Misinterpreting a Failure to Disconfirm as a Confirmation: A Recurrent Misreading of Significance Tests

Thomas Mayer*

It is an old principle of significance testing that failure to disconfirm a hypothesis does not imply that this hypothesis is true. The inability to reject it could be due to the small size of the sample or to the high variance of the data. But, as shown below (see section IV), this principle is often violated, if not explicitly, then by implication. Thus the results may be stated correctly as "the data do not reject the hypothesis". But since the author then proceeds to write (particularly in the concluding section) as though his or her hypothesis had been confirmed the reader is left with the impression that the significance test has confirmed it. Essentially, the problem arises from confusing the correct procedure of using as the null the proposition that the maintained hypothesis is false with the incorrect procedure of using as the null the proposition that the maintained hypothesis is true.

Presumably, one reason why the principle that failure to disconfirm does not imply confirmation of the hypothesis is because it seems to conflict with the widely accepted methodological rule that the only way a scientific hypothesis can ever be confirmed is by repeated failure to disconfirm it. Another reason may be the strong wish to obtain an unequivocal conclusion, even when the correct interpretation of the significance test only allows one to say the data do not reject the hypothesis outright, that they leave open the

possibility that it may be correct. (Cf. Leamer, 1978, p. iv)

This paper tries to resuscitate the principle that failure to disconfirm does not imply confirmation, first by showing that violating it creates serious problems, and second by showing that the seeming conflict between it and the above-cited methodological rule disappears when the term "disconfirmed" is defined in a consistent way.

I. A Limit on the Interpretation of Disconfirmation.

Suppose that in testing the natural rate hypothesis, i.e., that the long-run aggregate supply curve is vertical, the regression coefficient of the expected inflation term, instead of being the predicted 1.0, is 0.7 with a standard error of 0.2. Is it legitimate to say that the data have not disconfirmed the hypothesis? Is it also legitimate to go further and claim that the data corroborate the hypothesis, so that someone who previously attributed, say a 50 percent probability to its validity should now attribute a greater probability to it? Many (most?) economists would probably reply "yes" to both questions.

But both replies are problematic. First, if the natural rate hypothesis is true then the probability of obtaining (on a one-tailed test) purely because of sampling error, a t value of 1.5 for the difference between the predicted and actual values of the coefficient is only 7 percent. While this test therefore has not disconfirmed the natural rate hypothesis at the usual 5 percent level, it has shown that this hypothesis is unlikely to be true.

More generally, treating the failure to disconfirm at the 5 percent level as a confirmation of the maintained hypothesis creates several problems. One is that it

generates the following paradox. Suppose an economist tests the hypothesis that x=y+2, and finds that, although in her data set x=5.0 and y=6, one cannot reject at the 5 percent level the null hypothesis that this difference between the predicted value of y of x+2 and the estimated x-1 is only due to sampling error. She therefore concludes that the data do not disconfirm confirm her hypothesis, and thus provide confirming evidence for it. Her brother tests the contrary hypothesis that x=y-2, and since he uses the same data set, also finds that x=5.0 and y=6. Since the difference between his predicted and estimated coefficients is also not significant, he too, concludes that the data have confirmed his hypothesis. Who is right? (Cf. Mirowski, 2001, p. 195) In this example with its two conflicting mull hypotheses the problem created by treating both of them as confirmed merely because they were not disconfirmed is apparent. But if it is wrong to treat failure to disconfirm as a confirmation in this case were both nulls could be formulated, it must also be wrong to do so in the more frequent case where there is only one null available to be tested

Another paradox arises if a hypothesis is tested more than once. Suppose that on the first test an estimated coefficient that the hypothesis implies is zero, is positive with a t value of 1.5. Suppose further that on a second test using a different and independent data set it is again positive with a t value of 1.4. If failure to reject is interpreted as confirmation, then the second test must be treated as strengthening the plausibility of the hypothesis, since two tests have now failed to reject it. But the correct message of the

second test is just the opposite. The hypothesis has now been rejected at the 5 percent level, since the probability of sampling errors producing coefficients that differ from the predicted value in the same direction with t's as high as1.4 and 1.5 in two independent samples is less than 5 percent.

