Strategic Management Journal
Strat. Mgmt. J., **33**: 108–113 (2012)
Published online EarlyView in Wiley Online Library (wileyonlinelibrary.com) DOI: 10.1002/smj.975
Received 6 October 2011; *Final revision received 10 October 2011*

RESEARCH PROSPECTIVES

THE SEARCH FOR ASTERISKS: COMPROMISED STATISTICAL TESTS AND FLAWED THEORIES†

RICHARD A. BETTIS*
Kenan-Flager Business School, The University of North Carolina, Chapel Hill, North Carolina, U.S.A.

This paper discusses repeated tests and the resultant reporting of statistical significance when it is actually not present. These errors interact with professional norms such as biases against both replication studies and 'non-results' to undermine the efficacy of our base of empirically tested theory. This raises serious issues for the future of strategic management research. Suggestions are made for dealing with these issues substantively and in terms of professional norms. Copyright © 2011 John Wiley & Sons, Ltd.

INTRODUCTION

> False positive and exaggerated results in peer reviewed scientific studies have reached epidemic proportions in recent years. The problem is rampant in economics, the social sciences and even the natural sciences, ... (Professor John Ioannidis, Chaired Professor of Medicine and Director, Stanford Prevention Research Center, (Ioannidis, 2011)

Like many other academics, I am from time to time invited to present my current research in strategic management at other universities. One of the most pleasurable aspects of such occasions is meeting with individual faculty members and Ph.D. students for informal research discussions. In the last several years I have left a few of these visits feeling intellectually troubled by some of these discussions.

One recent incident captures and summarizes the source of my discomfort. I asked an obviously talented second year Ph.D. student at a top 25 business school the usual, uncreative, conversation starter, 'so what are you studying?' His reply of 'I look for asterisks' momentarily confused me. He proceeded to tell me how as a research assistant under the direction of two senior faculty

Keywords: statistical methodology; significance levels; data snooping; theory testing; replication
*Correspondence to: Richard A. Bettis, Kenan-Flager Business School, The University of North Carolina, Strategy and Entrepreneurship Department, McColl Building CB #3490, Chapel Hill, NC 27599-3490, U.S.A. E-mail: r_bettis@unc.edu
† The views expressed in this essay are those of the author and are not intended to reflect any editorial policies of the *Strategic Management Journal*. Will Mitchell, a fellow coeditor and friend has expressed concern that some readers may stop after reading the introduction in the mistaken belief that the paper is an *SMJ* Editorial and not an *SMJ* Research Prospective. If you are in this category I urge you to read on beyond the introductory anecdote.

members he searched a couple of large databases for potentially interesting regression models within a general topical area with 'asterisks' (10% or better significance levels) on some variables. When such models were found, he helped his mentors propose theories and hypotheses on the basis of which the 'asterisks' could be explained. With the details of the conversation changed, I have had similarly sincere, but methodologically naïve comments from other faculty and Ph.D. students. In discussions of the serious scientific downsides of such an approach to empirical research and theory building, I have found that a common rationale of some scholars is one or another variation on 'this is the game we play to get published.'

The core issue here is repeated statistical tests, or more precisely the reporting of grossly inappropriate significance levels without considering all the tests that have been run. My primary intention in writing this essay is to increase recognition of the important issues we face because of repeated statistical tests. I also hope it will stimulate a productive conversation in the profession.

COMPROMISED STATISTICAL MODELS: REPEATED TESTS

In order to examine the issues raised by repeated statistical tests, it is instructive to keep three broad categories of repeated testing in mind. The first results from the fact that multiple (perhaps numerous) researchers with similar interests will be using the same databases at any time and also across time. A second is the 'tuning' of an *ex ante* model with some additional tests after the planned analysis. In this regard, I am not without sin. The final category involves indefensible *ex ante* data snooping described as 'searching for asterisks' above. These categories probably capture most of repeated statistical testing that occurs in strategic management.

