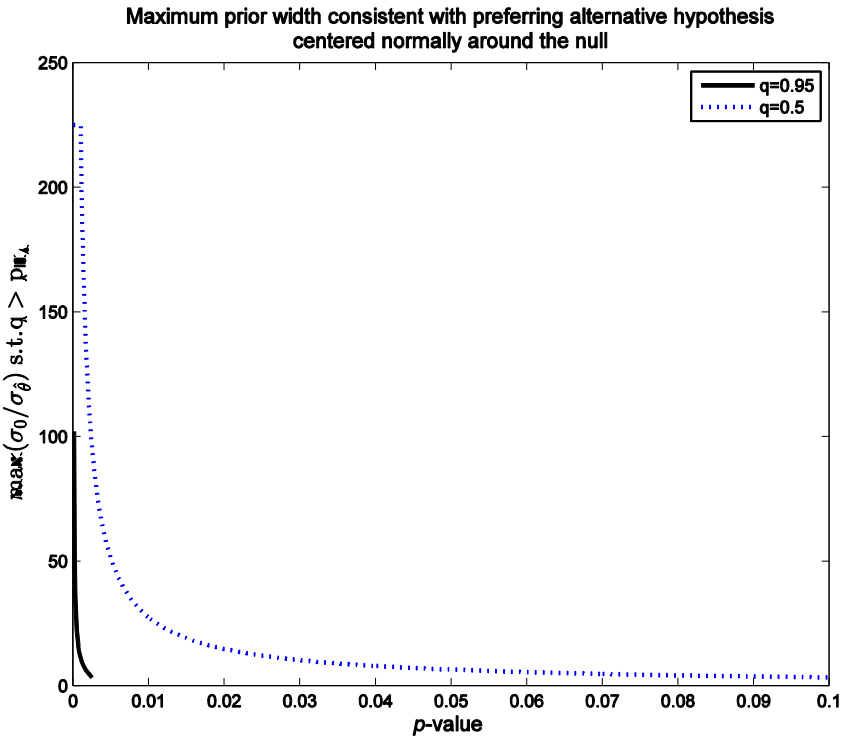

Figure 9

Suppose an econometrician wants to choose the alternative whenever it is the more likely outcome ($q = 0.5$), starting from a neutral position ($\pi_{H_0} = \pi_{H_A}$). Figure 9 tells us that the usual frequentist criteria likely leads to the right decision, in the following sense. Strong evidence against the null, a 1 percent *p*-value, points toward the alternative so long as the standard deviation of the implicit alternative is no more than 224 standard errors wide. Even weak

evidence, a 10 percent *p*-value, points toward the alternative for implicit priors as much as 3.3 standard errors wide.

As was true using a uniform alternative, while in this case the usual criteria leads to the right decision if the econometrician's standard is preponderance of the evidence, the same conclusion is not true if a stronger standard of evidence is required. Suppose the decision rule is to pick the alternative only when the gold standard $p_{H_A} > 0.95$ is met. What Figure 9 shows is that even what is usually thought to be very strong evidence against the null, 1 percent *p*-value, is inconsistent with choosing the alternative. There is *no* normal alternative for which a "strongly significant" frequentist rejection of the null is correctly interpreted as strong evidence for the alternative. In fact, the largest *p*-value consistent with $p_{H_A} > 0.95$ is 0.0025.

5.3 One-sided hypotheses

In the next section, we shall see that the choice of a very diffuse prior is generally problematic when comparing the usual point null to a broad alternative. It turns out that one-sided tests do not have a difficulty with such diffuse priors. In fact, the usual "incorrect" interpretation of the *p*-value as the probability of the null turns out to be correct. (See Casella and Berger (1987.) Since econometricians mostly conduct two-sided tests (for better or worse), this is in part a curiosum. However, looking at one-sided tests also helps understand why, as we see in the next section, problems develop when one hypothesis has a diffuse prior for a particular parameter when the other hypothesis does not.

A classical one-sided hypothesis might be specified by

$$\begin{aligned} H_0&: \theta > \theta_0 \\ H_A&: \theta < \theta_0 \end{aligned} \tag{26}$$

Continuing with the case $\hat{\theta} \sim N\left(\theta, \sigma_{\hat{\theta}}^2\right)$ and $t = \frac{\hat{\theta} - \theta_0}{\sigma_{\hat{\theta}}}$, the usual one-sided *p*-value is given by $\Phi(t)$.

A convenient prior for use in invoking Bayes law is to let θ be uniform for a distance c above θ_0 for the null and uniform for a distance c below θ_0 for the alternative, as in

$$\begin{aligned} \pi(\theta|H_0) &= \begin{cases} \frac{1}{c}, \theta_0 < \theta < \theta_0 + c \\ 0, otherwise \end{cases} \\ \pi(\theta|H_A) &= \begin{cases} \frac{1}{c}, \theta_0 > \theta > \theta_0 - c \\ 0, otherwise \end{cases} \end{aligned} \tag{27}$$

The marginal likelihoods are given by

$$\begin{aligned} p(y|H_0) &= \frac{1}{c}\left[\Phi\left(\frac{\theta_0 + c - \hat{\theta}}{\sigma_{\hat{\theta}}}\right) - \Phi\left(\frac{\theta_0 - \hat{\theta}}{\sigma_{\hat{\theta}}}\right)\right] \\ p(y|H_A) &= \frac{1}{c}\left[\Phi\left(\frac{\theta_0 - \hat{\theta}}{\sigma_{\hat{\theta}}}\right) - \Phi\left(\frac{\theta_0 - c - \hat{\theta}}{\sigma_{\hat{\theta}}}\right)\right] \end{aligned} \tag{28}$$

Notice that, unlike in the earlier case of two-sided tests, the term $1/c$ appears in both marginal likelihoods in a form that exactly cancels when forming the Bayes factor. A little substitution gives the Bayes factor,

$$BF = \frac{\Phi(t) - \Phi\left(t - \frac{c}{\sigma_{\hat{\theta}}}\right)}{\Phi\left(t + \frac{c}{\sigma_{\hat{\theta}}}\right) - \Phi(t)} \tag{29}$$

The width c matters, so the *p*-value does not give a generally correct expression for the probability of the null. For example, $\lim_{c \to 0} BF = 1$ (by l'Hôpital's rule) so $\lim_{c \to 0} p_{H_0} = 0.5$. The more interesting case is the completely diffuse prior, $c \to \infty$, where equation (29) shows the *p*-value to be correctly interpreted as the probability of the null. We have $\lim_{c \to \infty} BF = \frac{\Phi(t)}{1 - \Phi(t)}$, which gives $\lim_{c \to \infty} p_{H_0} = \Phi(t)$.

In summary, unlike the case for two-sided tests, *p*-values do give a correct probability statement about the null for a very reasonable specification of the prior.

6. Diffuse alternatives and the Lindley "paradox"

There is a bit of a folk theorem to the effect that use of a very diffuse prior (a prior that puts roughly equal weight on all possible parameter values) generates Bayesian results that are close to classical outcomes. The essence of the argument is that with a diffuse prior Bayesian results are dominated by the likelihood function, and that the same likelihood function drives the classical outcomes. The folk theorem is often correct: Bayesian highest posterior density intervals with diffuse priors are sometimes very similar to classical confidence intervals. However, this is irrelevant to correctly choosing between hypotheses. Application of Bayes theorem is required to correctly choose between hypotheses. And it turns out that in the presence of very diffuse priors classical test statistics are themselves irrelevant for choosing between hypotheses.