Third, if one treats the failure to reject at the 5 percent level

as a confirmation it is tempting to treat a failure to reject at the 1 percent level as even stronger confirmation. But that is wrong when significance tests are used incorrectly to indicate confirmation instead of just disconfirmation. It amounts to rejecting a hypothesis only if, due to sampling error, there is a 99 percent probability that it is wrong, rather than rejecting it if there is as much as a 5 percent probability that it is wrong.

Another situation in which the misinterpretation of significance tests can lead to serious error arises when one of the coefficients of a regression has the wrong sign, but is insignificant at the 5 percent level. It is then tempting to dismiss the wrong sign as due merely to sampling error, and to continue to use this regression as a building block for the maintained hypothesis. But suppose the coefficient is significant at, say the 12 percent level. Since it is then unlikely that a sampling error would result in such a large coefficient with the wrong sign, the wrong sign -- even though insignificant at the 5 percent level -- is a warning that the regression may be misspecified.

The error that results from the tendency to accept hypotheses because they cannot be rejected at the 5 percent level also arises in connection with a variety of econometric

procedures, such as an adjustment for heteroscedasticity. The standard procedure is to make such an adjustments only if the assumption that the data are normally distributed can he rejected at the 5 percent level. But what justifies acting on the assumption that the errors are normally distributed unless it can be shown that there is a less than 5 percent probability that they are not? [i] If the estimation error that results from failure to adjust for heteroscedasticity when the data are heteroscedastic is equal to the estimation error that results from making such an adjustment when the data are normally distributed, then a 50 percent probability level seems preferable to a 5 percent level. Whether, in fact, these two estimation errors are equal, or if not, which one is greater, presumably depends on the specific adjustment for heterogeneity that is made. Other examples are tests for cointegration and for Granger causality. If Granger tests show that one-way causality cannot be rejected at the 5 percent level, that does not justify acting as though two-way causality can be ruled out. A similar problem occurs with the common practice of truncating the number of lags on the basis of the Akaike AIC criterion or some similar criterion. Given the absence of a practical alternative the use of such a criterion may be unavoidable, but that does not enhance the credibility of the results obtained that way. Testing for robustness seems appropriate[ii]

II. The Limited Role of Significance Tests

All the above difficulties result from the failure to respect the limited role of significance tests. Strictly speaking, these tests by themselves are not intended to answer the question

whether a hypothesis is correct or incorrect. Their only function is to measure the probability that, assuming the truth of a certain hypothesis, random sampling errors could account for the difference between the estimated value of a coefficient and the value predicted by the hypothesis. (see Chow, 1996, p. 188.) One can then use this information, along with other information, such as the plausibility of the coefficient's sign and its magnitude (see McCloskey, 1985, Chapter 8; McCloskey and Ziliak, 1996), to decide whether to admit the hypothesis into the corpus of verified knowledge. For that we rightly set a high hurdle; we will admit it only if (among other things) it is consistent with the data, and if this consistency is highly unlikely to be due merely to sampling error. The conventional 5 percent hurdle is a severe one, intentionally set to reject the hypothesis in doubtful cases, and thus to generate more Type II than Type I errors. This bias against new hypotheses presumably reflects an asymmetric loss function; with so many hypotheses clamoring for admittance, a Type II error causes less damage than does a Type I error.[iii] Returning to the initial example, to say that the hypothesis of a vertical Phillips curve is confirmed because there is a greater than 5 percent probability that it is not wrong, is implicitly to reverse this decision about Type I and Type II errors. Such an upside down use of significance tests occurs whenever we say or imply that a hypothesis has been confirmed because it has not been rejected at the 5 percent level.[iv]

