A joint Initiative of Ludwig-Maximilians-Universität and Ifo Institute for Economic Research

CESifo Working Papers

A FREQUENT MISUSE OF SIGNIFICANCE TESTS

Thomas Mayer

CESifo Working Paper No. 549

August 2001

CESifo
Center for Economic Studies & Ifo Institute for Economic Research
Poschingerstr. 5, 81679 Munich, Germany
Phone: +49 (89) 9224-1410 - Fax: +49 (89) 9224-1409
e-mail: office@CESifo.de
ISSN 1617-9595

CESifo *Working Paper No. 549*
August 2001

A FREQUENT MISUSE OF SIGNIFICANCE TESTS

Abstract

Economists sometimes interpret the failure of a significance test to disconfirm a hypothesis as evidence that this hypothesis is valid. Six examples of this are cited from recent journals. But this is a misinterpretation of what significance tests show. While in general it is correct that every failure to disconfirm a hypothesis adds to its credibility, the term "disconfirm" is defined differently for this purpose than it is in the context of significance tests.

Keywords: Significance tests, t values, t coefficients, confirmation.

JEL Classification: C1, B4.

Thomas Mayer
3054 Buena Vista Way
Berkeley, CA 94708
U.S.A.
tommayer@bayarea.net

A FREQUENT MISUSE OF SIGNIFICANCE TESTS:

Thomas Mayer*

Abstract

Economists sometimes interpret the failure of a significance test to disconfirm a hypothesis as evidence that this hypothesis is valid. Six examples of this are cited from recent journals. But this is a misinterpretation of what significance tests show. While in general it is correct that every failure to disconfirm a hypothesis adds to its credibility, the term "disconfirm" is defined differently for this purpose than it is in the context of significance tests.

This paper makes no claim to establish some hitherto unknown result. Instead, it shows that in our everyday thinking and practice we often ignore a principle that presumably many of us learned in our basic statistics course. A like reason for this is that this principle seems to conflict with another widely accepted principle. But this apparent conflict is spurious because a critical term is defined differently in these two principles.

I. Treating a Failure to Disconfirm as Confirmation

Suppose that in testing the hypothesis that the long run aggregate supply curve is vertical the regression coefficient of the lagged 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 if you previously attributed, say a 50 percent probability to its validity you should now attribute a greater probability to it? Most economists would probably answer yes to both questions.

The first of these answers is questionable, while the second is wrong. This is so, because failure to disconfirm a hypothesis (e.g., the long-run Phillips curve being vertical) does not imply that this hypothesis has been confirmed. If it did one could in some cases generate a paradox by first showing that the data do not disconfirm a hypothesis, say that a>b, and then also showing that they do not disconfirm the opposite hypothesis that a<b. (Cf. Mirowski, 2001, p. 195)

Perhaps economists have lost sight of this principle for two reasons. One is the familiar (Popperian) principle that the only way the empirical validity of a hypothesis can tentatively be established is by the repeated failure of relevant tests to disconfirm it[1]. But, as discussed below, this principle does not apply here because it defines "failure to disconfirm" very differently from the way it is defined in the context of significance tests.

The second reason is a misinterpretation of what significance tests do. They are not intended to answer the question whether a hypothesis is correct or incorrect. Instead, their function is to help decide whether we should admit this hypothesis into the corpus of verified knowledge. For that we rightly set a high hurdle; we will admit it only if it is consistent with the data, and if this consistency is highly unlikely - usually interpreted to mean a less than 5 percent probability - to be due merely to sampling error. The 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.[2]

But in the above example to say that the hypothesis of a vertical Phillips curve is corroborated 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 is not just a peculiarity of the particular example with which I started the paper. It occurs whenever we say or imply that a hypothesis has been confirmed because it has not been rejected at the 5 percent level. And as shown below, this is done with some frequency.[3]

Put another way, significance tests themselves do not test hypotheses, but only show what the probability is that, given a certain hypothesis, such as the null hypothesis, random sampling errors account for the difference between the predicted and the actual values of a coefficient (see Chow, 1996, p. 188. Together with a conventional (and essentially arbitrary) cut-off point, such as a 5 percent probability level, they allow us in some cases to reject the hypothesis.[4] But if they do not allow us to reject the hypothesis then they also do not tell us that it is true. But how about the general principle, that if a test fails to disconfirm a hypothesis it enhances its credibility at least to some extent? This is correct if these tests are hard and not soft tests, so that the statement "fails to disconfirm" is interpreted correctly. Suppose, for example, that the t value of the difference between the predicted and the estimated coefficients is, say 1.5. This is conventionally interpreted as a failure to reject, and hence presented as a confirmation of the hypothesis. Often this is done implicitly: the author merely states that "the data do not reject" the hypothesis, and leaves it to the reader to carry away the impression that it has been confirmed. But with a t value of 1.5 the probability that (on a one-tailed test) a sampling error accounts for a difference between the predicted and observed value of the coefficient as great as 1.5 standard errors is only 7 percent, so that the results of this test should be counted as weak evidence against the hypothesis instead of a confirmation.