Unconnected research projects

The nature of the problems with some reported significance levels in unconnected research projects can be understood by starting with a simple thought experiment.[1] This thought experiment is also useful for understanding the consequences of all situations involving repeated tests. We test n independent null hypotheses that are all actually true, but assume their truth is unknown to us as it would be in an actual statistical test. (Notice that independence of hypotheses cannot be assumed in many situations.) Assume each of these hypotheses is tested on a separate subset of data. Under these conditions, the probability of rejecting at least one true null hypothesis is $p = 1 - (1 - \alpha)^n$. When n $= 1$ and $\alpha = 0.01$ we have, of course, $p = 0.01$. Now let's set n to 100 and let $\alpha = 0.05$. We find that $p = 0.9941$ and we are almost certain to reject a true null.[2] Hence, we will very likely have *at least* one significant hypothesis test and, given the strong bias in strategic management against non-results, it will be published and, thus, considered by some researchers as an empirically 'confirmed' or 'proven' result.'[3] This raises an important but neglected point—a rejected null is not 'proof' or 'confirmation,' but only a certain kind of evidence in support of the alternative. Replication of this hypothesis test on other samples would very likely provide counterevidence, but would not be publishable because professional norms generally preclude publication of replication studies and what are usually called 'non-results.'

In fact, the norm of only publishing studies with statistical significance is, through a perverse logic, likely to result in what is considered an interesting or counterintuitive result when the null is erroneously rejected. The adjectives 'interesting' and 'counterintuitive' may simply be logical consequences of the fact that the null hypothesis is true. Certainly, it should often be counterintuitive and/or interesting when a statistically supported alternative hypothesis is 'discovered' when the null is actually true. Also, notice that the probability of error here is quite extreme (p=0.9941) and that a much lower 'n' will often provide a false rejection of the null.

[1] This 'thought experiment' was suggested by and is based on an example that appears in Hand, Manilla, and Smyth (2001: 113).

[2] Furthermore, this calculation can be turned upside down to gain a positive result by assuming a p and n, then solving for α. For example, for $p = 0.05$ and $n = 100$, we calculate that $\alpha = 0.0005$

[3] When we move from this simple demonstration in the direction of more realistic and important situations, the analysis becomes much more complicated. The hypotheses are likely to have dependence among at least some of them. Furthermore, some hypotheses may be true and some false, and the samples may be overlapping to varying degrees. Obviously, the results that parallel the above thought experiment will be complex and dependent on many factors. Simulations can be used to understand aspects of such situations, but no general results are available.

DOI: 10.1002/smj

Such a 'thought experiment' is actually a likely and even common occurrence. Today and in the past, there are many scholars in strategic management examining independent (and dependent) hypotheses in different subsamples of major databases (e.g., patents and Compustat), blissfully ignorant of each other. In other words, the real equivalent of this thought experiment is actually being run currently with real data by the author and readers of this paper.

Ex post tuning of *ex ante* models

Repeated testing in unconnected research projects is unpreventable and largely unconscious as researchers are working independently. *Ex post* tuning of *ex ante* models to make them look better, while certainly not unconscious, seems difficult for most to resist. It is quite common in varying degrees. For example, dropping some insignificant variables or trying a few additional independent variables to improve the results after the data have been analyzed in terms of the *ex ante* model and hypotheses often occurs. This is still troubling since it impacts statistical probabilities. It is difficult to justify unless the resultant papers make the nature and extent of the additional testing clear to readers and referees so that they can form a judgment about the weight of the evidence presented by the analysis.

Searching for asterisks

Data snooping[4] or searching for asterisks is the most damaging form of repeated testing, since the aim is to reject null hypotheses while consciously ignoring the many models and tests that have been conducted and, thus, reporting greatly exaggerated levels of significance. In some cases, it is likely that in excess of a thousand tests have been run in search of a 'publishable model.'

To get an idea of the magnitude of the problems that are caused by data snooping, it is useful to briefly review a simple simulation of this problem that appears in a popular textbook on statistical modeling by D.A. Freedman (2009: 74, 75, 79, and 80). Freedman's simulation includes 50 independent variables and a single dependent variable. All 51 variables are *pure noise*, N(0,1). To simulate data snooping, Freedman runs the regression with 100 rows of data and then tests all of the resultant coefficients for significance at the 10 percent level. The significant independent variables are retained ('the keepers') and a new regression model of Y is created with them alone. The experiment is then run 1,000 times with new samples of noise for the variables each time. Only 19 of these runs had no significant coefficients. As a typical (actually median) example, Freedman cites a run of the experiment that has an R^2 of 0.20 and what he describes as 'dazzling t-statistics' (-1.037, 3.637, 3.668, -3.383, and 2.536). Remember that all the variables are *pure noise*! In summary he states that:

