Significance Tests Harm Progress in Forecasting

J. Scott Armstrong
The Wharton School, University of Pennsylvania
Philadelphia, PA 19104

E-mail: armstrong@wharton.upenn.edu

This paper was published in the International Journal of Forecasting, 23 (2007), 321-327.

Abstract

Based on a summary of prior literature, I conclude that tests of statistical significance harm scientific progress. Efforts to find exceptions to this conclusion have, to date, turned up none. Even when done correctly, significance tests are dangerous. I show that summaries of scientific research do not require tests of statistical significance. I illustrate the dangers of significance tests by examining an application to the M3-Competition. Although the authors of that reanalysis conducted a proper series of statistical tests, they suggest that the original M3 was not justified in concluding that combined forecasts reduce errors and that the selection of the best method is dependent upon the selection of a proper error measure. I show that the original conclusions were justified and that they are correct. Authors should try to avoid tests of statistical significance, journals should discourage them, and readers should ignore them. Instead, to analyze and communicate findings from empirical studies, one should use effect sizes, confidence intervals, replications/extensions, and meta-analyses.

Key words: accuracy measures, combining forecasts, confidence intervals, effect size, M-competition, meta-analysis, null hypothesis, practical significance, replications.

Tests of statistical significance became popular in the social sciences only in the last century. While only about 10% of empirical papers published in psychology journals included such tests in the early 1900s, this grew to over 90% by 1970 (Hubbard & Ryan 2000). In contrast, as noted by Schmidt and Hunter (1997), statistical significance is seldom used in the physical sciences such as physics and chemistry and many scientists in these areas regard its use as unscientific.

Statistical significance in commonly used in empirical research on forecasting, as can be seen by picking up any copy of a forecasting journal. For example, I coded the 12 papers in the first issue of the 2005 volume of the *International Journal of Forecasting.* Of these, 11 used tests of statistical significance. (I was an author in one of these.)

Despite its popularity, many researchers have warned against the use of tests of statistical significance. Schmidt and Hunter (1997, p. 38) concluded, " . . . reliance on significance testing is logically indefensible and retards the research enterprise by making it difficult to develop cumulative knowledge."

In this paper, I draw upon prior literature to summarize problems with statistical significance testing. I then demonstrate that such tests are unnecessary when summarizing scientific evidence. Finally, I apply these concepts to a recent study that applied statistical significance testing to the M3-competition.

Problems with Statistical Significance

Researchers have argued that tests of statistical significance are confusing and misleading since 1931 at least. The criticisms have been more common in recent years. They find that

- researchers publish faulty interpretations of statistical significance in leading economics journals (McCloskey and Ziliak 1996), in psychology (Cohen 1988), and other areas (Hubbard & Bayarri 2003).
- journal reviewers misinterpret statistical significance (e.g., see the experiment by Atkinson, Furlong, and Wampold 1982).
- general readers misinterpret tests of statistical significance even after training; for example, my experience is that few students in MBA programs are able to properly interpret tests of statistical significance.

The confusion harms the development of science in a number of ways. One is that there is a bias against publishing papers that fail to reject the null hypothesis (Hubbard & Armstrong 1992). Papers that fail to reject null hypotheses might contain important findings.

Another problem arises from the selection of null hypotheses. A null hypothesis is generally selected for its simplicity, not for its truth or for its importance to decision making. For example, in trying to forecast the

effect of prices of automobiles on automobile sales, one would test the null hypothesis that price is unrelated to sales. By finding that the results reject the null hypothesis, one learns nothing new. Cohen (1988) refers to such meaningless null hypotheses as "nil hypotheses." It is possible to develop meaningful null hypotheses under some conditions; however, as explained below, this is unnecessary.