III. The Definition of " Disconfirmation"

The bias in favor of tolerating a Type II error rather than a Type I error means that in

the context of significance tests the phrase "failure to disconfirm" is defined in a special sense. It is a sense that treats as "not disconfirmed" many cases, for example cases of t= 1.9, which in the ordinary usage of that phrase would be treated as "probably disconfirmed". It follows that the way the phrase "not disconfirmed" is used in the context of significance tests differs sharply from the way it is used in the general methodological rule that the only way to confirm an empirical hypothesis is to subject it to various hard tests, and to show that none of them disconfirm it.[v] For this purpose "not disconfirmed" has to be defined in a way that avoids the bias that inheres in using significance tests to confirm the maintained hypothesis. Suppose, for example, that the t value of the difference between the predicted and the estimated coefficients is, say 1.5, so that (with a one-tailed test) there is a 7 percent probability that sampling error can account for the observed difference. In the context of the methodological rule, unlike in the context of significance tests, this would count, if anything, as a disconfirmation. What does count towards confirmation for the methodological rule is a case in which the difference between the predicted and the estimated coefficients is so small relative to the variance that it could easily be due to sampling error. If such a result is found on many independent tests, then one can treat the hypothesis as confirmed.

For such cases there is no convention like the 5 percent rule of significance tests, nor are there unequivocal definitions of "easily" or of "many". That may perhaps be as it should be. Confirmation is inevitably a matter of degree and judgment, and explicitly

allowing for a subjective element may be preferable to an arbitrary rule. Or, it may be possible to develop a criterion like the 5 percent convention for determining how small the t values obtained from a given number of tests have to be before we decide to treat the hypothesis as confirmed.

IV. Some Examples of Misused Significant Tests

The following examples show that although it may be well-known in principle, the proposition that failure to disconfirm at the 5 percent level does not imply confirmation at the 5 percent level, practice is something else. A survey of papers in the American Economic Review and the Review of Economics and Statistics in 1999 and 2000 turned up six papers in which the failure of a coefficient to be significant at the 5 percent level is incorrectly interpreted as implying that its true value is zero.[vi] This excludes papers that in a formal sense use significance tests incorrectly, but in which the t value of the relevant coefficient is shown to be low, which, as discussed below, provides a limited degree of confirmation. In one example Robertson (1999, p. 760) reports: "Dramatic movements in the peso could bias the effects ... To evaluate this possibility, I regressed changes in the peso on changes in the U.S. production-worker average wage series ... . I found no significant correlation. This ... suggests that this ... bias in not important." But at most it would suggest this only if the t value of the coefficient is so low that the coefficient's nonzero value could easily be accounted for by sampling error. And Robertson does not tell us whether this is the case.

Viscusi and Hamilton (1999) try to determine the variables (and their relative importance) that underlie the EPA's decisions about which Superfund sites to clear up. Among their variables Viscusi and Hamilton include the size of the currently existing population, as well as the potential future population in the affected area. They find that the currently affected population has insignificant coefficients in all four variants of their regression, and therefore conclude that:

> These results suggest that the presence of current exposed populations to health risks generally does not enter EPA's decision with respect to the stringency of the cleanup. ... Rather than setting more stringent standards with a higher cost per case of cancer for current exposed populations, EPA incurs as high a cost per case of cancer when there are only potential future populations at risk. ... The presence of a risk to people based on current land-use patterns rather than hypothetical future uses did not increase the stringency of the regulation. (Viscusi and Hamilton, 1999, pp. 1017, 1021, 1025.)

But in two of their four regressions the t values of the relevant coefficients are 1.3 and 1.5, which are significant at the 10 percent and 7 percent levels respectively. They should therefore have warned the reader that their conclusion that the EPA fails to take the existing population at risk into account is disconfirmed at the 10 percent level in one of their four regression models, and at the 7 percent level in one other.