> 'A little bit of data-snooping goes a long way: t-statistics with [t] $>$ 2 are the rule not the exception—in regressions on the keeper columns. If we add an intercept to the model, 'the' F-test will give off-scale P-values (Freedman, 2009: 79)'[5]

Ex post tuning of models and data-snooping without reporting the actual number of tests run and/or the estimated significance levels[6] based on all the tests run, seems inappropriate if we are to make claims that strategic management is establishing scientific evidence regarding strategies, firms, and senior managers. Unfortunately, consciously or unconsciously, both have occurred and resulted in what is wrongly taken by many to be 'established' or 'proven' theory.

FLAWED THEORY

Perhaps the most deleterious effect of repeated statistical testing is that flawed theory is put in place. This results not just from repeated tests

[4] This is sometimes mistakenly called 'data mining,' which is now an important academic and applied field on the interface of statistics and computer science that uses both statistical and algorithmic techniques to extract economically valuable predictive relationships from very large databases of many interdependent variables. It is interesting to note that textbooks in this field generally recommend careful avoidance of the problems of repeated statistical tests (e.g., Hand, Manilla, and Smyth, 2001).

[5] In his discussion of this simulation, Freedman notes that there are texts that recommend data snooping and gives the example of Hosmer and Lemeshow (2000). However, he states that the procedure this book suggest of a 'preliminary screen at a significance level of 25 percent will actually inflate the level of R^2 and F beyond the levels found in his simulations.

[6] In all but the simplest cases, estimated significance levels would be crude at best for the reasons explained in Footnote 3.

themselves, but from their interaction with some institutional features of the strategic management field, including (1) a bias against publication of replication studies, (2) no reporting of so-called 'non-results,' and (3) the belief by many scholars, referees, and journal editors that all journal articles should both develop and test theory. As a result of the first two, it has become a common practice to assume, at least implicitly, that one study finding a significant coefficient is adequate to establish the efficacy of a hypothesis as 'established' or 'proven' theory.[7] This is completely inconsistent with the nature of the evidence that statistical tests provide (e.g., Berk, 2004; Freedman, 2009).

In many cases, once such a significant coefficient is found, the theory being tested is metaphorically hoisted up to the appropriate place in the theoretical superstructure of the field and firmly welded into the 'gap' between other 'proven' hypotheses for which significant coefficients have previously been found. There it subsequently provides connections onto which further theoretical beams can be welded as appropriate hypotheses are developed and tested. This is the way that at least some of what we consider theory is built today, one or a few significant coefficients at a time. Given the problems with repeated tests, I worry that some, and perhaps a lot, of the theory we are building may be more like a house of cards than a strong and enduring edifice of tightly welded steel beams.[8]

It is also hard to understand why the process of theory building has become primarily tied to empirical theory testing *within the same paper*. (This has certainly not been the case in the natural sciences.) Many such papers are undoubtedly fine contributions, but it would seem reasonable to have more of a mix of theory development papers, theory testing papers, and hybrid combinations of the two. This omits exploratory research and phenomenon-based research including research that provides careful descriptions, both of which have important epistemological benefits. Furthermore, there are excellent theory building tools available such as qualitative research, computational modeling (simulation), and game theory. However, some journals even have explicit statements to the effect that all papers must do both theory development and theory testing to qualify for review.[9] A few even go so far as to propose that all papers must also have important implications for practice to qualify for review. I doubt that this impressive research trifecta is truly achieved to a meaningful degree in many papers, though it may be an admirable set of goals in searching for the Holy Grail of published research.

Tying theory building only to statistical testing in the same paper discriminates against development of more comprehensive top-down theories such as the 'behavioral theory of the firm' or the 'resource-based view,' both of which were initially developed largely independently of statistical testing. Instead we primarily assume a bottom-up process of building from small results with statistical significance to collections of interconnected results (each of which individually has statistical significance) that may or may not constitute a coherent theory as aggregation increases.[10] Why not encourage work that builds from the bottom up and work that builds from the top down?

