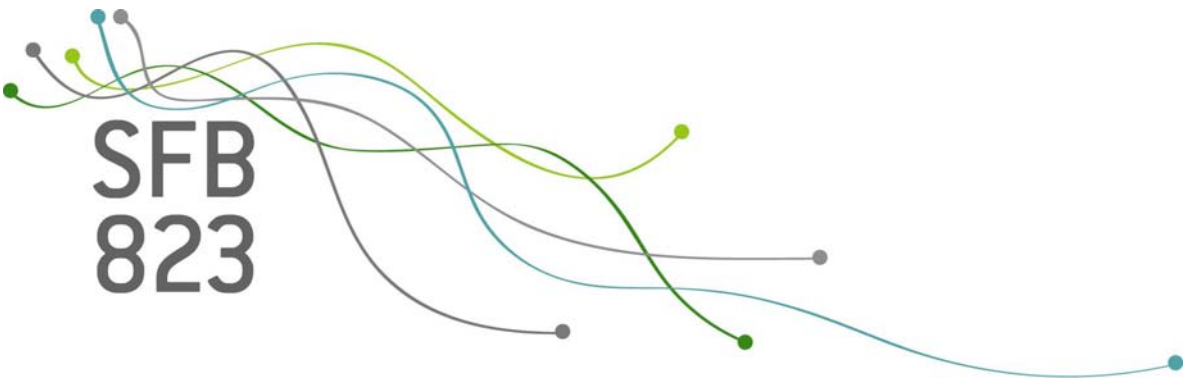
SFB 823

Discussion Paper

Das Signifikanztest-Ritual und andere Sackgassen des Fortschritts in der Statistik

Walter Krämer

Nr. 32/2011



Walter Krämer

Das Signifikanztest-Ritual und andere Sackgassen des Fortschritts in der Statistik[1]

Zusammenfassung

Dies ist eine Bestandsaufnahme des aktuellen Stellenwertes der Statistik in den Wirtschaftswissenschaften. Sie konstatiert eine gewisse Diskrepanz zwischen dem, was Ökonomen während ihres Studiums lernen, und dem, was sie an Statistik später im Beruf und in der Forschung wirklich brauchen. Insbesondere wird das sogenannte „Signifikanztestritual“ einer kritischen Prüfung unterworfen. Werden statistische Signifikanztests mit dem Hintergedanken unternommen, eine vorgefasste Meinung zu bestätigen, öffnen Sie Missbrauch und grobem Unfug Tür und Tor. Dagegen scheint bei sogenannten Spezifikationstests, die dazu dienen, etablierte Theorien anzugreifen, noch ein großes Potenzial für eine vermehrte sinnvolle Hilfestellung der Statistik zu bestehen.

1. Einleitung und Überblick

Wie misst man den Fortschritt in der Statistik? Auf jeden Fall anders als in der Physik oder in der Medizin. Die Natur- und Lebenswissenschaften, wie auch die Sozialwissenschaften inklusive Ökonomie und Soziologie, streben nach einem besseren Verständnis dieser Welt. Die Statistik dagegen stellt nur die Werkzeuge bereit, die Welt besser zu verstehen; das Verstehen selbst ist eine Angelegenheit der jeweiligen Sachwissenschaften, denen die Statistik dient.

[1] Die Arbeit entstand im Rahmen des DFG-Sonderforschungsbereiches 823: Statistik nichtlinearer dynamischer Prozesse in Wirtschaft und Technik. Ich danke einem anonymen Gutachter für wertvolle Kommentare, die zur Präzisierung meiner Thesen beigetragen haben. In dem Umfang, wie letztere auch Meinungen wiedergeben, sind das ausschließlich die des Verfassers.

Der vorliegende Artikel nimmt eine teilweise, vor allem auf wirtschaftliche Anwendungen zielende Bestandsaufnahme dieses Werkzeugvorrats vor. Erfüllen die Modelle und Methoden, die man während des Studiums der Statistik lernt, tatsächlich ihren intendierten Zweck? Oder hält ein missbräuchlicher Gebrauch ihrer Werkzeuge sogar den Erkenntnisfortschritt in den Sachwissenschaften zurück? Bzw. kann es als Folge einer Überdosierung statistischer Verfahren sogar zu schädlichen Nebenwirkungen im Sinne von geradezu falschen Sachaussagen kommen?