We've seen above that econometricians who use a classical decision rule to choose between hypotheses are not considering all alternatives equally. See the derivations above of the "implicit alternatives." In fact as the width of the alternative grows without limit, the probability of the null approaches one. Bayesians call this the "Lindley paradox" after Lindley (1957),[6] although there is nothing paradoxical about the result. Note again that in equation (19) $\lim_{c\to\infty} p_{H_0} = 1$ and in equation (25) $\lim_{\sigma_A\to\infty} p_{H_0} = 1$. In other words, the econometrician who wishes to treat all possible values of $\theta \neq \theta_0$ as equally likely may as well simply announce that the null hypothesis is true without doing any computation, thus saving a great deal of

[6] See also Bartlett (1957).

electricity. We can't take a completely agnostic position on parameter values and then conduct a meaningful hypothesis test.

Since this result may seem paradoxical, it is worth exploring further. Figure 10 shows the density for an estimated regression coefficient (the regression is described in more detail below) together with two illustrative uniform priors for the alternative, one broader than the other.[7] The more agnostic we are about a reasonable range for the alternative, the higher the probability in favor of the null. Here, the broader alternative gives $p_{H_0} = 0.58$ versus $p_{H_0} = 0.42$ even though the data is the same.

The probability of $\hat{\theta}$ under the null depends on the height of the density of $\hat{\theta}$ evaluated at θ_0. The probability of $\hat{\theta}$ under the alternative requires taking the conditional probability at a particular point, $\Pr(\hat{\theta}|\theta, H_A)$, multiplying by $\pi(\theta|H_A)$, and integrating to find $\Pr(\hat{\theta}|\theta)$ (see equation (17)). A wider alternative necessarily reduces the height of the prior density as the area under the prior integrates to 1. So while the wider the range of $\pi(\theta|H_A)$, the greater the area of the conditional probability that gets included, but the lower the value of $\pi(\theta|H_A)$. As we move from the solid to the dashed alternative, the area under the included portion of the bell curve increases, but only by a small amount. The increase in the area under the curve is outweighed by the decrease in the height of the alternative. In the illustration in Figure 10, most all of the density (99 percent) lies within either interval, so the probability of the $\hat{\theta}$ under the alternative is roughly $1/c$. Since $1/c$ is smaller for the wider alternative, the probability of $\hat{\theta}$ under the alternative is lower for the wide interval than for the narrower one. Further widening

[7] See Dickey (1977) for a discussion of the Lindley paradox for a normal coefficient with a uniform prior. Dickey also discusses a generalization to the uniform prior and to a Student *t*- likelihood.

of the alternative simply leads to a reduced calculation of the probability of the alternative and therefore an increase in p_{H_0}.

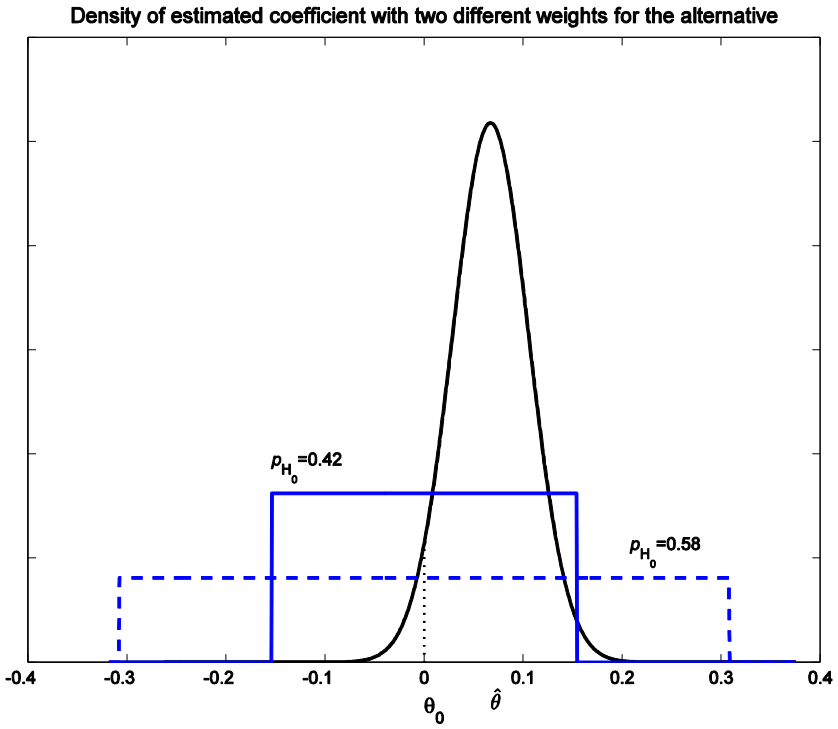

Figure 10

7. Is the stock market efficient?

The mathematics of Bayes theorem makes clear that specification of the alternative matters. To the classical econometrician the preceding discussion of the limiting values of the probability of the null as the width of the alternative increases may seem nihilistic. As a practical matter, we usually have something meaningful to say about what might be interesting alternatives. This means that the Lindley paradox often has little bite in application. For example, in the coin toss model, it was reasonable to bound the probability of a coin landing heads-up to be between zero and one. The difference between p_{H_0} and p-value generally does matter with reasonable priors.

As an example with more economic content, consider testing for weak form efficiency in the stock market. A classic approach is to estimate

$$r_t = \alpha + \theta r_{t-1} + \varepsilon_t \tag{30}$$

where r_t is the return on the market and under weak form efficiency $\theta = 0$.

I estimate equation (30) using the return on the S&P 500, once using monthly data and once using a much larger sample of daily data (Table 1). The underlying price data is the FRED II series SP500. The return is defined as the difference in log price between periods. Monthly observations use the price on the last day of the month. Since the objective is illustration rather than investigating the efficient market hypothesis, day of the week effects, January effects, heteroskedasticity, etc. are all absent. (Pace.) The *t*-statistic using the monthly data is 1.74. Tradition would label this result "weakly significant," suggesting that the null is probably not true but that the evidence is not overwhelming. Using the daily data, with many more observations, the *t*-statistic is 3.09, with a corresponding *p*-value of 0.002. So for this data the classical approach rejects the null in favor of the alternative. Whether this is the correct conclusion depends on how we define the alternative.

	Data	S&P 500 returns, monthly 1957M03 - 2012M08	S&P 500 returns, daily 1/04/1957 - 8/30/2012
(1)	observations	666	14,014
(2)	Coefficient on lagged return $(\hat{\theta})$ (std. error) [t-statistic]	0.067 (0.039) [1.74]	0.026 (0.008) [3.09]
(3)	p-value	0.082	0.002
	Probability of weak form efficiency, i.e. $\theta = 0$, single-point null		
(4)	Prior on alternative for lag coefficient $U[-.15,.15]$	0.42	0.11
(5)	Prior on alternative for lag coefficient $U[-.31,.31]$	0.58	0.20
(6)	Prior on alternative for lag coefficient $U[-1,1]$	0.82	0.44
(7)	BIC approximation	0.85	0.50
	Implicit prior to reject weak form efficiency		
(8)	Reject with probability>0.5	$\theta \sim U[-0.22,0.22]$	$\theta \sim U[-1.25,1.25]$
(9)	Reject with probability>0.95	$\emptyset$	$\theta \sim U[-0.07,0.07]$

Table 1

The fact that the prior matters is inconvenient. Reaching the wrong conclusion is more inconvenient. In lines (4), (5), and (6) of Table 1, I give the results for three different alternative widths. Lines (4) and (5) give the values underlying Figure 10. While the specific widths are mostly just for illustration, they make the point that p_{H_0} is very different from the usual p-value. Using monthly data, the frequentist conclusion would be moderate evidence against weak form efficiency, but for both priors in lines (4) and (5) the correct conclusion is that there is little evidence one way or another as to whether the market is weak form efficient. Using daily data, a frequentist would overwhelmingly reject weak form efficiency. The Bayes theorem conclusion

is rather more mild. The reports in lines (4) and (5) give clear evidence against market efficiency, but the evidence is well short of the 95 percent gold standard.

