Are All Economic Hypotheses False?*

J. Bradford De Long
Department of Economics
Harvard University
Cambridge, MA 02138
and
National Bureau of Economic Research

and

Kevin Lang
Department of Economics
Boston University
270 Bay State Road
Boston, MA 02215
and
National Bureau of Economic Research

November, 1989

*We thank, without implicating, Gary Koop, Kevin Murphy, and Jim Stock for stimulating discussions.

I. Introduction

In classical hypothesis testing a null hypothesis is posed against an alternative. A critical significance level (typically .05) is chosen in advance. If a preselected test statistic is observed to fall within a prespecified range that under the null has a probability equal to the critical significance level then the null hypothesis is rejected in favor of the alternative. If the test statistic is observed to fall in a range that has a probability greater than the critical value, the null hypothesis "fails to be rejected"; our confidence in the correctness of the null hypothesis is increased given that the data do not speak strongly against it.

Under the null, the distribution of test statistics is known. Their cumulative distributions are given by the marginal significance levels associated with them: a test statistic has an 8 percent chance of falling below the value of the .08 significance level. Our knowledge of the distribution of test statistics allows as to examine a set of tests that fail to reject null hypotheses and ask the following question: does the distribution of these test statistics conform to what we would expect if a prespecified fraction of the null hypotheses were in fact true?

In other words, we can use the results of classical hypothesis tests to in turn perform a classical test of the hypothesis that a fraction π of null hypotheses are true. We perform such a test, and find that we can reject at the .05 level the null hypothesis that more than about one-third of unrejected null hypotheses--more

than one-eleventh of all tested hypotheses--are true. Moreover, our point estimate is that none of the unrejected nulls is true.

We consider three prime contenders as explanations for this finding. First, "So what? We know that all hypotheses are false--they are only approximations"; second "So what? We all know that data mining negates the classical distribution of the t-statistic, so why be surprised that test statistics fail to conform to the distribution classical theory claims they should have?"; third "Hmm! Perhaps papers that fail to reject their null hypotheses survive referees and get published if and only if the null hypotheses they test are false." We are sympathetic to all three views, and all three have significant implications for the way economists do and evaluate empirical research. We, however, argue that the last provides the most important explanation for our findings.

Section II outlines our basic approach to determining the fraction of unrejected null hypotheses which are, in fact, false. Section III presents our data. The principal findings are described in section IV. Section V assesses the three potential explanations for our findings. Section VI summarizes our conclusions, and poses the peculiar dilemma our findings pose for lines of empirical research which rely on failures to reject null hypotheses as confirmatory evidence.

II. Our Approach

Most empirical work in economics tests a null hypothesis:

(1) H_0: $f(\alpha) = 1$

against an alternative:

(2) H_1: $f(\alpha) < 1$

where α is a vector of parameters generating the data, and $f()$ is some function, which we without loss of generality take to be the marginal significance level of the calculated test statistic, and thus to be uniformly distributed over [0, 1]. An estimate a of α is obtained, and analyzed using some pre-chosen critical level x (typically set at .05). If:

(3) $f(a) \leq x$

then we say that we reject the null hypothesis H_0—conclude provisionally that it is false—in favor of the alternative hypothesis H_1, which we provisionally conclude is true. If:

(4) $f(a) > x$

we then say that we have failed to reject the null hypothesis H_0--and our confidence in H_0 is increased.

As we all learned in our first statistics class, such a decision procedure is subject to two types of errors. First, we can erroneously reject a true null hypothesis H_0 because an unlikely realization of the underlying random process has led to a low value of $f(a)$. It has become standard to set the critical value x at .05, so that when the null hypothesis H_0 is true such type II errors occur only 5% of the time--the size of the test is 5%.

Second, we can erroneously fail to reject a false null hypothesis H_0 when the alternative H_1 is in fact true. As a rule the critical value x is not adjusted for the probability of such type I errors. Typically, the alternative hypothesis H_1 is diffuse and the test statistic $f(a)$ has a different distribution

for each point in H_1. Calculating the distribution of $f(a)$ under the alternative in order to construct hypothesis tests of a specified high power--and low chance of a type I error--requires, it is argued, more knowledge of the distribution of $f(a)$ under H_1 than the data can provide.

