Forecasting without significance tests?

Andrey V Kostenko and Rob J Hyndman

November 5, 2008

Abstract

Statistical significance testing has little useful purpose in business forecasting, and other tools are to be preferred. For selecting or ranking forecasting methods (especially those based on models) there exist simple but powerful and practical alternative approaches that are not tests. It is suggested that forecasters place less emphasis on p values and more emphasis on the predictive ability of models.

1 Introduction

In the beginning there was Arbuthnot (1667–1735), and he begat the first statistical test of significance. Nicholas Bernoulli, de Moivre and other wise men saw that it was good, and the process began. Then there was a new generation, the generation of K. Pearson's χ^2-test and Gosset's t-test, Fisher's *significance testing* and Neyman and E. Pearson's *hypothesis testing*. It was good for Fisher, and it was good for Pearson and Neyman, and it was never good for both. A controversy began. Then there were further generations, the generations of knowledgable developers, obedient users and conscientious, if not rebellious, writers shifting the debate to the masses. This is a very brief account of the story of statistical significance, as inspired by Le Cam's (1986) short story of the central limit theorem. Apart from some of the same personalities at the origin, the two have very little in common; while no-one resists the usefulness of the subject of his story, many opine that there is more harm than good coming from significance tests. We will tell you why, and what can be done about it.

2 What's hypothesis testing?

A statistical hypothesis is a guess about the population from which the data were supposedly drawn. There are a number of different approaches to testing statistical hypothesis, and it is useful to understand the difference. For a broader and deeper discussion, see Berger (2003) or Christensen (2005) among many others. Excellent expositions of the views of Fisher on significance testing and of Neyman and Pearson on hypothesis testing are provided by Hall and Selinger (1986) and Mayo (1992), respectively.

In forecasting, hypothesis testing is often used for testing features of a fitted model (tests that one or more coefficients in a regression are zeros, or that residuals are normally distributed and independent). Versions of the Diebold-Mariano (1995) test for equality of predictive accuracy are also commonly used.

2.1 Testing according to Fisher

Consider an example. A friend of yours once mentioned that his last winter sales of product G were 120 units on average. You recently started a similar business, and your weekly sales records are

$\{118, 102, 121, 110, 137, 107, 113, 91, 111, 75, 128, 105\}$, with the sample mean of 109.833 and sample standard deviation of 16.325.

You wonder whether, on average, your weekly sales might be as good as those of your friend. Assuming that the counts follow a Poisson distribution with a large parameter value, you approximate them with a normal distribution. Therefore, your hypothesised model is that weekly sales are independent and normally distributed with mean μ and standard deviation σ, and your null hypothesis is $H_0 : \mu = 120$. You do not have an explicit alternative hypothesis, as all you want to know is how much support do the data have for your hypothesised model with known μ and unknown σ.

To conduct a goodness-of-fit test, you need a random variable called a test statistic, its fully specified distribution under the null hypothesis and its observed value. In this example, you can use a T statistic that follows a Student-t distribution with $n - 1$ degrees of freedom, n being the number of observations. The data are summarised by the observable value of the test statistic $t = \frac{109.833-120}{16.325/\sqrt{12}} = -2.16$. Finally, a p value is computed as the probability of T being at least as extreme as t:

$$p = 2\Pr(T > |t|) = 1 - \Pr(T \geq -2.16) + \Pr(T < -2.16) = 0.054.$$

The test may be considered done at this stage. The outcome of the test is the measure of evidence against the null hypothesis conveyed by the p value, the probability of having something as or more extreme than what you actually have. The smaller this probability, the stronger the evidence against the hypothesis. The p value of 0.054 means that if the population means of the two businesses are identical, there is only a 5.4% chance of observing a sample mean at least as extreme as that observed.

Reporting a p value is the logical endpoint of a Fisherian test, leaving one free to believe either in 5.4% chance of coincidence of the population means or that they do really differ. By choosing an appropriate threshold p value, one may formalise the belief. If the threshold p value is greater than the observed p value, the hypothesis is rejected (the test is significant), meaning that there is statistical evidence against the hypothesis (the data are found to be inconsistent with the hypothesised model).

All Fisherian tests of significance follow the logic illustrated. The intent is to look for the extent to which data (as summarised by a test statistic) contradict the model corresponding to the null hypothesis. The p value is used to measure the contradiction (sufficiently small p values contradict the model). It is this contradiction that lies at the heart of all Fisherian tests.

2.2 Testing according to Neyman and Pearson

Fisherian tests only have one hypothesis—the null hypothesis. Neyman-Pearson tests were designed to choose between two explicitly stated hypotheses: the hypothesis of interest H_1 (the alternative hypothesis) and the complementary hypothesis H_0 (the null hypothesis). Another main difference between Fisherian tests and Neyman-Pearson tests is that the latter always involve a prescribed significance level (denoted by α) for determining whether the null hypothesis is rejected in favour of the alternative hypothesis.