Loeb and Page's (2000, p. 394) study of teachers' salaries summarizes previous work on their topic as follows: "Only nine of the sixty teacher salary studies cited ... [in a survey paper] produced wage coefficient estimates that were both positive and statistically significant. One interpretation of the literature is that teacher wages are

unrelated to student outcomes." But if in most of the other 51 studies the coefficient has the right sign and is significant at, say the 20 percent level, that would be plausible evidence that teachers' wages do affect student outcomes. Moreover, on the admittedly strong assumption that the 60 studies are independent and their results normally distributed, 9 of them, that is 15 percent, being significant with the correct sign is more than one would expect if the true value of the wage coefficient were zero.

In estimating the effect of corruption on foreign direct investment (FDI) from different countries Shang-Jin Wei (2000) includes in some regressions a dummy variable for FDI from Japan, the U.K. and the U.S. Since its coefficient "is not significant at the 10% level," he concludes that FDIs from these countries are "just as sensitive ... as FDIs from other source countries." (p. 6.) But the t values of these dummies are very close to unity (1.04, 0.96, 1.06, 1.02 in different regressions), which implies on a one-tailed test (given a normal distribution) that there is only about a one third chance that the true value of the coefficient is zero.

Papell et al. (2000, p. 313) in discussing the stability of the natural rate of unemployment tell us that: "ten of the eleven countries have at most two significant breaks, providing almost equally strong evidence [as their evidence previously given against the hypothesis of no break] that there have been only a few permanent changes in postwar unemployment." But the fact that for additional breaks there is no evidence that is significant at the 5 percent level is not "strong evidence" against such breaks.

McConnell and Perez-Quiros (2000, p. 1467) in their analysis of the decline in the variability of the growth rates of U.S. GDP in the 1980s report that: "in all cases we cannot reject the null of no break and therefore conclude that the variance break is not attributable to a change in the constant and the AR component of the model." But for two of their four test statistics the p value of the coefficients are 0.67 and 0.62, suggesting that the absence of a variance break in the AR components is by no means well established.

V. Alternative Procedures

The previous example of testing the natural rate hypothesis illustrates the difficulty of using a significance test to confirm rather disconfirm a hypothesis. To confirm the hypothesis that the long-run Phillips curve does not have a positive slope would require showing that one can reject at the 5 percent level any positive value large enough to be economically meaningful. The data are not likely to oblige.

In some cases it is possible to obtain an unambiguous answer by reversing the maintained hypothesis. Suppose that there are only two possible values for a coefficient, zero and unity, with the maintained hypothesis predicting the former. Suppose further that the only possible alternative hypothesis predicts a value of zero. If that value can be rejected at the 5 percent level, then the maintained hypothesis is confirmed. (See Fisher. 1935, pp. 18-20.) Such cases where a coefficient can have only one of two values are not common. But they do exist.

A much more widely applicable solution is to abandon the Neyman-Pearson

variant of significance testing, at least in those cases in which the maintained hypothesis is not disconfirmed, in favor of the Fisherian variant. Researchers can then report their p values or confidence intervals, so that they - and their readers can decide from this information, perhaps in combination with prior information, how credible the hypothesis is.[vii] Despite its subjectivity this is preferable to claiming erroneously that the failure of a significance test to disconfirm a hypothesis at the 5 percent level implies that this hypothesis has been confirmed. And it is also better than having the researcher deciding behind the scenes whether the p value warrants advocating the hypothesis.

A possible solution to the problem of whether to adjust for heteroscedasticity and non-stationarity when significance tests do not reject homoscedasticity and stationarity at the 5 percent level is to run at least the more important regressions both with and without these adjustments and report both.

V. Conclusion

Perhaps because of a tendency to use significance tests mechanistically, it is easy to confuse the proposition that such a test does not reject a hypothesis at the 5 percent level with the proposition that it has been confirmed it at the 5 percent level.

If one wants to rely on the principle that it is failure to disconfirm that establishes the credibility of hypotheses, then one has to define such failures much more stringently than is normally done by when looking at a t test, and one needs to look at the p value or the confidence limits obtained from many tests. Since this is usually impractical it may be

appropriate to de-emphasize significance tests, particularly since they guard only against sampling error. (Cf. McCloskey, Chapter 9.) And the frequency with which empirical papers in economics reach conflicting results suggests that sampling errors are not the main source of our difficulties.