Economists' statistical tests, therefore, typically have a known size of 5% but an unknown power q. There is a tight bound on the chance of a type II error. If the null hypothesis H_0 is true in a fraction π of hypothesis tests, then the fraction of hypothesis tests that produce a type II error--land in the upper right box of figure 1--is $.05\pi$, which must be less than or equal to .05. By contrast, there is no analogous tight bound on the chance of a type I error—of failing to reject a false null hypothesis H_0 and landing in the lower left box of figure 1.

Figure 1
Possible Outcomes

	Fail to Reject H_0	Reject H_0	
H_0 true	.95 π	.05 π	π
H_1 true	(1-q)(1-π)	q(1-π)	(1-π)
	1 - q + (q-.05)π	q + (.05-q)π	

In this paper, we examine a large number of hypothesis tests that have been carried out in the past few years in order to learn about the fraction π of null hypotheses H_0 that are true and about the average power q of economists' hypothesis tests. We conclude

that π is essentially zero: that only a very small fraction of the null hypotheses in published articles are true. Failures to reject nulls are therefore almost always due to lack of power in the test, and not to the truth of the null hypothesis tested.

Recall that if the null hypothesis H_0 is true, the function $f()$ is uniformly distributed over [0, 1] and satisfies:

$$(5) \quad P\{f(a) \geq p\} = 1 - p$$

Under the alternative H_1, $f(a)$ has some unknown cumulative distribution G:

$$(6) \quad P\{f(a) \geq p\} = 1 - G(p)$$

We will assume that the density g(p) under the alternative is decreasing in p, so that [1-G(p)]/[1-p] falls monotonically from 1 at p=0 to g(1) at p=1. This is simply an assumption that the hypothesis test is not a biased test--that the chances of obtaining a value of $f(a)$ below x and rejecting the null are greater under the alternative than under the null. We require that for any significance level we be more likely to reject the null when it is false than when it is true.

We use (5) and (6) to write the unconditional distribution of $f(a)$ in terms of the distribution G, and the unknown fraction π of null hypotheses H_0 that are in fact true:

$$(7) \quad P(f(a) \geq p) = \pi(1-p) + (1-\pi)(1 - G(p))$$

Since the cumulative distribution $G(p) \leq 1$ for all p in [0, 1]:

$$(8) \quad \pi \leq \frac{P(f(a) \geq p)}{1 - p}$$

Equation (8) allows the construction of an upper bound on the fraction π of null hypotheses that are true. For every critical

value p, the fraction of reported test statistics with marginal significance levels at or above p provides us with an estimate of the numerator of (8); if one-half of all null hypotheses tested are true, then at least one-tenth of marginal significance levels $f(a)$ ought to be above 0.8.

The bound (8) is tightest for values of p near one, for their the density g(p) under the alternative is lowest. The bound (8) becomes trivial for values of p fixed near zero: at p=0 equation (8) becomes $\pi \leq 1/1$, which is always satisfied.

It would not be surprising to find that most null hypotheses in economics are false. After all, economists typically develop models which imply that a given parameter is non-zero, and pit it as alternative against the null hypothesis that the parameter is zero. Thus null hypotheses are formulated in such a way that it is intended that they be rejected. And nearly three-quarters of economics articles in our sample reject their central null. Therefore in this paper we concentrate on those hypotheses that the authors concluded were <u>not</u> rejected. We determine the fraction of these <u>unrejected</u> null hypotheses that were in fact false.

<u>III. Data</u>

We collected our data by reading recent issues of major economics journals to find articles in which the central null hypotheses set forth by the authors had not been rejected. We limited ourselves to empirical papers which tested substantive economic hypotheses: thus we did not include tests of the

"exogeneity of the instruments" or other specification tests unless they were the principal test of a substantive economic hypothesis which was the central focus of the paper. Similarly, we did not include tests of whether, for example, the elasticity of labor supply was equal to zero unless a theory posited in the paper suggested this value was of particular interest.

For each paper, we tried to ascertain the most central hypothesis tested and, when in doubt, chose the first test presented. Thus if an author first presented OLS results and then instrumental variables results which she or he argued were to be preferred, we used the IV results; but if the IV results were included merely to show the robustness of the findings, we used the OLS results. Occasionally a test statistic was simply reported as significant or insignificant; we were unable to make use of this information.

As much as possible we tried to conform to the author's sense of what was the single most important or reliable specification. The choice of test was, however, to some extent arbitrary. This raises the possibility of "coding bias": perhaps our judgments of what was the principal hypothesis test of a paper were unduly influenced by our expectations of the results of this project. A better experimental design--one common among psychologists--would have been to have the data coded by assistants who have no point of view about or stake in the outcome of the research project.