Importantly, there are serious agency problems when the same authors are in charge of both developing theories as formal hypotheses and testing them. Publication is a powerful incentive for opportunistic behavior regarding the relationship of the two, especially when it comes to finding significant coefficients. This runs the risk of becoming 'searching for asterisks.' As one senior scholar and friend explained to me, 'for any significant coefficient there has to be a theory behind it.' Sometimes that theory has a contrived feel to me as disparate literatures and theories are combined to generate a particular functional form, though the assumptions among the constituent theories may be inconsistent or even contradictory.

DISCUSSION

There is a strong and more general critique of the abuse of statistical testing that has been rapidly

[7] It is also problematic that empirical research papers often imply causality to theory tests based on significant coefficients, though the statistical models and tests typically used can only establish association.

[8] For a very different and interesting take on theory testing in strategic management see Miller and Tsang (2011).

[9] This has resulted in Kafkaesque situations where editors and referees demand that stochatic simulation studies must include statistical tests of formal hypotheses to be considered research.

[10] For an illustration of this point applied to the theory of chief executive officer impact on firm financial performance see Blettner, Chaddad, and Bettis, 2012.

DOI: 10.1002/smj

developing in recent years across several fields. This critique is now becoming a topic for discussion in the public media (e.g., Freedman, 2010; Lehrer, 2010). It is both in our own self-interest and incumbent on us as professional scholars to strive to ensure the highest level of professionalism in our application of research methodology. Otherwise, the usefulness of our empirical models and empirically established theory will be increasingly questioned by thoughtful managers and ultimately by society in general. After all, governments, firms, private institutions, and individuals are making significant investments in our research, but have so far not required much in terms of accountability. At the same time, I recognize the nature of inertia in any academic field and the strongly felt pressures by junior faculty to 'play the publication game' to get tenure. Change, if it is to occur, will not be easy; but, I have some suggestions.

Baseline data

We simply do not have any baseline data on how big or small are the problems introduced by repeated statistical tests in strategic management. One logical way to generate such data would be to undertake the replication of highly cited statistical studies from our top journals using different samples.

An example of such analysis from top journals in medicine is instructive. In the *Journal of the American Medical Association*, Ioannidis (2005) reports on 49 highly cited (more than 1,000 citations) clinical research studies from three top medical journals. Subsequent studies were of comparable or larger size and similarly or better controlled design. Of the original 49 studies, 45 claimed to find that the particular medical intervention examined was effective. Of these seven (16%) were contradicted by the subsequent studies, and seven others had found effects that were stronger than the replications. It gives one pause to consider that this is happening in medical research that moves from publication immediately into clinical practice. It is indeed fortunate that replication occurs in medicine.

Given the relative care associated with statistical testing in medical science, 16 percent contradicted results seems a conservative lower bound on what could be expected in strategic management with replication, and the actual contradiction rate could be much higher.[11] If the rate is 16 percent, this could undermine considerably more than 16 percent of what is currently taken as 'established' theory given the interdependencies when significant coefficients are incorporated as accepted theory into the development and testing of further hypotheses.

Some better statistical practices

There are several ways in which our statistical practices relevant to repeated tests can be improved. Simply reporting the actual number of tests conducted during the study would be a big step forward. Of course in the case of 'searching for asterisks' this would obviously disqualify serious consideration if honestly reported. What I have in mind is very minor *ex post* tuning of models developed *ex ante* for studies otherwise in conformance with good statistical practice. Of course, there are some precise tests that correct for certain forms of repeated testing. Most prominent among these is the well-known Bonferroni correction (see Miller, 1991), which is included with virtually all major statistical software packages. When tests such as the Bonferroni are appropriate, there are obvious benefits to using them.

There is a set of primarily graphical techniques, many originally developed by Tukey (1977) and known collectively as exploratory data analysis, which allow one to creatively examine the data *ex ante* for potential relationships without raising the issue of repeated statistical tests. All comprehensive statistical software packages include these techniques. They are simple to use. They seem to be largely unutilized in our field. This is puzzling.