ENDNOTES

* E-mail: tommayer@bayarea.net I am indebted for helpful comments to Richard Burdekin and Clinton Greene, and to David Jacks for able research assistance. An earlier version of this paper was presented at the July 2001 meetings of the Western Economic Association.

1. The argument that most variables are distributed normally is open to the objection that while many are normally distributed in natural numbers, many others are log-normally distributed (see Aitchison and Brown, 1957). Moreover, we are dealing here with the distribution of errors, not of observed variables.

2. The elimination of insignificant variables in general-to-specific modelling raises a more complex question because in the Monte Carlo tests of Krolzig and Hendry (2001) and Hoover and Perez (1999, 2000) such modelling succeeded in recovering the data generating process. (For critical evaluations of Hoover and Perez's test see the symposium following their 1999 paper.) Moreover in the context of model selection in which general-to-specific modelling is used eliminating variables might be defended as buying simplicity at the cost of some accuracy.

3. Thus, a biologist, Davis Wolfe (2001, p. 27) writes: "Peer review of grant proposals and publications, along with many other subtler barriers, has been established to prevent one renegade scientist from leading us all over the cliff and into the dreaded Abyss of False Theories," Whether one should act on the presumption that the hypothesis is false depends, of course, also on the loss function.

4. That failure to reject does not imply acceptance has been known for a long time, though economists seem to be less aware of it than psychologists, who in general are much more critical of significance tests than are economists. For discussions by psychologists and sociologists see for instance Rozenblum, 1960, Morrison and Henkel, 1970. and "Open Peer Comment" (1998). Chow (1996 p. 11), who himself defends the use of significance tests, writes "the overall assessment of the ... [null-hypotheses significance test procedure] in psychology is not encouraging. The puzzle is why so many social scientists persist in using the process." He argues persuasively that these criticisms of significance tests are largely due to researchers trying to read too much into them. Frick (1995, p. 32) reports that: "The best-know attitude toward the null hypothesis is that it should never be accepted. A survey of 15 undergraduate research methodology textbooks [in psychology] revealed that 4 did not mention the possibility of accepting the null hypothesis, and 7 claimed that it should never be accepted ..." In the six undergraduate economic and business statistics texts that I looked at none warned about accepting the null merely because it cannot be rejected at the 5 percent level. However, a text by two statisticians (Snedecor and Cochran, 1980, p. 66, italics in the original) points out that: "it is not clear just what should be concluded from a nonsignificant result. A test of significance is most easily taught as a rule for deciding whether to accept or reject the null hypothesis. But the meaning of the word accept requires careful thought. A nonsignificant result does not prove that the null hypothesis is correct - merely that it is tenable."

5. Citing confirming instances does not suffice to verify the truth of a hypothesis that tries to establish a general law. This is due to the problem of induction; however many white swans we observe there is always the possibility that the next swan will be black. The force of this argument is, however, diminished by the fact that disconfirmation, too, can never be conclusive. Every disconfirmation makes use not just of the hypothesis being tested, but also of auxiliary hypotheses, such as that the microscope used for the observations is not defective. And when we obtain a disconfirmation how can we be certain that it is not such an auxiliary hypothesis that is at fault? (For a succinct discussion see D.W. Hands, 2001, pp. 96-99) However, the principle that failure to disconfirm provides more persuasive evidence than do instances in which the hypothesis is confirmed, can be justified on more pragmatic grounds. In many cases it is just too easy to find confirming instances; there are, for example, many instances in which the hypothesis that the election of a Democratic (or Republican) president is followed by a recession.

6. I also looked - without success - for such papers in the Journal of Political Economy, February 2000 to February 2001, and at one paper in the 2001 Quarterly Journal of

Economics. I did not read all papers in the journals that I covered, so I may have missed some misuses of significance tests. I did not include cases where a significance test was used to decide whether to adjust for non-stationarity or for heteroscedasticity. That is not always reported and, in any case, it is such a common practice that it does not need documentation.