II. Implications

Ignoring the limitation on what can be inferred from a failure to disconfirm at the 5 percent level has some undesirable consequences. First, and most obviously, many hypotheses are accepted that should not be. Second, the results of multiple tests can easily be misinterpreted. Suppose that on the first test a coefficient which the maintained hypothesis implies is zero has a t value of 1.5. Suppose that a second test using another data set produces the same result. The way significance tests are widely interpreted the second test is read as enhancing the plausibility of the hypothesis since two tests have now failed to reject it. But the correct message of the second test is that the hypothesis has now been rejected at the 5 percent level, since the probability of a sampling error as great as 1.5 standard errors on both of these tests is only 0.5 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 in the upside-down way discussed here. 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 is 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 should not be brushed aside.

None of this denies that significance tests are useful when they are applied the right way round,that is when they are interpreted so that the maintained hypothesis is treated as not confirmed unless the probability that sampling error accounts for the observation (e.g., a regression coefficient exceeding zero) is less than 5 percent. Even then, however, they are subject to many criticisms that have been leveled mainly by psychologists and sociologists.[5] Also, as McCloskey and Ziliak (1996)emphasize, one needs to pay attention to the economic and not just the statistical significance of a coefficient.

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 adjustments for heteroscedasticity. The standard procedure is to make such 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?[6] 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 if one makes 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 is greater, presumably depends on the specific adjustment for heterogeneity that is made.

Such a problem also problem exists with respect to tests for nonstationarity and cointegration. Similarly, if a Granger causality test shows 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.

III. Examples

The following examples show that although the proposition that failure to disconfirm at the 5 percent level does not imply confirmation at the 5 percent level may be well known in principle, 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.[7] This excludes papers that in a formal sense use significance tests incorrectly, but in which the t value of the relevant coefficient is low, so that there is some justification for arguing that the coefficient can perhaps be treated as though its true value is zero.

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. ... 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 is at the margin 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 40 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 dummy 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. Failure to be significant at the 5 percent level could, for example, be due to the size of the sample.

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.

IV. Alternative Procedures

If a hypothesis has not been rejected at the 5 percent level it is often desirable to look at the probability that this is due, not to the validity of the hypothesis, but to the sample being too small, (given the variance of the data) to provide anything but an ambiguous answer. In some cases it may be possible to obtain an unambiguous answer by testing and rejecting the null hypothesis. But that requires the often unattainable condition that the null can be stated precisely (see Fisher, 1935, pp. 18-20).

A different, and 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, in combination with prior information, how credible the hypothesis is.[8] 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 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 do at least the more important regressions both with and without adjustments. If they differ substantively, then both sets of results should be reported.

V. Summary

Perhaps because of a tendency to use significance tests mechanically, 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 confirmed it at the 5 percent level, As several examples demonstrate, this error has occurred in various papers. Hence, when their significance tests do not reject their hypothesis at the 5 percent level researchers should state the p values of their maintained hypotheses.

ENDNOTES

* University of California, Davis. E-mail: tommayer@bayarea.net I am indebted to David Jacks for able research assistance.

1. It could also be due to a natural tendency to do some things that are necessary to obtain a clear result, even if it means cutting corners (see Leamer, 1978, p. iv).

2. 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.

3. That failure to reject does not imply acceptance has been known for a long time, though economists seem to show less awareness of it than do psychologists and sociologists, who tend to frame the discussion in terms of the acceptance of the null hypothesis. (For discussions by psychologists and sociologists see Rozenblum, 1960, Morrison and Henkel, 1970. and "Open Peer Comment" (1998). Chow. a psychologist, (1896, p. ix) remarks that: "although the null hypothesis significance-test procedure ... is an integral component of data analysis in empirical research, many researchers have reservations about its validity or utility." Frick (1995, p. 132) 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 revealed that 4 did not mention the possibility of accepting the null hypothesis, and 7 claimed that it should never be accepted ..." In the half dozen undergraduate economic and business statistics texts that I looked at none warned about accepting the null. 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.

4. The problem of theory choice is extremely complex, and significance tests play only a modest role in it. It may sometimes be reasonable to reject a hypothesis that is confirmed at the 5 percent level in favor of one that is not.

5. Thus 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." Chow argues persuasively that these criticisms of significance tests are largely due to researchers trying to read too much into them.

6. 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).

7. 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.

8. 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 it measures the probability of sampling error contingent on the hypothesis being true.

REFERENCES

Aichison, J and Brown (1937) *The Lognormal Distribution*, Cambridge, Cambridge University Press.

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

Fisher, Ronald A. (1935) *The design of Experiments*. Endinburg, Oliver and Boyd.

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

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, pp. 393-408

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

McCloskey, Deirdre and Ziliak, Stephen (1996) "The Standard Error of Regressions," *Journal of Economic Literature,* March, pp. 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), *Journal of Brain and Behavioral Research*, vol. ?/, ?/, pp.

Papell,David, Murray, Christian and Ghiblawi, Hala (2000) *Review of Economics and Statistics*, 82, May, pp. 309-15

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

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

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, pp. 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, pp. 210-27

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

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