Die Antwort ist einmal nein und zweimal ja. Kein unbedingtes Nein und kein unbedingtes Ja, aber, wie im weiteren zu zeigen ist, doch ein begründeter Zweifel an dem Nutzen so mancher Methoden und Modelle, mit denen Statistiker heute ihren Kollegen und Kolleginnen in einigen Fachdisziplinen helfen zu können glauben. Hier ist zunächst eine traurige Diskrepanz zu konstatieren zwischen dem, was der typische Anwender der Statistik während seines Studiums übt und lernt, und dem, was er oder sie später in den Anwendungen oder auch als Wirtschaftsforscher wirklich braucht. Zumindest in den Wirtschaftswissenschaften nämlich zielt die Statistikausbildung mehr oder weniger weit an den Bedürfnissen der Praxis vorbei. Und auch die Methodenforschung erscheint eher ihren eigenen Gesetzen als den Nöten der Anwender zu gehorchen. Der folgende Abschnitt 2 fast diese schon mehrfach andernorts ausgeführte These (Krämer 2004, 2008) nochmals kurz zusammen.

Abschnitt 3 wendet sich dann im Detail einem der wichtigsten Werkzeuge der mathematischen Statistik, den statistischem Signifikanztests, zu. Hier mehren sich in den letzten Jahren Stimmen, die einen zunehmend unsachgemäßen Gebrauch dieses Werkzeugs konstatieren oder gar dessen Nützlichkeit ganz allgemein verneinen. Insbesondere in den Wirtschaftswissenschaften trifft der routinemäßige Gebrauch statistischer Signifikanztest auf zunehmenden Widerstand: "The progress of economic science has been seriously damaged [by the common practice of significance testing]." So schreibt Deirdre McCloskey (2002, S. 44), seit Jahrzehnten eine der wortgewaltigsten Kritikerinnen des, wie sie es nennt, Signifikanztestrituals. "You can’t believe anything that comes out of [it]. Not a word. It is all nonsense, which future generations of economists are going to have to do all over again. Most of what appears in the best journals of economics is unscientific rubbish. I find this unspeakably sad. All my friends, my dear, dear friends in economics, have been

wasting their time....They are vigorous, difficult, demanding activities, like hard chess problems. But they are worthless as science.“ Oder nochmals deutlicher, in ihrem letzten Buch mit Stephen Ziliak (2008, S. 40): ”If null-hypothesis significance testing is as idiotic as we and other critics have so long believed, how on earth has it survived?”

Eine vernichtendere Kritik eines der wichtigsten Werkzeuge der mathematischen Statistik ist wohl schwerlich vorzustellen. Und wie in Abschnitt 3 an zahlreichen Beispielen dargelegt, trifft sie in gewisser Weise auch tatsächlich zu. Auf der anderen Seite wird aber hier sozusagen das Kind mit dem Bade ausgeschüttet, und wird ein in den Anwendungen mehr als nützlicher Typus von Signifikanztests, sogenannte Spezifikationstests, zu Unrecht angeklagt. Der abschließende Abschnitt vier führt aus, dass hier noch ein großes Potenzial für weiteren Fortschritt in der ökonometrischen Methodenforschung liegt.