Because economists attach economically meaningful interpretations to parameters, we often have something to say about what might be reasonable values of those parameters. In this example, while formally the efficient market hypothesis is simply that $\theta = 0$, we really have a more nuanced view of what we learn from the lag coefficient. The larger the lag coefficient the more likely that there are easily exploitable trading opportunities.

This suggests a certain statistical paradox. If the investigator allows for a very wide alternative, we know that the probability of the null will be very large. In other words, entertaining *very* large deviations from the efficient market hypothesis is guaranteed to give evidence in favor of the efficient market hypothesis. Effectively, the prior is an ex ante position about what parameters should be considered reasonable, rather than a statement about the investigator's beliefs.[8] We understand that it is good practice to comment ex post on the "substantive significance" of parameter estimates.[9] Choosing between hypotheses calls for making such comment, i.e. specifying a prior for the alternative, ex ante.

In many applications, economic theory gives noncontroversial limits on the alternative. In equation (30) values of $|\theta| \geq 1$ imply an explosive process for stock returns, something which presumably everyone would be willing to rule out. In line (6) of Table 1, we see that this restriction again leads to the conclusion that the evidence is relatively inconclusive.

[8] Perhaps some Bayesians would disagree.

[9] McCloskey and Ziliak (1996) discuss the divergence between "good" and actual practice. Ziliak and McCloskey (2008) provide many examples of the consequences of considering statistical significance absent economic significance.

Since the prior specification for the alternative matters, what might an econometrician do in practice? Under ideal circumstances, economic theory offers some guidance. Alternatively, one can offer a range of calculations of p_{H_0} corresponding to a range of priors. In other words, rather than conclude simply that the null hypothesis is likely false or not, the investigator can offer a conclusion along the lines of "if you think the relevant alternative is X, then the null likely is (or isn't) true, but if you think the relevant alternative is Y, then the data says …" (The one thing an econometrician cannot do is estimate a parameter and then act as if a reasonable inferential range—a 95 percent confidence interval, or whatever—becomes a reasonable prior chosen ex post.)

Fortunately, a reasonable set of priors often lead to the same substantive conclusion. Figure 11 computes the probability $\theta = 0$ in equation (30) for a range of alternative specifications. For the monthly data where the frequentist conclusion was moderate evidence against weak form efficiency, a wide range of priors lead to the conclusion that the evidence is indecisive. For the daily data, where the conclusion was a very, very strong rejection of weak form efficiency, the same range of priors suggest moderate evidence against the efficient market hypothesis.

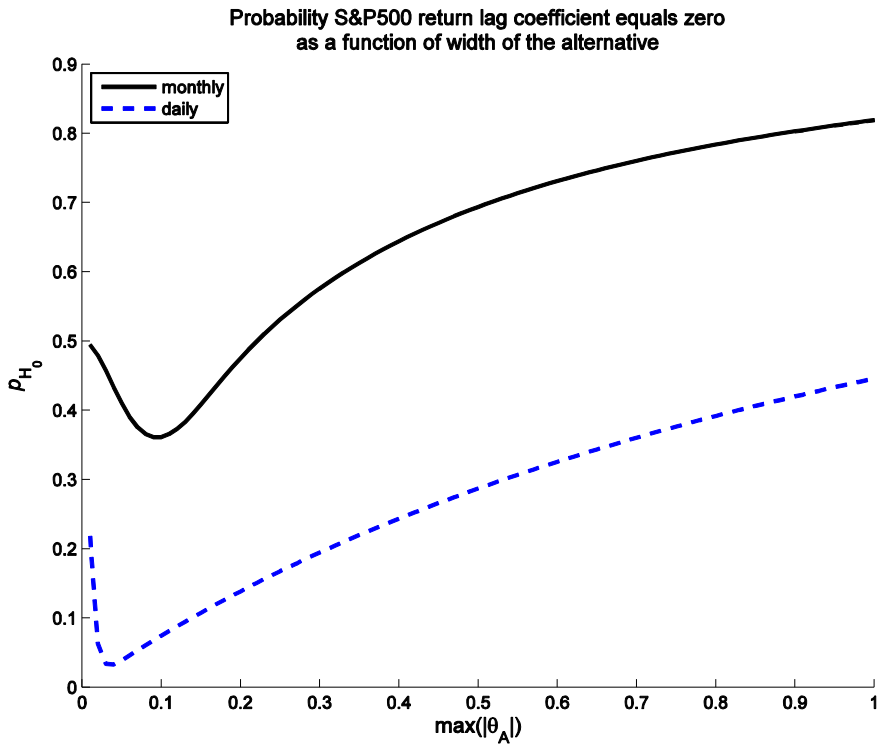

Figure 11

As before, we can compute the implicit prior being used by an econometrician using a frequentist decision rule. The widest alternative priors consistent with rejecting weak form efficiency are given in rows (8) and (9) of Table 1. For the monthly sample, the *p*-value is 0.08. But this moderate evidence against weak form efficiency is consistent with $p_{H_0} < 1/2$ only for alternatives between ± 0.22. Note that there is no prior consistent with finding a 95 percent probability against weak form efficiency.

The daily data provides much stronger evidence against weak form efficiency. If the decision rule is to choose the more likely hypothesis, the data rejects weak form efficiency for lag coefficients between ± 1.25. Since this is greater than a reasonable interval for θ, we can conclude that the daily data favors the alternative. If we want evidence at the standard 95

percent threshold, the widest possible interval is ± 0.07, in other words, the researcher must have started with a quite narrow prior.

8. Non-sharp hypotheses

The profession gets reminded from time to time to compare estimates to economically relevant parameter values, and that this is true for the null as well as the alternative. (See McCloskey (1985, 1992), with many references in the latter.) In particular, in most cases the economic theory encapsulated in a point null is valid in a small area around that null as well. Computing p_{H_0} allows one to formalize such a "non-sharp" null by defining the null over an interval rather than as a point. In the previous section, we derived $\Pr(\hat{\theta}|H_A)$ for an alternative covering a line segment. Bayes theorem then combined this value with the value of $\Pr(\hat{\theta}|H_0)$ for a point hypothesis. To compute Bayes theorem for a non-sharp hypothesis, we compute $\Pr(\hat{\theta}|H_0)$ for a line segment and compute $\Pr(\hat{\theta}|H_A)$ for a line segment with hole in the middle representing the null.