We began by examining what we take to be the four principal journals read by American economists: the American Economic Review, Econometrica, the Journal of Political Economy, and the Quarterly Journal of Economics. After examining two years' worth

of Econometrica and uncovering only two articles in which the central null hypothesis failed to be rejected, we substituted the Review of Economics and Statistics for Econometrica. Our data are therefore from the American Economic Review (1984-88), Econometrica (1986-87), the Journal of Political Economy (1984-87), the Quarterly Journal of Economics (1985-88), and the Review of Economics and Statistics (1986-88). The final sample consisted of 94 articles from RESTAT, 81 from the JPE, 73 from the AER, 16 from the QJE, and 12 from ECMA; 78 of the total of 276 central hypothesis tests failed to reject the null at the 0.1 level.

While it was not our objective in collecting these data to analyze the general treatment of hypothesis testing in the economics profession, we did discover some regularities that we think are worth reporting in their own right. First, in the vast majority of cases test statistics significant at the .1 but not the .05 level are treated as significant rejections of null hypotheses--often, but not always, justified by the similarity of results across specifications or by the finding of a "significant" coefficient in a subsequent properly-tuned specification. While this practice does not conform to the teachings of classical statisticians, it may nevertheless be sensible. Since in practice 0.1 appears to be the critical value for rejecting or failing to reject nulls, we treated "unrejected nulls" with marginal significance levels below 0.1 as rejections.

Perhaps the most striking serendipitous finding to us was the scarcity of hypothesis testing in the major journals. In the absence of the RESTAT and the papers and proceedings issue of the AER, papers organized around formal tests of central null

hypotheses would be scarce.

IV. Results

Table 1 presents the distribution of the probabilities associated with the test statistics in the papers we analyzed, along with the implied upper bounds on π, the fraction of null hypotheses which are true. We focus attention on the tighter bounds obtained for values of p fixed near one at 0.9 and 0.8. The conclusions are striking. In our sample, there are no values of $f(a)$ greater than 0.9. One-tenth of $f(a)$ values should fall into the range 0.9-1.0 when the null hypothesis H_0 is true. The implied estimate of π is therefore zero: no null hypotheses are true.

A less extreme estimate comes from examing the fraction of unrejected null hypotheses with $f(a) > 0.8$. Two-ninths of unrejected nulls should fall into this category when the null hypothesis is true; we actually find that only four out of the seventy-eight unrejected nulls (and the two hundred seventy-six nulls tested) do so. This produces a point estimate that 6.5 percent of all tested null hypotheses are true--that 23 percent of unrejected null hypotheses are true.

Table 1
Distribution of Reported Marginal Significance Levels

Marginal Significance Levels	Number of Hypothesis Tests	Estimated Upper Bound on True Nulls/ Total Hypotheses*	Estimated Upper Bound on True Nulls/ Unrejected Nulls*
1.0-0.9	0	0 %	0 %
0.9-0.8	4	6 %	23 %
0.8-0.7	7	11 %	42 %
0.7-0.6	7	13 %	52 %

0.6-0.5	6	14 %	54 %
0.5-0.4	11	17 %	66 %
0.4-0.3	11	20 %	75 %
0.3-0.2	14	23 %	86 %
0.2-0.1	18	28 %	100 %

*Estimated by the ratio of the number of hypotheses with marginal significance levels in this category or higher to the number of hypothesis tests that should fall in this category or higher if all null hypotheses or all unrejected null hypotheses were true.

An alternative way of approaching the issue is to assume that if each null hypothesis is true the events $W = \{a \mid f(a) > 0.9\}$ are independently distributed (there is overlap in the data used in different articles, so the independence axiom is likely to be violated; the sampling distributions derived under this assumption differ from the actual distributions). Under the null $P(W) = P(f(a) > 0.9) = 0.1$; under the alternative $P(W) = P(f(a) > 0.9) \leq 0.1$. We can therefore construct a test of the hypothesis that the unobserved fraction of all null hypotheses that are true is π or greater for any fixed π. If more than twenty-five of the seventy-eight unrejected null hypotheses in the articles in our sample are true, the odds of finding no W events--no cases in which $f(a) > 0.9$--given that $f(a) > 0.1$ are less than .05.

Therefore at conventional levels of significance we can reject the hypothesis that more than 25/78, or a little less than one-third, of the unrejected null hypotheses (or one-twelfth of tested null hypotheses) in our sample are true. (Our review of hypothesis testing suggests that 0.1 is a more conventional level, at least for the central null hypothesis under study; at this significance level we can reject the hypothesis that more than one-quarter of unrejected null hypotheses are true).