2. Das Mißverhältnis zwischen Angebot und Nachfrage in der wirtschaftsstatistisch/ökonometrischen Ausbildung und Methodenforschung

Bei einer Rückschau auf die wirtschaftswissenschaftlich relevante statistische Methodenforschung seit dem Zweiten Weltkrieg fällt als erstes auf, dass es immer wieder Moden gibt; solange eine dieser Moden andauert, fühlen sich Anwender fast schon zwangsweise verpflichtet, die jeweiligen Modelle und Methoden zu verwenden, ob sie nun zu den Daten passen oder nicht. Das fing mit den Simultanen Gleichungen an. Angestoßen durch mehrere einflussreiche Monographien der gleichermaßen einflussreichen Cowles-Kommission war die ökonometrische Methodenforschung fast zwei Jahrzehnte auf die Effekte Simultaner Gleichungen und den Umgang damit zentriert - wann ist ein solches System identifiziert, wie schätzt man es, wie leitet man daraus Prognosen ab? Und gnadenlos wurde den Studierenden der Unterschied zwischen der indirekten Kleinst-Quadrate-Methode, der "Limited Information ML-Schätzung" (LIML), der “Full Information ML-Schätzung (FIML) sowie der zweistufigen und dreistufigen KQ-Methode eingebläut. Heute weiß das keiner mehr, und kaum jemand verwendet noch viel geistige Energie auf die Schätzung von dergleichen Modellen; die Endogenitätsproblematik der Regressoren

in Simultanen Gleichungen hat sich als ein im Vergleich zu anderen eher minder wichtiges Problem und die jahrzehntelange Beschäftigung damit als eine große Verschwendung geistiger Ressourcen herausgestellt.

Dann wurde die Methodendiskussion einige Jahre von den rationalen Erwartungen dominiert, bis dann mit einem einflussreichen Aufsatz von Engle und Granger die Integrations- und Kointegrationsproblematik in das Zentrum des Interesses rückte. Lange Jahre war in ökonomischen Fachjournalen keine empirische, mit Zeitreihendaten arbeitende Untersuchung zu lesen, in der man sich nicht verpflichtet gefühlt hätte, auf Einheitswurzeln zu testen, und das oft auch dann, wenn von der Sache her unmöglich Einheitswurzeln vorliegen konnten (etwa bei Variablen, deren Werte in einem beschränkten Intervall liegen müssen). Dann wieder waren einige Jahre Modelle mit stochastisch variierenden Koeffizienten en vogue ("Markov-Switching"), und aktuell haben die Mikro-Ökonometrie und die Paneldaten-Problematik das Zepter in der Hand.

In der eher betriebswirtschaftlich orientierten Literatur hießen die Schlagwörter “Box-Jenkins-Verfahren”, “Kalman-Filter”, “Optionsbewertung” oder “Neuronale Netze”. Der Verfasser erinnert sich an die Reminiszenz eines bedeutenden amerikanischen Fachkollegen anlässlich der ISI-Tagung in Dublin 2011, wo letzterer aus einem Seminar von George Box aus den späten sechziger Jahren berichtete: kein Datensatz, gleich welcher Provenienz, hätte den Seminarraum ohne eine ARMA-Modellierung verlassen, ob sie nun passte oder nicht. Aber auch hier kühlte sich ein initialer Enthusiasmus bald wieder ab bzw. wandte sich anderen Moden zu. Daneben und parallel dazu hat auch die Verfügbarkeit immer größerer Datenmengen und Rechnerkapazitäten jeweils eigene Konjunkturen ausgelöst, von der Verlaufsdatenanalyse über neuere Verfahren der nichtparametrischen Statistik (Kernschätzer) und die Bootstrap-Revolution bis zu den jüngsten Entwicklungen in der Hochfrequenzmethodologie bei Finanzmarktdaten. Hier reagiert die statistische Methodenforschung auf eine zum Teil fast schon dramatische Ausweitung des Rechenkapazität- und Datenangebots (siehe etwa Rendtel 2011), aber auch hier darf man die Frage stellen: hat das den Erkenntnisfortschritt in der Sache substantiell vorangebracht? Es ist ja schön und gut, wenn man mit Bootstrap-Verfahren robuste

Konfidenzintervalle für seine Parameterschätzungen erhält, aber was soll das nutzen, wenn das Modell als solches hinten und vorne sozusagen klemmt?