Define the null as covering an area c_0 around θ_0 and call the area under the alternative c_A, so that $\pi(H_0) \sim U[\theta_0 - c_0/2, \theta_0 + c_0/2]$ and $\pi(H_A) \sim U[-c_A/2, \theta_0 - c_0/2] \cup U[\theta_0 + c_0/2, c_A/2]$. Equation (3) becomes

$$p_{H_0} = \frac{\left(\frac{A_0}{c_0}\right) \cdot \pi(H_0)}{\left(\frac{A_0}{c_0}\right) \cdot \pi(H_0) + (A_A/(c_A - c_0)) \cdot \left(1 - \pi(H_0)\right)}$$

$$A_0 = F_{\hat{\theta}|\theta}\left(\theta_0 + \frac{c_0}{2\sigma_{\hat{\theta}}}\right) - F_{\hat{\theta}|\theta}\left(\theta_0 - \frac{c_0}{2\sigma_{\hat{\theta}}}\right) \tag{31}$$

$$A_A = F_{\hat{\theta}|\theta}\left(\theta_0 + \frac{c_A}{2\sigma_{\hat{\theta}}}\right) - F_{\hat{\theta}|\theta}\left(\theta_0 - \frac{c_A}{2\sigma_{\hat{\theta}}}\right) - A_0$$

When $\hat{\theta}$ is normally distributed, we have

$$\begin{aligned} A_0 &= \Phi\left(t + \frac{c_0}{2\sigma_{\hat{\theta}}}\right) - \Phi\left(t - \frac{c_0}{2\sigma_{\hat{\theta}}}\right) \\ A_A &= \Phi\left(t + \frac{c_A}{2\sigma_{\hat{\theta}}}\right) - \Phi\left(t - \frac{c_A}{2\sigma_{\hat{\theta}}}\right) - A_0 \end{aligned} \tag{32}$$

Note that for $c_A \gg \sigma_{\hat{\theta}}$, $A_A \approx 1 - A_0$. Note also that $\lim_{c_0 \to 0} A_0/c_0 = \phi(t)/\sigma_{\hat{\theta}}$.

If we have $\pi(H_0) = 1/2$ and $c_A \gg \sigma_{\hat{\theta}}$, equation (31) simplifies to

$$p_{H_0} = \frac{A_0}{A_0 + \frac{c_0}{c_A - c_0}(1 - A_0)} \tag{33}$$

	Data	S&P 500 returns, monthly 1957M03 2012M08	S&P 500 returns, daily 1/04/1957 8/30/2012
	Probability of weak form efficiency, finite null		
(1)	Null $\sim U[-0.0001, 0.0001]$, alternative $\sim U[-.15, .15]$ excluding $U[-0.0001, 0.0001]$	0.42	0.11
(2)	Null $\sim U[-0.02, 0.02]$, alternative $\sim U[-.15, .15]$ excluding $U[-0.02, 0.02]$	0.43	0.68

Table 2

Table 2 illustrates what can happen if the null is changed to an interval using the market efficiency example. If the interval is very small, then of course the results are the same as for a point null. The probabilities for the point null (row (4) of Table 1) are the same as the probabilities from a very small interval (row (1) of Table 2)). A wider null—here $\theta_0 \sim U[-.02, .02]$ just as an illustration—can make a larger difference. Comparing row (4), Table

1 and row (2), Table 2 for the daily data as an example, the broader null ($\theta \in \pm 0.02$) changes the conclusion from notably against weak form efficiency to modest evidence in favor.

The ability to allow for a non-sharp null using Bayes theorem contrasts sharply with most frequentist procedures. The difficulty in testing for a non-sharp null in the frequentist framework relates directly to the issue of specifying a prior for the null, the issue being that a functional form for the prior is required. For many test statistics, computing the probability that the test will reject conditional on a given value of the true parameter, $\Pr(\tau = 1|\theta)$ is straightforward. The difficulty is that the size of the test on the non-sharp null is

$$\alpha = \int_{-\infty}^{\infty} \Pr(\tau = 1|\theta)\, \pi(\theta|H_0) d\theta \tag{34}$$

which requires $\pi(\theta|H_0)$ just as much as does use of Bayes theorem.

9. Bayes theorem and consistent estimation

The examples above use a finite number of observations. It is useful to see the intuition for why under the point null the statistic p_{H_0} consistently identifies the correct hypothesis so long as the true value of θ is included in the prior for the alternative. Assume, as is the usual case, that in a large sample the standard deviation is inversely proportional to the square root of the sample size. This means there is some constant κ_σ such that $\sigma_{\hat{\theta}} \approx \kappa_\sigma/\sqrt{n}$.

Consider first the uniform alternative. Equation (35) repeats equation (19) for convenience of reference.

$$p_{H_0} = \frac{\phi(t)}{\phi(t) + \frac{\sigma_{\hat{\theta}}}{c}\left[\Phi\left(t + \frac{c}{2\sigma_{\hat{\theta}}}\right) - \Phi\left(t - \frac{c}{2\sigma_{\hat{\theta}}}\right)\right]} \tag{35}$$

If the null is true, then the term $\left[\Phi\left(t+\frac{c}{2\sigma_{\hat{\theta}}}\right)-\ \Phi\left(t-\frac{c}{2\sigma_{\hat{\theta}}}\right)\right]$ converges in probability to 1, since $c/2\sigma_{\hat{\theta}}$ becomes arbitrarily large. Since $\sigma_{\hat{\theta}}/c \to 0$, $p_{H_0} \to \phi(t)/\phi(t) \to 1$. Under the alternative, the term $t-\frac{c}{2\sigma_{\hat{\theta}}} \approx \left(\theta-\theta_0-\frac{c}{2}\right)/\sigma_{\hat{\theta}}$. If $\theta < \theta_0+\frac{c}{2}$, which is to say that the alternative covers the true parameter, then $t-\frac{c}{2\sigma_{\hat{\theta}}}$ converges in probability to $-\infty$ and the term $\left[\Phi\left(t+\frac{c}{2\sigma_{\hat{\theta}}}\right)-\ \Phi\left(t-\frac{c}{2\sigma_{\hat{\theta}}}\right)\right]$ again converges to 1. Since the numerator converges to $\phi(\infty)=0$, p_{H_0} consistently identifies the alternative.

Now consider the normal alternative, which is slightly easier since the alternative prior always covers the true value θ. The Bayes factor is given in equation (24), copied here for convenience as equation (36).

$$BF_{0A} = \frac{\sqrt{1+(\sigma_A/\sigma_{\hat{\theta}})^2}\,\phi\left(\sqrt{\frac{(\sigma_A/\sigma_{\hat{\theta}})^2}{1+(\sigma_A/\sigma_{\hat{\theta}})^2}}\,t\right)}{\phi(0)} \tag{36}$$

For small $\sigma_{\hat{\theta}}$, the term $\sqrt{\frac{(\sigma_A/\sigma_{\hat{\theta}})^2}{1+(\sigma_A/\sigma_{\hat{\theta}})^2}} \to 1$. Under the null, the second factor in the numerator approaches $\phi(t)$. Since the first factor $\sqrt{1+(\sigma_A/\sigma_{\hat{\theta}})^2} \to \infty$, the Bayes factor goes to infinity and p_{H_0} converges to 1. Under the alternative, $t \to (\theta-\theta_0)/\sigma_{\hat{\theta}}$. The numerator of equation (36) converges to $\left(\frac{\sqrt{n}}{\kappa_\sigma}\right)\phi\left((\theta-\theta_0)\times\frac{\kappa_\sigma}{\sqrt{n}}\right)$. The second term is proportional to $\exp\left(-\frac{1}{2}\left((\theta-\theta_0)\times\frac{\kappa_\sigma}{\sqrt{n}}\right)^2\right)$, which goes to zero faster than $\sqrt{n}$ grows. The Bayes factor goes to zero, so p_{H_0} goes to zero.