7. See Rozenblum (1970). As Rosnow and Rosenthal (1989, p. 1277) remark: "God loves the 0.6 nearly as much as the 0.5." However, as Chow (1996, p. 39) points out, in interpreting the p value one needs to keep in mind that its measurement of the probability of sampling error is contingent on the hypothesis being true.

REFERENCES

Aitchison, J. and Brown, J. (1957) The Lognormal Distribution, Cambridge, Cambridge University Press.

Chow, Sui (1996) Statistical Significance, London, Sage Publishing.

Fisher, Ronald A. (1935) The Design of Experiments, Edinburgh, Oliver and Boyd.

Frick, Robert (1995) "Accepting the Null Hypothesis," Memory and Cognition, 23. (1), 133-8.

Hands, D. Wade (2001) Reflection Without Rules, New York, Cambridge University Press.

Hoover, Kevin and Perez, Stephen (1999) "Data Mining Reconsidered: Encompassing and the General-to-Specific Approach to Specification Search," Econometrics Journal, 2 , (2), 167-91.

Hoover, Kevin and Perez, Stephen (2000) "Three Attitudes towards Data Mining," Journal of Economic Methodology, 7, June, 195-210.

Krolzig, Hans-Martin and Hendry, David (2001) "Computer Automation of General-to-Specific Model Selection Procedures," Journal of Economic Dynamics and Control, 25, 831-866.

Leamer, Edward (1978) Specification Searches: Ad Hoc inferences with Nonexperimental Data,, New York, John Wiley.

Loeb, Suzanna and Page, Marianne (2000) "Examining the Link between Teacher Wages and Student Outcomes; The Importance of Alternative Labor Market Opportunities and Non-Pecuniary Variation," Review of Economics and Statistics 82, August, 393-408

McConnell, Margaret and Gabriel Perez-Quiros (2000), "Output Fluctuations in the United States: What has changed since the early 1980's?" American Economic Review, 90, December, 1464-76.

McCloskey, Deidre (1985) The Rhetoric of Economics, University of Wisonsin Press.

McCloskey, Deirdre and Ziliak, Stephen (1996) "The Standard Error of Regressions," Journal of Economic Literature, 34, March, 97-114.

Mirowsky. Philip (2001) "What Econometrics Can and Cannot Tell us about Historical Actors: Brewing, Betting and Rationality in London, 1822-44," in J. Biddle, J. Davis and S. Medema, Economics Broadly Considered, London, Routledge.

Morrison, Denton and Henkel, Ramon (1970) The Significance Test Controversy, Chicago, Aldine.

Open Peer Comments (1996) Brain and Behavioral Research, vol. 19, June, 188-228

Papell, David, Murray Christian and Ghiblawi, Hala (2000) "The Structure of Unemployment Review of Economics and Statistics, 82, May, 309-15

Robertson, Raymond (2000) "Wage Shocks and North American Labor-Market Integration," American Economic Review, vol. 9, September, 742-64.

Roosnow, Ralph and Rosenthal, Robert (1989) "Statistical Procedures and the Justification of Knowledge in Psychological Science," American Psychologist, 49, October, 1276-84

Rozenblum, William (1960) "The Fallacy of the Null Hypothesis Significance Test," reprinted in Denton Morrison, and Ramon Henkel (1970) The Significance Test Controversy, Chicago, Aldine, 216-230.

Snedecor, George and Cochran, William (1980) Statistical Methods, Ames, Iowa, Iowa State University Press.

Viscusi, W. K. and Hamilton, J. T. (1999)"Are Risk Regulators Rational? Evidence from Hazardous Waste Cleanup Decisions," American Economic Review. 89, September, 210-27.

Wei, Sang-Jin (2000) "How Taxing is Corruption on International Investment?" Review of Economics and Statistics, 82, February, 1-11.

Wolfe, David (2001) "The Empire under the Ground," Wilson Quarterly, 25, Spring, 18-27.