Our failure, for $\pi < 1/11$, to reject the null hypothesis that a

fraction π of null hypotheses are true is itself an economic hypothesis. If this article is published in an economics journal, the logic of our argument would imply that this null hypothesis is also false. Without a full-blown Bayesian analysis, we cannot make precise statements about our posterior distribution over the truth or falsity of null hypotheses. It is nevertheless worth pointing out that our test does have substantial power: if more than five of the seventy-eight unrejected null hypotheses were true, we would have less than a fifty-fifty chance of finding none with $f(a) > 0.9$.

V. Discussion

A rational Bayesian would use our result to draw what seem to be paradoxical inferences. On reading in a leading economics journal an article in which the central null hypothesis H_0 was not rejected, she or he would note that the sample data themselves did not appear to speak strongly against the null hypothesis. But she or he would also note that the experiment itself was drawn from a larger population--that of the subject matter of published economics articles--in which the null hypothesis is almost never true. This prior population information--that almost all null hypotheses are false--would dominate the posterior evaluation. If in a state of relative ignorance before reading the article, after finishing she or he would be highly confident that the null hypothesis under discussion was false even though the author of the article has failed to reject and provisionally concluded that the null hypothesis is true.

One possible response to this paradox is to say that this was something economists knew all along: all null hypotheses are false, because all null hypotheses are simple shorthand descriptions of a complex world. The key question instead is whether a null hypothesis is "good enough" for empirical work: whether the deviations between the null and the real world are sufficiently small as to make conclusions reached conditional on the null reliable guides to the world or are economically significant.

There is a good deal to this argument. It is essentially an argument against hypothesis tests and for confidence intervals--economists should report not whether or not they can reject the null but whether or not their confidence interval excludes (a) economically-insignificant values or (b) economically-significant values. With this we agree, and we think that Edward Leamer's (1978) and Donald McCloskey's (1987) arguments for reorienting the rhetoric of economics toward focusing on confidence intervals have the truth on their side. Reports of empirical work should present the map the data generate from priors to posteriors, and so should focus on confidence intervals and on the sensitivity of the results to small changes in specification (as in Leamer (1983), Leamer and Leonard, 1983) even if they do not present their results within a full-blown Bayesian framework (see Zellner, 1971).

It should, however, be noted that for the most part economists do not act as if they know that their hypotheses are false and are merely seeking to establish their quality as approximations. The practice of econometrics suggests that economists take their

hypotheses seriously. As one example, recall that the "unit root" literature has seen a great deal of effort devoted to determining the asymptotic distribution of test statistics under and testing the null hypothesis that the coefficients in a univariate autoregressive model of U.S. GNP summed to exactly one. Such a focus on the exact implications of a lower-dimensional subspace of possible parameter values for test statistics is difficult to understand if the null is viewed as only an approximation.[1]

In any event, the fact that all hypotheses are mere approximations does not completely account for our results. Economics articles are sprinkled with very low t-statistics--values of $f(a)$ very close to one--on nuisance coefficients. Very low t-statistics appear when the null hypothesis tested is a subsidiary one from the standpoint of the main thrust of the paper. Very low t-statistics appear to be systematically absent--and therefore null hypotheses are overwhelmingly false--only when the universe of null hypotheses considered are the central themes of published economics articles.

This suggests, to us, a publication-bias explanation of our finding. What makes a journal editor choose to publish an article which fails to reject its central null hypothesis, which produces a value of $f(a) > 0.1$ for its central hypothesis test? The paper must excite the editor's interest along some dimension, and it seems to us that the most likely dimension is that the paper is in apparent contradiction to earlier work on the same topic: either

[1]In fact, Christiano and Eichenbaum (1989) argue that the entire literature is badly posed because it has focused on whether or not processes contain a "unit root"; they suggest that this issue is seen as unimportant once one recognizes that the "implications of a broad class of dynamic models are reasonably robust to whether the forcing variables... are modeled as trend or difference stationary."

others working along the same line have in the past rejected the same null, or because theory or conventional wisdom suggests a significant relation.

When will there have been earlier papers along the same lines that rejected the null, or strong theoretical arguments that the null is false? When the null hypothesis is in fact false. Authors therefore face a catch-22: papers that fail to reject their central null hypothesis will be published only when editors think they are especially interesting, but editors will think that they are especially interesting only when the null hypothesis that they test really is false. Our paper can be interpreted as arguing that this social screening device is in fact quite powerful, so powerful that at most a very small proportion of failures to reject a null hypothesis can be taken at face value.