10. More General Bayesian Inference

Having to specify prior weights between the null and the alternative is likely not terribly objectionable since 50/50 can be regarded as neutral. The practical issue of picking a prior for a continuous alternative is less pleasant. Fortunately, we can borrow a large sample approximation from the Bayesian literature that requires no extra Bayesian computation and is equivalent to an implicit, relatively agnostic specification of weights. This is the Bayesian Information Criterion (BIC or SIC) due to Raftery (1986a,b), who also advocated for the perspective here—that p_{H_0} should replace *p*-values. Econometricians often use the BIC for auxiliary choices. (Choosing lag lengths for unit root tests is a common example.) The BIC allows for a trivial computation of p_{H_0}, without requiring an explicit statement of the prior. In the first subsection here, I present the BIC. In the second subsection, I give a light introduction to its derivation.

10.1 Use of the BIC

In a regression, or other estimator where using a *t*-test is asymptotically valid, Kass and Raftery (1995) show that the BIC, B, can be written as in the first line in equation (37). We can use the BIC to calculate the probability of the null, thus approximating equation (19) or equation (25) using the expression in the second line,

$$\begin{aligned} B &= -t^2 + \log n \\ p_{H_0} &= \frac{\exp(.5 \cdot B)}{1 + \exp(.5 \cdot B)} \end{aligned} \tag{37}$$

where t is the *t*-statistic against the null and n is the number of observations.

In the daily data stock market example $t = 3.09$, which implies the BIC approximation $p_{H_0} = 0.49996$. This is slightly more conservative than the probability under the restriction that returns are stationary ($-1 < \theta < 1$). For the monthly data, the BIC gives essentially the same result as the stationarity restriction. It is a common result that the BIC is conservative in the sense of requiring strong evidence to move away from the null (Raftery (1999)).

The conservative nature of the BIC can be seen further in the following large sample approximation to the width of the uniform prior considered earlier in the paper. Startz (2014) shows that in a large sample, say $n > 60$, use of the BIC is equivalent to using the uniform prior with width

$$c = \sqrt{2\pi n \sigma_{\hat{\theta}}^2} \tag{38}$$

For example, in line (6) of Table 1 we consider the width $c = 2$. For the monthly data with 666 observations and $\sigma_{\hat{\theta}} = 0.039$ equation (38) gives $c = 2.52$. The implicit value of the uniform prior is close to the explicit prior used in line (6), which is why the probabilities of the null given by the two calculations are similar.

10.2 A Light Derivation of the BIC

The BIC is derived using an approximation due to Schwarz (1978). (Hence the alternative name "SIC.") A brief introduction may be helpful and also will serve to set out notation for the section on Bayesian inference below. The heart of Bayesian analysis is Bayes theorem, which in this context is that the posterior equals the likelihood times the prior over the marginal likelihood.

$$\Pr(\theta|y,H) = \frac{\Pr(y|\theta,H)\pi(\theta|H)}{p(y|H)} \tag{39}$$

The denominator—called the marginal likelihood—equals the integral of the numerator. To see this, integrate both sides with respect to θ, remember that the density on the left has to integrate up to 1.0, and note that the denominator is constant with respect to θ, as in

$$\int \Pr(\theta|y,H)d\theta = 1 = \frac{\int \Pr(y|\theta,H)\pi(\theta|H)d\theta}{p(y|H)} \tag{40}$$

Note for later the distinction between equation (39) where the information we condition on is the available data, y, and the outcome is an estimate of the probability of the parameter θ, and the earlier equation (1), where we condition on the frequentist estimator $\hat{\theta}$ and the outcome is an estimate of the probability of the null hypothesis.

What we want is $p(H_0|y)$. Using Bayes theorem this is

$$p(H_0|y) = \frac{p(y|H_0)\pi(H_0)}{\Pr(y)} \tag{41}$$

If we divide equation (41) by the analogous equation for the alternative, we get the posterior odds ratio, PO.

$$PO = \frac{\Pr(H_0|y)}{\Pr(H_A|y)} = \frac{\Pr(y|H_0)}{\Pr(y|H_A)} \times \frac{\pi(H_0)}{\pi(H_A)} = BF \times \frac{\pi(H_0)}{\pi(H_A)} \tag{42}$$

Note that $\Pr(y)$ cancels, just as $\Pr(\hat{\theta})$ did in getting from equation (1) to equation (2). The ratio of the marginal likelihoods $\Pr(y|H_0)/\Pr(y|H_A)$ is called the Bayes factor, BF. If we take $\pi(H_0) = 1/2$ then, as before, the posterior odds ratio equals the Bayes factor.

The key to understanding why we can have a large sample approximation to the Bayes factor that does not involve the prior is the fact that in the numerator of equation (41) the likelihood function grows with the sample size while the prior remains constant, so for a large sample the product is entirely dominated by the likelihood. The Schwarz approximation begins

by applying the central limit theorem in order to say that in a large sample the likelihood function is approximately normal. Next one takes a second-order Taylor series approximation around the maximum likelihood estimate. A good explanation is given in Greenberg (2008), who gives the following Laplace approximation to the log Bayes factor

$$\log BF \approx .5B + \left[\log\frac{\pi_0\left(\hat{\theta}_{0,mle}\right)}{\pi_A\left(\hat{\theta}_{A,mle}\right)} + \frac{1}{2}\log\frac{|V_0|}{|V_A|} + \frac{k_0 - k_A}{2}\log 2\pi\right] \tag{43}$$

$$B = \left[2 \cdot \log\frac{\Pr\left(y\middle|\hat{\theta}_{0,mle}\right)}{\Pr\left(y\middle|\hat{\theta}_{A,mle}\right)} - (k_0 - k_A)\log n\right] \tag{44}$$

where j is the null or alternative, $j \in \{0, A\}$, $\widehat{\theta_j}$ is the maximum likelihood estimate, $\pi_j\left(\hat{\theta}_{j,mle}\right)$ is the prior evaluated at the maximum likelihood estimate, and V_j is the maximum likelihood estimate of the variance of $\hat{\theta}_j$. B is twice the difference in the Schwarz approximation, approximates twice the log Bayes factor, and grows with the number of observations. The second term in equation (43), dropped in the BIC, does not depend on the number of observations but does depend on the prior. Dropping the second term gives an approximation error that is small relative to the sample size, but which does not vanish.

A quick look at the first term in equation (44) explains why the BIC does a good job of picking out the more likely hypothesis. Suppose first the null is true. Then the first term is the likelihood ratio, which is distributed $\chi^2(k_A - k_0)$. The expected value of the first term is $k_A - k_0$, 1 in the examples in the paper, while the second term is proportional to $\log n$. So the approximate Bayes factor is proportional to n and the probability of the null converges to 1.0. In contrast, suppose the alternative is true. The likelihood ratio is a constant plus a difference in the respective sum of squared residuals, so the first term is dominated by a sum of order n

(which is much larger than $\log n$). The approximate Bayes factor is roughly proportional to e^{-n} so the probability of the null converges very rapidly to zero.

The BIC relies on the idea that as the number of observations grows the posterior is completely dominated by the likelihood function so the role of the prior becomes negligible (Schwarz (1978)). In a maximum-likelihood model with d restrictions the BIC is given by[10]

$$B = -2\left[\log \Pr\left(y\middle|\hat{\theta}_{mle}, H_A\right) - \log \Pr\left(y\middle|\hat{\theta}_{mle}, H_0\right)\right] + \mathrm{d} \cdot \log n \tag{45}$$

The first term is the usual likelihood ratio statistic. The second term is sometimes called a sample size penalty, but the "penalty" is inherent in the Bayes factor—not ad hoc.