The α level is the probability of rejecting H_0 when it is true (a type I error or false positive). The opposite error is to accept H_0 when it is false (a type II error or false negative), which has probability β. The power of the test is $1 - \beta$, the probability of a true positive. The intent is to find a test for a given α that maximises the power. All calculations are done assuming H_0 being true.

For the example at hand, we choose a significance level $\alpha = 0.05$ (the choice is purely arbitrary) and then find the values of the test statistic t^* for which $\Pr(T > |t^*|) = \alpha/2$ holds. These "critical values" $(-2.20, 2.20)$ form an interval called an *acceptance region*, its complement (the rest of the real line) being *a rejection region*. A rejection region is chosen so that the probability of data falling into this region is α when H_0 is true.

The reject-accept decision depends on where the observable value of the test statistic falls. In the example, $t = -2.16$ falls in the acceptance region, so the outcome is failure to reject H_0 in favour of H_1 at the 5% level of significance. The same analysis can be done via the p value (0.054) compared to

the α level (0.05), but p values play no useful role in Neyman-Pearson tests, with the rejection being always claimed at the prescribed α significance level, which is the only thing that should be reported. By choosing a suitable α one can always reject the null hypothesis, thereby "approving" the hypothesis of interest. The Neyman-Pearson theory provides no guidelines for choosing the appropriate α, the value of $\alpha = 0.05$ being customarily used.

3 What's the controversy?

The significance test controversy is whether testing for significance is at all useful. A whole host of problems revolves around significance tests, many of them stemming from misuse and misinterpretation by users. We mention a few of the most notable problems in our view.

3.1 Mismarriage of the century

As with any aged research, tests of significance have accumulated a great deal of healthy criticism (e.g., Berger, 2003), but the mainstream of the controversy seems to lie in their practical use and interpretation. One issue which can account for a number of problems is that what most contemporary users were taught to use is an anonymous hybrid that neither Fisher nor Neyman with E. Pearson would have recognised as their original brainchildren. Fisher, in particular, was most unhappy to see the incorporation of his evidential p values into the Neyman-Pearson testing framework. Hubbard and Bayarri (2003) provide a detailed analysis of the confusion over p values and α levels from the two approaches to testing. There have been attempts in the statistical literature to unify and reconcile different approaches to hypothesis testing (e.g., Lehmann, 1993; Schervish, 1996; Berger, 2003), but the practice of hypothesis testing appears to be unchanged.

3.2 Misinterpreting p values

A p value is the probability of getting a test statistic (a data summary) at least as extreme as the observed one, assuming the null hypothesis is true. We write this using the short-hand notation $p = \Pr(\text{data}|H_0)$. A p value is not the probability of the null hypothesis being true, as it is commonly interpreted. That is, $\Pr(\text{data}|H_0) \neq \Pr(H_0|\text{data})$. Such probability statements about a hypothesis being true given the data (evidence) are beyond the realms of classical hypothesis testing. It is the task of Bayesian statistics to yield statements about $\Pr(H_0|\text{data})$.

3.3 Misapplying tests

Testing a hypothesis is not the only way to approach problems, and often it may not be the right way. (An opinion piece on applying significance tests to tackle problems they were never designed for is found in Chatfield (2006, p. 7)). Fortunately, there are alternatives to testing hypotheses. For example, statisticians agree that, when possible, estimation by confidence intervals is preferred to hypothesis testing, as they provide more interpretable information (e.g., Hall and Selinger, 1986). In the example given above, a 95% confidence interval for your mean sales is $(99.5, 120.2)$.

Many argue that it is always the size that matters, but situations may exist when the effect size does not matter. As Jones and Tukey (2000) wrote: "The scale of the outcome measure may be so untrustworthy that the primary interest then may appropriately reside in just the direction of [an] effect rather than in its size". This is an example where no harm may be caused by the thoughtful use of the significance testing.

More often, however, hypothesis tests provide no good answer to the right question. For example, you may want to check a simple hypothesis that a short sequence of small non-negative integers have come from a Poisson distribution. The null hypothesis then may be $H_0 : \text{R} \sim \text{Poi}(\lambda = 0.5)$. Suppose we failed to reject this null hypothesis, nor similar hypotheses with $\lambda = 0.4$ or $\lambda = 0.6$. Which of the

three is best? According to Descartes' dictum "when it is not in our power to follow what is true, we ought to follow what is most probable", which is the one to follow?

3.4 Mislearning from tests

The usefulness of p values, and tests which use them either explicitly or otherwise, is limited. Testing for significance often provides little help in understanding the data. The rejection of the hypothesised model could be provoked (among other things) by the fact that the data are incompatible with the test assumptions (e.g., independence, normality, equality of variance for all observations). In addition, p values are dependent on sample sizes, and they convey no information about the size of the effect being measured, which may well be of no practical importance whatsoever. To sum up, (small) p values help us learn nothing about the data, the model and the importance of the result.