Still another problem is that statistical significance testing distracts attention from key issues. It leads researchers to think they have completed the analysis even though much remains to be done. The focus should be on assessing effect size (e.g., an estimate of practical significance, such as how much improvement in accuracy is provided by a given method under certain conditions). Effect sizes are important for the advancement of science. For example, when scientists tried to estimate the distance from the earth to the sun, they presented their estimates and confidence intervals, not tests of statistical significance. Forecasting practitioners also need effect sizes and prediction intervals.

Here is an example of how statistical significance has misled researchers. In a study on predicting how nonrespondents would have replied to mail surveys, Wright and Armstrong (2006) found that the use of statistical significance testing led researchers to conclude that adjustment for non-response bias was unnecessary in 92% of the 38 studies in our sample. In fact, none of the studies in the sample made an adjustment for non-response bias. Judging from the original study, we estimate that non-response bias should have been predicted and adjusted for 111 of the variables in those studies.

Significance tests can also harm practice. In their simulation study, Goodwin and Lawton (2003), found that the use of significance tests harmed accuracy in comparison with a rule that one should correct all judgmental forecasts to remove systematic bias.

The above discussion of problems looks to be one-sided. Are there good reasons for using significance tests? I have circulated working drafts of this paper and received peer review from nine people. In addition, three anonymous reviewers have commented. I made further attempts at gaining peer review by posting drafts on my homepage and by soliciting feedback for academic email lists. Schmidt and Hunter (1997) had conducted a much more extensive search. In doing so, they examined objections to the conclusion that significance testing should be avoided. They discuss and refute the eight most commonly stated objections and found no merit in another 79 objections. These objections were obtained in their three-year effort to have researchers submit cases in which significance tests made a contribution; no cases were found. They concluded, "Statistical significance testing retards the growth of scientific knowledge; it never makes a scientific contribution."

Statistical Significance is Unnecessary for Communicating Scientific Advances

Armstrong (2001a) presents evidence on 139 forecasting principles. The principles would be the same had tests of statistical significance been removed from all of the papers that were used. In my own chapters, I provided meta-analyses for role-playing (p. 26), judgmental bootstrapping (p. 182), and combining (p. 428). Only effect sizes were used. I challenge readers to show how tests of statistical significance would have led to better estimates.

More recently (Armstrong 2006), I reviewed evidence on which methods have been shown to improve the accuracy of forecasts over the past quarter century. The word "significance" does not appear in the paper.

I have spent many years summarizing evidence-based persuasion principles. This involves summarizing evidence on many hundreds of research papers. The word "significance" appears only in the paragraph where I explain why it is unnecessary to report on significance.

What should one do without tests of statistical significance? There are better ways to report findings. To assess—

- importance, use effect sizes.
- confidence, use prediction intervals
- replicability, use replications and extensions.
- generality, use meta-analyses.

Significance Testing of the M3-Competition

The dangers involved with testing statistical significance apply even when the testing is properly done and fully reported. I use a recent application of statistical significance testing in forecasting to illustrate the dangers.

Koning, Franses, Hibon and Stekler (2005), referred to hereafter as KFHS, used tests of statistical significance to support their claim that such tests should be used in forecasting. They examined four conclusions from the M3-Competition (Makridakis and Hibon 2000) that had been reached without testing for statistical significance. By using significance tests, KFHS found that two of the conclusions were in agreement with the original study, while two differed. I examined the conclusions that differed; one relates to combining forecasts, and the other to the selection of an accuracy measure.

Combined forecasts

The authors of the M3-Competition concluded that a combined forecast (of three extrapolation methods) was more accurate because its error was 4.7% less than that from the typical method used to create the

combined forecast. After statistical testing of these results, however, KFHS (which includes one of the original M3-Competition authors) stated " . . . the M3 conclusion about the combination of [forecasts] being statistically superior has not been proven." How might a practicing forecaster interpret this conclusion? I expect that many would conclude from this statement that combining forecasts will not improve accuracy and, since it costs more, that they should not combine forecasts.

The null hypothesis used in KFHS is that combining does not improve accuracy when judged against all of the components. This null hypothesis only of interest to practitioners if they are certain *a priori* which of the methods is most accurate, a highly unusual situation. Furthermore, there may be conditions under which the combined forecast is expected to be more accurate than the most accurate component.