Perhaps the most general approach would be the use of the well-established technique of cross-validation or split samples. Here the original sample is split into two mutually exclusive subsamples. One is examined statistically *ex ante* for an appropriate model, which can subsequently be carefully validated (no repeated tests) using the second subsample. This approach is not perfect, but the problems are relatively minor. Split samples are widely used in the discipline of data mining (see Footnote 4), where the first sample is often known as

[11] Of course, causes of contradicted results besides repeated testing such as idiosyncratic characteristics of the specific sample originally tested, may be involved in the medical study discussed above or similar potential studies in strategic management.)

DOI: 10.1002/smj

the training sample, and the second sample as the validation sample. This technique, as applied in data mining, has apparently been the source of many predictive relationships that yielded substantial profits.

Professional norms

As discussed above, it is an established norm to consider a significant coefficient conclusive in 'proving' the theory embodied in the particular hypothesis. This is totally inconsistent with the probabilistic and evidential nature of statistical testing. We simply need to make the facts of statistical testing correspond with how test results are reported and interpreted in our profession.

Publishing replication studies based on different samples from the original would be an important change that would go a long way toward constructing a self-correcting research reporting system. Closely related to this is some form of publication and/or reporting for studies that are not replications but yield no significant coefficients ('non-results'). It is unclear to me what form this publication could take, but electronic publication opens up simple ways for researchers to make this information available in a highly condensed form indexed by topic, perhaps with citation and/or brief summary in a special section of top journals.

Perhaps the most important norm that needs to be changed is the training of Ph.D. students regarding the theory and application of statistical tests. Some Ph.D. students receive little rigorous training in probability theory and the theory of statistical tests, but are urged to take only courses that primarily emphasize the pragmatic use of one or more general statistical software package. These packages all include many esoteric models and tests for which one only needs to push the 'enter' key. Learning to appropriately use and interpret powerful statistical software, without a rigorous understanding of the theory of statistical testing, is a prescription for disaster and can turn both the science and art of rigorous statistical research into a paint-by-numbers exercise fraught with oversights and errors. Of course, such training is necessary but not sufficient, if data-snooping is encouraged by senior faculty. Therein lays the crux of the whole matter. As long as a significant number of faculty use, encourage, or ignore data-snooping, little will change. The choice is yours and mine.

ACKNOWLEDGEMENTS

This essay is dedicated to the memory of CK Prahalad, a friend and coauthor on four papers early in our careers. I learned a great deal from him, though we disagreed intellectually on some important research issues. He was always skeptical of the dependence of strategic management research on statistics. I suspect he would have liked this paper.

I thank Dani Blettner, Dirk Martignoni, Will Mitchell, Ed Zajac, Chris Bingham, Kent Miller, Torben Andersen, Jeff Edwards, Aleksandra Rebeka, Changhyun Kim, Kevin Miceli, and Rajat Khanna for helpful comments and suggestions while I was preparing this paper.

REFERENCES

Berk RA. 2004. *Regression Analysis: A Constructive Critique*. Sage:Thousand Oaks, CA.

Blettner D, Chaddad F, Bettis RA. 2012. The CEO performance effect: challenges of empirical modeling and a complex fit perspective. *Strategic Management Journal* **33**: forthcoming.

Freedman DA. 2009. *Statistical Models: Theory and Practice* (2nd edn). Cambridge University Press: Cambridge, UK.

Freedman DH. 2010. Lies, damned lies and medical science. *Atlantic Magazine*, November.

Hand D, Mannila H, Smyth P. 2001. *Principles of Data Mining*. MIT Press:Cambridge, MA.

Hosmer DW, Lemeshow S. 2000. *Applied Logistic Regression* (2nd edn). Wiley:Hoboken, NJ.

Ioannidis PAJ. 2011. An epidemic of false claims. *Scientific American* **304**(6): 16.

Ioannidis PAJ. 2005. Contradicted and initially stronger effects in highly cited clinical research. *Journal of the American Medical Association* **294**(4): 218–228.

Lehrer J. 2010. The truth wears off: is there something wrong with the scientific method? *New Yorker*, 13 November.

Miller KD, Tsang EWK. 2011. Testing management theories: critical realist philosophy and research methods. *Strategic Management Journal* **32**(2): 139–158.

Miller RG. 1991. *Simultaneous Statistical Inference*. Springer-Verlag:New York.

Tukey JW. 1977. *Exploratory Data Analysis*. Addison-Wesley:Reading, PA.