The expression for B also adds to our intuition that as the sample size grows the null is increasingly preferred unless the test statistic is also growing. If the null is true, then the likelihood ratio statistic in the square brackets should be small no matter the sample size. (The expected value of t^2 under the null is about 1.0.) In contrast, if the alternative is true then the likelihood ratio statistic grows in proportion to the sample size. So if the test statistic has not gotten quite large as the sample grows, then the evidence is against the alternative.

While the BIC is very handy, it does include an approximation error precisely because it leaves out the prior. Use of the BIC is not without controversy among Bayesians precisely because it involves using an implicit rather than explicit prior for $\pi(\theta|H_A)$. See Weakliem (1999a) and the discussion in Firth (1999), Gelman (1999), Raftery (1999), Weakliem (1999b), and Xie (1999). However, it is hard to imagine that using the BIC is anything other than a huge improvement over not computing p_{H_0} at all.

[10] This expression is due to Kass and Raftery (1995). To compute the probability of the hypothesis that all k coefficients in a regression equal zero one can use $B = -k \cdot F + k \cdot \log n$, where F is "the" *F*-statistic for the regression.

Approximations to p_{H_0} computed from the BIC as in equation (37) have the practical advantage of being non-manipulatable by the investigator's choice of prior, thus providing something of a scientific level playing field. For example, Xie (1999) writes

> The BIC requires neither the input of a prior on the part of the researcher nor adjustment according to frequency distributions. The simplicity with which the BIC can be readily applied to social science research is of high value in practice, for it allows researchers with different theoretical and statistical orientations to communicate research results without the danger of 'twisting' statistics to fit one's own preconceptions.

10.3 Departures from the Bayesian Approach

While the heart of this paper is the insistence on the importance of respecting Bayes theorem, and while use is made of Bayesian tools, the recommendations here are a step apart from a Bayesian approach. The differences are two-fold. First, I condition only on the information in the frequentist estimator, $\hat{\theta}$, where a complete Bayesian analysis uses the complete set of data, y. Second, as is discussed in the next section, Bayesians often eschew hypothesis testing in favor of a more explicit decision-theoretic approach.

The theme of this paper is to convince frequentists—which as a practical matter means just about all empirical economists—to use Bayes theorem to choose between hypotheses. For this reason, the discussion has been framed entirely in terms of conditioning on $\hat{\theta}$. In general $\hat{\theta}$ is not a sufficient statistic for the data, and therefore there is a loss of information inherent in the approach taken here. However, in the classical linear regression framework with known error

variance and normal errors, it can be shown that $\{\hat{\theta}, \sigma^2_{\hat{\theta}}\}$ is a sufficient statistic so no loss of information occurs.[11]

Consider the following demonstration, which is a very slight extension of the argument given by Poirier (1995, page 221). A set of statistics $\widehat{\Theta}$ is jointly sufficient if and only if the likelihood can be partitioned into a function $h(\cdot)$ independent of θ and a function $d(\cdot)$ depending on y only through $\widehat{\Theta}$ as in

$$\Pr(y|\theta) = d\left(\widehat{\Theta}\middle|\theta\right) \times h(y) \tag{46}$$

For the standard regression model

$$y_i = \theta x_i + \varepsilon_i, i = 1, \ldots, n, \varepsilon_i \sim iidN(0, \sigma^2_\varepsilon) \tag{47}$$

the likelihood function is

$$\Pr(y|\theta, x) = (2\pi\sigma^2_\varepsilon)^{-\frac{n}{2}} \exp\left[-\frac{1}{2\sigma^2_\varepsilon}\sum_{i=1}^{n}(y_i - \theta x_i)^2\right] \tag{48}$$

Letting $\hat{\theta}$ be the maximum-likelihood (and least squares) estimator, writing $\sum_{i=1}^{n}(y_i - \theta x_i)^2 = \sum_{i=1}^{n}\left(y_i - \hat{\theta}x_i\right)^2 + \sum_{i=1}^{n}\left(\left(\hat{\theta} - \theta\right)x_i\right)^2$ and substituting gives

$$\begin{aligned}\Pr(y|\theta, x) = &\left((2\pi\sigma^2_\varepsilon)^{-\frac{n}{2}} \exp\left[-\frac{1}{2\sigma^2_\varepsilon}\sum_{i=1}^{n}\left(\left(\hat{\theta} - \theta\right)x_i\right)^2\right]\right) \\ &\times \left(\exp\left[\sum_{i=1}^{n}\left(y_i - \hat{\theta}x_i\right)^2\right]\right)\end{aligned} \tag{49}$$

The first factor in equation (49) corresponds to $d(\cdot)$ and the second factor corresponds to $h(\cdot)$. So the regression coefficient $\hat{\theta}$ in this case is a sufficient statistic and there is no loss when compared to a complete Bayesian analysis.

[11] In general, the set of sufficient statistics for the linear regression model is comprised of the estimated coefficients, the standard error of the regression, the product moment matrix $X'X$, and the number of observations. (See Geweke (2005), page 33.) For a reasonable size sample, little harm is done by treating $\{\hat{\theta}, \sigma^2_{\hat{\theta}}\}$ as containing all the needed information.

As a practical matter, in the vast majority of empirical work taking $\hat{\theta} \sim N\left(\theta, \sigma_{\hat{\theta}}^2\right)$ is surely a minor sin compared to other approximations required for analysis.

There is a slightly different argument that can be made in favor of a complete Bayesian analysis involving multiple parameters. The examples earlier consider the scalar parameter θ. In cases where there are multiple parameters or where the distribution of the parameter of interest depends on nuisance parameters, a complete Bayesian approach may simply be easier to implement. However, for the most standard econometric estimators the approach taken here is likely to be satisfactory, at least in a large sample. For example while I treat $\sigma_{\hat{\theta}}^2$ as known, in practice it needs to be estimated. Ignoring this in a large sample is fairly harmless. In the standard Bayesian regression model, if one begins with normal-gamma priors for $\{\theta, \sigma_\varepsilon^2\}$ with ν degrees of freedom for the prior for σ_ε^2, the marginal posterior for θ is t with $\nu + n$ degrees of freedom. (See Koop (2003) pp. 19-20.) For large n the t- distribution is, of course, indistinguishable from the normal.

11. The General Decision-Theoretic Approach

11.1 Wald's Method

One suspects that most empirical economists would prefer to not have to specify prior distributions. For the reasons demonstrated above, choosing between hypotheses necessitates priors; the choice being whether the priors are explicit or implicit. One also suspects that most empirical economists, faced with the necessity of choosing a prior, would prefer a very diffuse prior. Because of the Lindley paradox, very diffuse priors don't "work" for hypothesis testing. In principle, and sometimes in practice, the solution is to eschew hypothesis testing in favor of a decision-theoretic framework.

Wald (1950) proposed the following decision-theoretic framework.[12] Suppose the investigator wishes to minimize expected losses over a loss function $L(A,\theta)$, where A is the action.[13] The investigator solves

$$\min_A \int L(A,\theta)p\left(\theta\middle|\hat{\theta}\right)d\theta \tag{50}$$

What matters in equation (50) is the posterior. Unlike the situation we've seen for hypothesis choice, the Lindley paradox does not arise in calculating Bayesian posteriors. One can specify a very diffuse prior, therefore allowing the posterior to be determined by the data. The Lindley paradox arises because a diffuse prior sends the marginal likelihood for the alternative, the integral of the likelihood times the prior, toward zero. In contrast, in computing the posterior the growth in the marginal likelihood is cancelled by the growth in the likelihood times the prior, leaving a well-behaved posterior.