Yet another alternative explanation of our results is that we have ignored another well-known fact: applied econometricians do not follow classical procedures, therefore t-statistics are misleading and reported marginal significance levels incorrect. Most of us suspect that most empirical researchers engage consciously or unconsciously in data mining. Researchers share a small number of common data sets; they are therefore aware of regularities in the data even if they do not actively search for the "best" specification. There seems to be no practical way of establishing correct standard errors when researchers have prior knowledge of the data, or when they report only their favorite results--the distribution of the highest of ten t-statistics is not well known.[1]

[1]Especially sobering is the ease with which Hendry (1980) uses spurious variables to generate close within-sample fits and accurate beyond-sample

One possible reaction is to adjust standard errors by some multiplicative factor which "compensates" for this abuse of classical procedures. Along these lines, we can use our data to ask the question: "by what factor would we have to divide reported t-statistics so that one-ninth of unrejected nulls would exhibit a marginal significance level of .9 or more? The answer is about 5.5. The t-statistic of 2 rule of thumb would then suggest that only unadjusted t-statistics of 11 or more should be taken seriously, in which case hypothesis testing--especially in macroeconomics--would become largely uninformative, and empirical work would play only a very minor role in determining the theories that economists believe. Some claim that at present empirical work does play a very minor role in determining the theories that economists believe (see McCloskey, 1987).

While we have sympathy with this reaction--and neither of us takes reported t-statistics at face value--we do not think that this is ultimately the proper road to take. While we readily believe that researchers data-mine to produce t-statistics above 1.64 or below 1.96, we see little reason to expect this bias to permeate results well outside of this range. Our skepticism is perhaps enhanced by the nihilistic implications regarding the role of empirical work should we set the required level of significance at an unadjusted t-statistic of 11.

VI. Conclusions

At the simplest level our findings reinforce previous calls for

---

predictions.

economists to concentrate on the magnitudes of coefficients, and to report confidence levels and not significance tests. If all or almost all null hypotheses are false, there is little point in concentrating on whether or not an estimate is distinguishable from its predicted value under the null. Instead, we wish to cast light on what models are good approximations--which requires that we know ranges of parameter values which are excluded by empirical estimates.

It appears to us that a number of researchers have implicitly taken the view that explicit testing of hypotheses convinces no one, preferring to develop a "persuasive collage" of evidence. They attempt to establish a set of empirical regularities and interpret them as favorable or unfavorable to a substantive economic hypothesis. While we have some sympathy with this view, we nevertheless believe that there is a role for hypothesis testing because of the discipline in places on argument. But hypothesis test should concentrate on implications that are robust to minor changes in specification. And the key question should not be: can I reject zero? Instead it should be: can I reject all small (or all large) values for this parameter?

Our findings also pose a very peculiar epistemological problem for those inter-related literatures which have relied heavily on the failure to reject point nulls--tests of efficient markets, of the effects of anticipated variables, and of unit roots. These three literatures account for about one-third of the unrejected null hypotheses in our sample. A rational Bayesian, however, reading each each paper that fails to find effects of anticipated money concludes that previous work has given the profession has

strong priors that anticipated money has effects and is more convinced that anticipated money does have effects, and reading each paper that fails to find profitable trading rules is more convinced that such profitable trading rules exist. How can one do convincing empirical work in support of these null hypotheses if each published paper that fails to reject the central nulls only provides evidence to rational readers that they are false?

REFERENCES

Lawrence Christiano and Martin Eichenbaum (1989), "Unit Roots in Real GNP: Do We Know, and Do We Care?" (Cambridge: NBER Wkg. Paper 3130).

David Hendry (1980), "Econometrics: Alchemy or Science?" Economica 47, 188 (November): 387-406.

Edward Leamer and Herman Leonard (1983), "Reporting the Fragility of Regression Estimates," Review of Economics and Statistics 65 (May): 306-17.

Edward Leamer (1978), Specification Searches (New York: John Wiley and Sons).

Edward Leamer (1983), "Let's Take the Con Out of Econometrics," American Economic Review 73 (March): 31-43.

Donald McCloskey (1987), The Rhetoric of Economics (Minneapolis: University of Minnesota Press).

Arnold Zellner (1971), An Introduction to Bayesian Inference in Econometrics (New York: John Wiley and Sons).