3.5 Misunderstanding the need

In forecasting, hypothesis tests are often based on point forecasts and certain (hypothesised) models. In contrast, Chatfield (2005) reiterates that more forecasters should produce interval forecasts (prediction intervals), and more attention should be given to issues of model uncertainty. The former concern is naturally satisfied when forecasts are based on a statistical model, and the latter issue immediately arises upon recalling George Box's maxim "All models are wrong, but some are useful" (Box and Draper, 1987, p. 424). They are all wrong because approximations to an unknown underlying process can be good or bad, but never right. Consequently, the question of interest is not whether a certain model can be rejected (they all can with enough data), but whether among all models that lie there are those that lie a little, those that won't mislead when used for forecasting.

4 What's for forecasters?

In the predictive approach to statistics (McLean, 2000), problems of statistical analysis are viewed as prediction problems, and the central theme is a statistical (probability) model. A key idea in the approach is to use only those models that work, only those producing reasonably good predictions. This is clearly a problem of model selection rather than significance testing.

A powerful practical alternative to hypothesis testing for forecasting methods is to select models using information criteria (e.g., Hyndman et al., 2008, ch. 7). The information criterion approach has nothing to do with the testing paradigm: there is no null hypothesis, no arbitrary threshold α levels and no notion of significance. This approach to model selection can be readily extended to model ranking and model averaging so that Descartes' dictum can be used. The approach can also be extended to forecasting methods that are not based on models with computable likelihoods (e.g., Anderson et al., 2000). However, that is altogether another story.

It is not that testing the slope coefficient for zero or residuals for normality and independence is never appropriate; or that tests for equality of predictive accuracy are never useful. Rather, there is widespread confusion, inappropriate use and interpretation of significance testing. While statisticians admit the confusion (Hubbard and Bayarri, 2003), and while the discipline of statistics is making a continuing effort to provide the world with a unified, problem-free testing methodology (e.g., Berger, 2003), it is perfectly reasonable for a non-statistician to brush aside significance tests altogether, tending to use non-test alternatives wherever possible.

5 Conclusion

So do we need significance tests in business forecasting? Not necessarily. Forecasting can live without p values and α levels. Instead, we suggest attention be paid to finding models that forecast well.

References

Anderson, D. R., K. P. Burnham and W. L. Thompson (2000) Null hypothesis testing: Problems, prevalence, and an alternative, *The Journal of Wildlife Management*, **64**(4), 912–923.

Berger, J. (2003) Could Fisher, Jeffreys and Neyman have agreed on testing? (With discussion), *Statistical Science*, **18**(1), 1–32.

Box, G. E. P. and N. R. Draper (1987) *Empirical model-building and response surfaces*, Wiley.

Chatfield, C. (2005) Time-series forecasting, *Significance*, **2**(3), 131–133.

Chatfield, C. (2006) Confessions of a pragmatic forecaster, *Foresight: The International Journal of Applied Forecasting*, **6**, 3–9.

Christensen, R. (2005) Testing Fisher, Neyman, Pearson, and Bayes, *The American Statistician*, **59**(2), 121–127.

Diebold, F. X. and R. S. Mariano (1995) Comparing predictive accuracy, *Journal of Business and Economic Statistics*, **13**(3), 253–263.

Hall, P. and B. Selinger (1986) Statistical significance: Balancing evidence against doubt, *Australian & New Zealand Journal of Statistics*, **28**(3), 354–370.

Hubbard, R. and M. J. Bayarri (2003) Confusion over measures of evidence (p's) versus errors (α's) in classical statistical testing. (With discussion.), *The American Statistician*, **57**(3), 171–182.

Hyndman, R. J., A. B. Koehler, J. K. Ord and R. D. Snyder (2008) *Forecasting with exponential smoothing: the state space approach*, Springer.

Jones, L. V. and J. W. Tukey (2000) A sensible formulation of the significance test, *Psychological Methods*, **5**(4), 411–414.

Le Cam, L. (1986) The central limit theorem around 1935, *Statistical Science*, **1**(1), 78–96.

Lehmann, E. L. (1993) The Fisher, Neyman-Pearson theories of testing hypotheses: One theory or two?, *Journal of the American Statistical Association*, **88**(424), 1242–1249.

Mayo, D. G. (1992) Did Pearson reject the Neyman-Pearson philosophy of statistics?, *Synthese*, **90**, 233–262.

McLean, A. (2000) The predictive approach to teaching statistics, *Journal of Statistics Education*, **8**(3).

Schervish, M. J. (1996) P values: What they are and what they are not, *The American Statistician*, **50**(3), 203–206.