Prior research has been done to address the issue of whether combining is more accurate than the typical component. This is the typical issue facing the forecasters. Armstrong (2001) summarized 30 studies published prior to the M3-competition. Full disclosure on this analysis is provided in Armstrong (2001b, p. 417).

Table: Error Reductions from Combining Ex Ante Forecasts

Study	Methods	Components	Criterion	Data	Situation	Validation Forecasts	Forecast Horizon	Percent error reduction
Levine (1960)	intentions	2	MAPE	annual	capital expenditures	6	1	18.0
Okun (1960)	"	2	"	"	housing starts	6	1	7.0
Landefeld & Seskin (1986)	"	2	MAE	"	plant & equipment	11	1	20.0
Armstrong et al. (2000)	"	4	RAE	"	consumer products	65	varied	5.5
Winkler & Poses (1993)	expert	4	Brier	cross-section	survival of patients	231	varied	12.2
Thorndike (1938)	"	4 to 6	% wrong	"	knowledge questions	30	varied	6.6
Makridakis et al. (1993)	"	5	MAPE	monthly	economic time series	322	1 thru 14	19.0
Richards & Fraser (1977)	"	5	"	annual	company earnings	213	1	8.1
Batchelor & Dua (1995)	"	10	MSE	"	macroeconomic	40	1	16.4
Kaplan et al. (1950)	"	26	% wrong	cross-section	technology events	16	varied	13.0
Zarnowitz (1984)	"	79	RMSE	quarterly	macroeconomic	288	1	10.0
Sanders & Ritzman (1989)	extrapolation	3	MAPE	daily	public warehouse	260	1	15.1
Makridakis & Winkler (1983)	"	5	"	monthly	economic time series	617	18	24.2
Makridakis et al. (1993)	"	5	"	"	"	322	1 thru 14	4.3
Lobo (1992)	"	5	"	quarterly	company earnings	6,560	1 thru 4	13.6
Schnaars (1986)	"	7	"	annual	consumer products	1,412	1 thru 5	20.0
Landefeld & Seskin (1986)	econometric	2	MAE	annual	plant & equipment	7	1	21.0
Clemen & Winkler (1986)	"	4	MAD	quarterly	GNP (real & nominal)	45	1 thru 4	3.4
Shamseldin et al. (1997)	"	5	MAPE	annual	rainfall runoff	22	1	9.4
Lobo (1992)	expert/extrap	2	MAPE		company earnings	6,560	1 thru 4	11.0
Lawrence et al. (1986)	"	3	"	annual monthly	economic time series	1,224	1 thru 18	10.7
Sanders & Ritzman (1989)	"	3	"	daily	public warehouse	260	1	15.5
Lobo & Nair (1990)	"	4	"	annual	company earnings	768	1	6.4
Landefeld & Seskin (1986)	intentions/econ	2	MAE	annual	plant & equipment	11	1	11.5
Vandome (1963)	extrap/econ	2	MAPE	quarterly	macroeconomic	20	1	10.1
Armstrong (1985)	"	2	"	annual	photo sales by country	17	6	4.2
Weinberg (1986)	expert/econ	2	"	cross-section	performing arts	15	varied	12.5
Bessler & Brandt (1981)	exprt/extrap/econ	3	"	quarterly	cattle & chicken prices	48	1	13.6
Fildes (1991)	"	3	MAE	annual	construction	72	1 & 2	8.0
Brandt & Bessler (1983)	"	6	MAPE	quarterly	hog prices	24	1	23.5
							Unweighted average	**12.5**

As can be seen, there are no tests of statistical significance in the exhibit. Evidence for the principles is summarized completely by effect sizes, in this case by "percent of error reduction" in the last column. In all cases, combined forecasts were more accurate than the typical method. The gains were substantial; there was, on average, an error reduction of more than 12%. Sometimes the combined forecast was even more accurate than the best of the components, although this is not noted in the table. Had tests of statistical significance been use in the assessment of value of combining, it might have led to the conclusion that

some of the small sample studies did not have sufficient data to be significant. The only effect here would be to discard useful information. Significance tests could not possibly affect the finding that combining reduces errors when compared to the error of the typical component, nor would they add any useful information.