For a uniform prior for $\left|\theta-\hat{\theta}\right| \leq c/2$, if we let $I(\theta,\hat{\theta})$ be an indicator function equal 1 if θ is inside the support of the prior we can write the posterior as

$$\begin{gathered} I(\theta,\hat{\theta}) = \begin{cases} 1, \left|\theta-\hat{\theta}\right| \leq c/2 \\ 0, \left|\theta-\hat{\theta}\right| > c/2 \end{cases} \\ p\left(\theta\middle|\hat{\theta}\right) = \frac{\frac{1}{\sigma_{\hat{\theta}}}\phi(t)\frac{1}{c}I(\theta,\hat{\theta})}{\frac{1}{c}\left[\Phi\left(t+\frac{c}{2\sigma_{\hat{\theta}}}\right)-\Phi\left(t-\frac{c}{2\sigma_{\hat{\theta}}}\right)\right]} \\ \lim_{c\to\infty} p\left(\theta\middle|\hat{\theta}\right) = \frac{1}{\sigma_{\hat{\theta}}}\phi(t) = f_n\left(\theta;\hat{\theta},\sigma_{\hat{\theta}}\right) \end{gathered} \tag{51}$$

[12] For work foreshadowing Wald, see Simon (1945).

[13] Note that choosing a hypothesis based on p_{H_0} and p_{H_A} can itself be thought of as a decision problem. See Li et. al. (2014). Le et. al. also propose loss functions for choosing between hypotheses that do not suffer from the Lindley paradox.

While the denominator in equation (51) goes to zero as c grows, the ratio is well-behaved. The c in the numerator and denominator cancel. For large c, both $I(\theta, \hat{\theta})$ and the term in square brackets in the denominator go to one. In other words, for large c the posterior approaches $\theta \sim N\left(\hat{\theta}, \sigma_{\hat{\theta}}^2\right)$.

For the normal prior, the results are essentially as given in equations (21) and (22). If the prior is $\theta \sim N(\theta_0, \sigma_0^2)$, then the posterior is

$$\begin{aligned} \theta|\hat{\theta} &\sim N\left(\tilde{\theta}, \tilde{\sigma}^2\right) \\ \tilde{\theta} &= \frac{\sigma_0^2 \hat{\theta} + \sigma_{\hat{\theta}}^2 \theta_0}{\sigma_0^2 + \sigma_{\hat{\theta}}^2} \\ \tilde{\sigma}^2 &= \frac{\sigma_{\hat{\theta}}^2 \sigma_0^2}{\sigma_{\hat{\theta}}^2 + \sigma_0^2} \end{aligned} \tag{52}$$

As $\sigma_0^2 \to \infty$, $\tilde{\theta} \to \hat{\theta}$ and $\tilde{\sigma}^2 \to \sigma_{\hat{\theta}}^2$, so in this case too the posterior is approximately $N\left(\hat{\theta}, \sigma_{\hat{\theta}}^2\right)$. Notice that this is exactly the usual frequentist distribution, swapping the roles of θ and $\hat{\theta}$. This suggests that even though one ought to employ whatever information is available in a prior, a decision-theoretic analysis that uses the frequentist distribution as if it were a posterior may not go too far astray.

Economists like decision-theoretic problems in principle, and occasionally in practice as well. For a Bayesian, the posterior summarizes all the information necessary as input to the decision problem. A prior for these problems is important if an informative prior is available, but a diffuse prior raises no paradoxes.

However, nothing about decision-theoretic problems opens up a backdoor approach to choosing between hypotheses with a diffuse prior. To the extent that specification searches are part of empirical investigations, hypothesis tests are an important tool. The argument

presented in this monograph is that *if* a choice is to be made between competing hypotheses, the choice should be made in a way that respects Bayes theorem.

11.2 Akaike's Method

For problems in which it is sensible to choose between competing hypotheses, application of Bayes theorem provides a method for assessing relative probabilities. However, Bayes theorem is not the only method for making model comparisons. In particular, there are information-theoretic methods based on the Akaike Information Criterion (AIC) that provide relative model probabilities that are analogous to equation (37). In Monte Carlo studies AIC-based methods outperform BIC-based methods in some circumstances, and vice versa. Burnham and Anderson (2002) provides the background for AIC-based methods and their connection to the Kuller-Leiber distance metric. See in particular sections 6.3 and 6.4 of Burnham and Anderson (2002) for a discussion of model selection and comparisons between AIC and BIC based methods.

12. A Practitioner's Guide to Choosing Between Hypotheses

I give here a quick summary of the three formulas an applied econometrician might want to use to compare hypotheses of the form $\theta = \theta_0$ versus $\theta \neq \theta_0$ where the estimate $\hat{\theta}$ is normally distributed with mean θ and variance $\sigma_{\hat{\theta}}^2$. The canonical example would be the test of whether a regression coefficient equals zero.

The best practice is to report the probability of the null hypothesis given the statistical estimate, $p_{H_0} \equiv \Pr\left(H_0 \middle| \hat{\theta}\right)$, which is given approximately by

$$p_{H_0} \approx \frac{\phi(t)}{\phi(t) + \left[{}^{c}/_{\sigma_{\hat{\theta}}}\right]^{-1}} \tag{53}$$

where $\phi(\cdot)$ is the standard normal density and t is the usual *t*-statistic. The constant c, which is chosen by the investigator as the width of an interval centered on θ_0 which contains the true value of θ, should also be reported.

The second best practice is to report the probability p_{H_0} using a large sample approximation based on the Bayesian Information Criterion, making it unnecessary to be explicit about the choice of c needed to implement equation (53).

$$p_{H_0} \approx \frac{\exp\left(\frac{1}{2}[-t^2 + \log n]\right)}{1 + \exp\left(\frac{1}{2}[-t^2 + \log n]\right)} \tag{54}$$

where t is the usual *t*-statistic and n is the number of observations.

If the investigator wishes to only use classical hypothesis testing, then the fallback, third best, practice is to conduct the usual hypothesis test and to report the implicit prior value of c consistent with rejecting the null hypothesis. For a rejection at the $1 - \alpha$ level of significance the implicit value of c is the maximum c solving

$$\alpha = \frac{\phi(t)}{\phi(t) + \frac{\sigma_{\hat{\theta}}}{c}\left[\Phi\left(t + \frac{c}{2\sigma_{\hat{\theta}}}\right) - \Phi\left(t - \frac{c}{2\sigma_{\hat{\theta}}}\right)\right]} \tag{55}$$

where $\Phi(\cdot)$ is the standard normal cdf. Note that equation (55) does not always have a solution. If there is no solution, then any rejection of the null hypothesis ought to be reconsidered.

Application of any of equations (53), (54), or (55) involves some assumptions and some approximations, all of which were discussed above.

13. Summary

Classical econometric techniques tell us the probability of seeing a given estimate if the null is true, but do not tell us the probability that the null is true given the observed estimate. It is the latter that is relevant for distinguishing between economic hypothesis. In practice, the distinction can be quite large. Using our standard test statistics to choose between hypotheses *is very likely to lead to the wrong conclusion.* What's more, the error can go in either direction.

The key insight is that considering the distribution of an estimator under the null is not enough; the distribution under the alternative matters as well. A traditional, "statistically significant" rejection of the null hypothesis tells us that the realized estimate is very unlikely under the null. But the realized estimate may, or may not, be even more unlikely under the alternative. The point that in considering hypotheses what matters is the relative weight of the evidence was put this way in 1893 by Sherlock Holmes, "That is the case as it appears to the police, and improbable as it is, all explanations are more improbable still." (Doyle (1893), p. 11.)