While the authors of KFHS were probably aware of the prior research, is it unlikely that the readers would be able to make this connection and to also realize that KFHS were testing whether the combined method was better than the *best of the components* not as in the original M3 paper (and in the above exhibit) whether combining is better than the typical method. In fact, the M3 competition yielded a 4.7% error reduction in comparison with the typical component.

Accuracy measures

KFHS wrote, "We discovered that there is a significant agreement among the results obtained from the various accuracy measures that have been used in the evaluations of the various competitions." From this, readers might conclude that when comparing forecasting methods, the choice of an accuracy measure makes little difference.

Again there are problems with the choice of a null hypothesis and the failure to use prior knowledge. For example, the analyses presented in Armstrong and Collopy (1992) showed that the selection of the most accurate extrapolation method varied substantially depending upon the choice of an accuracy measure. This was especially so when the number of series was modest. While that paper mentioned statistical significance on two occasions, we think that its use was harmless. Certainly, it was not needed to reach our conclusion.

Discussion

KFHS concluded that "An important implication . . . is that our analysis demonstrates the necessity for undertaking rigorous statistical testing in any evaluation of forecast accuracy . . . If this is not done, it is possible that one might draw erroneous conclusions from the descriptive statistics." They claimed this was because the conclusions from the statistical analysis differed from those based on descriptive statistics.

In contrast, I argue that tests of statistical significance should never be used when reporting scientific findings. The original M3-competition was unusual in departing from the norm by avoiding tests of statistical significance and it was also proper to do so. When one looks at the evidence to date, combining improves accuracy and the choice of an accuracy measures does matter. The significance tests by KFHS have only contributed to confusion on these issues.

My conclusions relate to the use of statistical significance in the development of knowledge about forecasting. This does not rule out the possibility that statistical significance might help in other areas such as in (1) aiding decision makers by flagging areas that need attention (e.g., quality control charts), (2) as part of a forecasting procedure (e.g., helping to decide whether to apply a seasonality adjustment or when to damp trends), or (3) serving as a guide a scientists who is analyzing a problem (e.g., as a quick way to highlight areas that need further study).

Researchers should avoid tests of statistical significance when reporting their findings. As the journal reviewing system exists, however, researchers feel a need to include tests of significance. I have had discussions with co-authors on this topic and the decision has usually been to include some tests in order to gain favor with reviewers, but to do so in a way that is unlikely to harm the use of the findings.

I join with others (e.g., Shea 1996; Hunter 1997) in suggesting that journals ban or at least discourage tests of statistical significance. The balance of evidence to date supports such a change. The challenge for the discussants is to provide cases where the use of statistical significance has contributed to scientific knowledge. Despite efforts to locate such studies, I am unaware of any.

Conclusions

Readers should be clear about the conclusions from the KFHS analysis of the M3-Competition. First, it did not refute the original M3 conclusion that combined forecasts are more accurate than those from the typical method. Second, it did not refute the conclusion that the choice of an accuracy measure is important when choosing among forecasting methods.

Tests of statistical significance carry dangers when doing research related to the advancement of scientific knowledge about forecasting. To analyze and communicate findings from empirical studies, use effect sizes, confidence intervals, replications/extensions, and meta-analyses.

Practitioners can take immediate action. They should ignore tests of statistical significance. Researchers can take this step when they write books or present papers. When writing for journals, they may have to seek ways of reducing the potential harm of reporting significance tests. They should also omit the word significance; findings that reject the null hypothesis are not significant in the everyday use of the term, and those that reject are not insignificant.