So, what is to be done? Proper calculations that are respectful of Bayes theorem ought to supplement, or perhaps supplant, classical test statistics and *p*-values in reports of empirical results. Report the probability of the null or alternative, along with the relevant value of c using equation (13) or σ_A using equation (14). Where that is not feasible, at the least one should report the probability implied by the BIC. Those who insist on using traditional hypothesis tests should at least calculate their implicit priors according to equation (55) and report that as well. For regressions and most standard econometric techniques, especially for the usual hypotheses about the value of a single coefficient, the required calculations are straightforward.

References

Arrow, Kenneth J. 1960. "Decision Theory and the Choice of a Level of Significance for the *t*-Test," Olkin, Churye, Hoeffding, Madow, and Mann, eds., *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling,* Stanford University Press, Stanford, CA.

Bartlett, M.S. 1957. "A Comment on D.V. Lindley's Statistical Paradox," *Biometrika,* 44, 533-534.

Bayes, Thomas. 1763. "An Essay towards Solving a Problem in the Doctrine of Chances," *Philosophical Transactions,* Royal Society of London, vol. 53, 269-271.

Burnham, Kenneth P. and David R. Anderson, 2002. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach,* 2nd edition, Springer, New York.

Casella, George and Roger L. Berger, 1987. "Reconciling Bayesian and Frequentist Evidence in the One-Sided Testing Problem," *Journal of the American Statistical Association,* Vol. 82, No. 387, 106-111.

Degroot, Morris H. 1973. "Doing What Comes Naturally: Interpreting a Tail Area as a Posterior Probability or as a Likelihood Ratio," *Journal of the American Statistical Association,* Vol. 68, No. 344, 966-969.

Dickey, James M. 1971. "The Weighted Likelihood Ratio, Linear Hypotheses on Normal Location Parameters," *The Annals of Mathematical Statistics*, Vol. 42, No. 1, 204-223.

_____. 1977. "Is the Tail Area Useful as an Approximate Bayes Factor?," *Journal of the American Statistical Association*, Vol. 72, No. 357, 138-142.

Doyle, Arthur C. 1893. "Silver Blaze," *The Memoirs of Sherlock Holmes,* reprinted 1993 Oxford University Press, New York.

Firth, David and Jouni Kuha. 1999. “Comments on ‘A Critique of the Bayesian Information Criterion for Model Selection,’” *Sociological Methods & Research*, 27(3), 398-402.

Gelman, Andrew and Donald B. Rubin. 1999. “Evaluating and Using Methods in the Social Sciences: A Discussion of ‘A Critique of the Bayesian Information Criterion for Model Selection,’” *Sociological Methods & Research*, 27(3), 403-410.

Geweke, John. 2005. *Contemporary Bayesian Econometrics and Statistics*, John Wiley & Sons, Hoboken, NJ.

Greenberg, Edward. 2008. *Introduction to Bayesian Econometrics*, Cambridge University Press, Cambridge.

Hausman, Jerry A. 1978. “Specification tests in econometrics,” *Econometrica,* Vol. 46. No. 6, 1251-1271.

Ioannidis, John P.A. 2005. “Why most published research findings are false,” *PLoS Medicine,* Vol. 2, Issue 8, 696-701.

Jeffreys, Harold. 1961. *Theory of Probability*, 3rd ed., Oxford University Press, Oxford.

Kass, Robert E. and Adrian E. Raftery. 1995. “Bayes Factors,” *Journal of the American Statistical Association*, Vol. 90, No. 430, 773-796.

Koop, Gary. 2003. *Bayesian Econometrics,* John Wiley & Sons.

Laplace, Pierre. 1774. “Mémoire sur la probabilité des causes par les événemens,” *Savants étranges* 6, 621-656

Leamer, Edward E. 1978. *Specification Searches: Ad Hoc Inference with Nonexperimental Data*, John Wiley and Sons, New York.

_____. 1983a. “Model Choice and Specification Analysis,” *Handbook of Econometrics,* Z. Griliches and M.D. Intriligator eds., vol. 1, North Holland Publishing Co., 285-330.

_____. 1983b. “Let’s Take the Con out of Econometrics,” *American Economic Review,* Vol. 73, No. 1, 31-43.

Li, Yong, Tao Zeng, and Jun Yu, 2014. “A New Approach to Bayesian hypothesis testing,” *Journal of Econometrics,* vol. 178, No. 3, 602-612.

Lindley, D.V. 1957. “A Statistical Paradox,” *Biometrika,* 44, 187-192.

McCloskey, Donald N. 1985. “The Loss Function Has Been Mislaid: The Rhetoric of Significance Tests,” *American Economic Review,* vol. 75. No. 2, 201-205.

_____. 1992. “Other things equal: The bankruptcy of statistical significance,” *Eastern Economic Journal,* vol. 18. No. 3, 359-361.

_____, and Stephen T. Ziliak. 1996. “The Standard Error of Regressions,” *Journal of Economic Literature,* Vol. XXXIV, 97-114.

McGrayne, Sharon B. 2011. *The Theory That Would Not Die: How Bayes' Rule Cracked the Enigma Code, Hunted Down Russian Submarines, and Emerged Triumphant from Two Centuries of Controversy,* Yale University Press, New Haven, CT.

Neyman, J. and E. S. Pearson. 1933. “On the Problem of the Most Efficient Tests of Statistical Hypotheses,” *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, Vol. 231, 289-337.

Pearson, E.S. 1938. “‘Student’ as a Statistician,” *Biometrika,* 30, 210-250.

Poirier, Dale. 1995. *Intermediate Statistics and Econometrics,* M.I.T. Press, Cambridge.

Raftery, Adrian E. 1986a. “Choosing Models for Cross-Classifications,” *American Sociological Review*, Vol. 51, No. 1,145-146.

_____. 1986b. “A Note on Bayes Factors for Log-linear Contingency Table Models With Vague Prior Information,” *Journal of the Royal Statistical Society, Series B*, vol. 48. pp. 249-250.

_____. 1995. “Bayesian Model Selection in Social Research,” *Sociological Methodology*, 25, 111-163.

_____. 1999. “Bayes Factors and BIC: Comment on ‘A Critique of the Bayesian Information Criterion for Model Selection,’” *Sociological Methods & Research*, 27(3), 411-427.

Schwarz, Gideon. 1978. “Estimating the Dimension of a Model,” *The Annals of Statistics*, 6, 461-464.

Simon, Herbert. 1945. “Statistical Tests as a Basis for ‘Yes-No’ Choices,” *Journal of the American Statistical Association,* Vol. 40., No 229, 80-84.

Startz, Richard. 2013. “On the Implicit Uniform BIC Prior,” *Economics Bulletin*, Vol. 34, No. 2, April 2014.

Wald, Abraham. 1950. *Statistical Decision Functions,* Wiley, New York.

Weakliem, David L. 1999a. “A Critique of the Bayesian Information Criterion for Model Selection,” *Sociological Methods & Research*, 27(3), 359–397.

____ 1999b. “Reply to Firth and Kuha, Gelman and Rubin, Raftery, and Xie,” *Sociological Methods & Research*, 27(3), 436-443.

Xie, Yu. 1999. “The Tension Between Generality and Accuracy,” *Sociological Methods & Research*, 27(3), 428-435.

Ziliak, Stephen T. and Deirdre N. McCloskey. 2008. *The Cult of Statistical Significance,* University of Michigan Press, Ann Arbor, MI.