In contrast with the KFHS conclusion on the value of significance testing, I join the ranks of those recommending that significance tests be eliminated from scientific journals. Such tests are difficult to do properly and to interpret correctly. Even when properly done, they do not lend themselves easily to

decision-making. Finally, even if properly done and properly interpreted, significance tests are unnecessary for scientific progress.

References

Armstrong, J. S. (2006). Findings from evidence-based forecasting: Methods for reducing forecast error. *International Journal of Forecasting* (Forthcoming). (In full text under Working Papers at forecastingprinciples.com)

Armstrong, J. S. (2001a). *Principles of Forecasting: A Handbook for Researchers and Practitioners*. Boston: Kluwer Academic Publishers.

Armstrong, J. S. (2001b). Combining forecasts. In J. S. Armstrong (ed.), *Principles of Forecasting*. Boston: Kluwer Academic Publishing.

Armstrong, J. S. & Collopy, F. (1992). Error measures for generalizing about forecasting methods: Empirical comparisons. *International Journal of Forecasting*, 8, 69-80.

Atkinson, D. R., Furlong, M. J. & Wampold, B. E. (1982). Statistical significance, reviewer evaluations, and the scientific process: Is there a statistically significant relationship? *Journal of Counseling Psychology* 29, No. 2, 189-194.

Cohen, J. (1988). The earth is round ($p < .05$). *American Psychologist*, 49, 997-1003.

Goodwin, P. & R. Lawton (2003). Debiasing forecasts: How useful is the unbiasedness test? *International Journal of Forecasting*, 19, 467-475.

Hubbard, R. & Armstrong, J.S. (1992). Are null results becoming and endangered species in marketing? *Marketing Letters*, 3 (April), 127-136

Hubbard, R. & Bayarri, M. J. (2003). Confusion over measures of evidence (*p*'s) versus errors (*a*'s) in classical statistical testing (with comments). *The American Statistician*, 57, (August), 171-182.

Hubbard, R. & Ryan, P. A. (2000). The historical growth of statistical significance testing in psychology – and its future prospects. *Educational and Psychological Measurement*, 60, 661-681.

Hunter, J. E. (1997). Needed: A ban on the significance test. *Psychological Science*, 8, (1), 1-20.

Koning, A. J., Franses, P.H., Hibon, M. & Stekler, H. O. (2005). The M3 competition: Statistical tests of the results. *International Journal of Forecasting*, 21, 397-409.

Makridakis, S. & Hibon, M. (2000). The M3-Competition: Results, conclusions and implications. *International Journal of Forecasting*, 16, 451-476.

McCloskey, D. N. & Ziliak, S. T. (1996). The standard error of regressions. *Journal of Economic Literature*, 34, 97-114.

Schmidt, Frank L. & Hunter, J. E. (1997). Eight common but false objections to the discontinuation of significance testing in the analysis of research data, in Harlow, Lisa L., Mulaik, S. A. & Steiger, J. H. *What if there were no Significance Tests?* London: Lawrence Erlbaun.

Shea, C. (1996). Psychologists debate accuracy of significance test. *The Chronicle of Higher Education*, 42 (August 16), A12 & A17.

Shrout, P. E. (1997), Should significance tests be banned? Introduction to a special section exploring the pros and cons. *Psychological Science*, 8, (1), 1-2 (Special Section follows on pages 3-20).

Wright, M. & Armstrong, J.S. (2006). Verification of citations: Fawlty towers of knowledge? Working paper (available in full text at jscottarmstrong.com)

Acknowledgments: Many people provided useful suggestions. In thanking them, I do not mean to imply that they agree with my positions. They include Kay A. Armstrong, Eric Bradlow, Chris Chatfield, Fred Collopy, Robert Fildes, Kesten Green, Ray Hubbard, Keith Ord, and Tom Yokum. I also solicited suggestions from the four authors of the Koning, et. al paper. Bryan LaFrance and Saurabh Bajpai provided editorial assistance.