Allen diesen Moden ist gemeinsam, dass in gewissen Kontexten durchaus drängende Probleme auch da gesehen werden, wo ganz andere Aspekte wichtig wären - die Methodenforschung entwickelt eine Eigendynamik und drängt sich den Anwendern sozusagen auf.[2] Die haben es aber in erster Linie mit Datenschutz und Datenbankgestaltung, mit fehlenden Daten und verzerrten Stichproben, mit Fragebögen und Meßfehlern oder ganz generell der mangelnden Passgenauigkeit der verfügbaren Informationen zu tun: wie definiert und misst man Armut, Ungleichheit und Arbeitslosigkeit? Und wie die durchschnittliche Änderung der Preise? Oder den Wohlstand einer Volkswirtschaft? Dass hier das übliche Sozialprodukt als Indikator allein nicht ausreicht, bewegt inzwischen ja sogar die hohe Politik; anläßlich der 100-Jahr-Feier der Deutschen Statistischen Gesellschaft in Leipzig 2011 war dies eines der beiden zentralen Themen der ganzen Veranstaltung. Ob dagegen der eine oder der andere Schätzer effizienter oder der eine oder der andere Test tatsächlich mächtiger ist, berührt den empirisch tätigen Statistiker nur am Rande.

Aber genau darauf ist die statistische Methodenausbildung vielfach fokussiert. Stunden um Stunden knapper Vorlesungszeiten werden den Themen Konsistenz und Effizienz, asymptotische Normalverteilung und Gütevergleichen gewidmet, die elementaren Probleme der Gewinnung und Interpretation der Daten werden dabei sozusagen als gelöst unterstellt. Eine effiziente Ausbildung wie auch Forschung in Statistik hätte aber ihre Ressourcen da zu konzentrieren, wo ein zusätzlicher Mitteleinsatz den größten Nutzen im Sinne von Erkenntnisfortschritt bringt. Und das scheint mir nicht da zu sein, wo aktuell die Forschung schwerpunktmäßig stattfindet und wo aktuell der Nachwuchs in Statistik unterwiesen wird. Oder um mit Nobelpreisträger Trygwe Haavelmo (1958) zu sprechen, der schon vor mehr als 50 Jahren die folgende, leider viel zu wenig beachtete Einsicht zu äußern wagte: "The concrete results of our efforts at quantitative measurement often seem to get worse the more refinement of tools ... we call into play." Oder nochmals deutlicher der in letzter Zeit in vielfacher Funktion immer wieder aufgefallene Lawrence H. Summers (1991, S. 129). Er argumentiert, dass "formal econometric work, where elaborate

[2] Auch als Maslow-Prinzip bekannt: "It is tempting, if the only tool you have is a hammer, to treat everything as if it were a nail" (1966, S. 15).

technique is used to apply theory to data or isolate directions of causal relationships when they are not obvious a priori, virtually always fails. The only empirical research that has contributed to thinking about substantive issues and the development of economics is pragmatic emprical work, based on methodological priciples directly oposed to those that have become fashionable in recent years."

Wenn also Egeler u.a. (2011) als Vertreter der deutschen Amtsstatistik schreiben: "Selbst das reine Erhebungsgeschäft kommt heute nicht ohne wissenschaftliche Methoden aus, von der Stichprobentheorie bis zur auch stark psychologisch geprägten Interview- und Fragebogengestaltung", dann ist das zwar richtig, aber genau diese Methoden werden an den Universitäten kaum gelehrt.

3. Die fehlende Signifikanz der Signifikanz

Das Testen von Hypothesen ist eine Standardprozedur der mathematischen Statistik. Jeder Student der Wirtschaftswissenschaften lernt sie spätestens im zweiten Semester. Und wenn er oder sie später empirisch arbeitet, sind die Ergebnisse wahrscheinlicher als andersherum mit einer Reihe von Signifikanztests garniert - ohne eine Batterie von t-Tests scheinen wissenschaftliche Arbeiten in Ökonomenjournalen heute nicht mehr publizierbar.

Dabei sagt ein zum Niveau 5 % signifikanter t-Test doch nur: wenn die jeweilige Nullhypothese richtig wäre – und das ist ein dickes wenn –, dann wäre die Wahrscheinlichkeit für das beobachtete oder für noch extremere Ergebnisse höchstens 5%. Das ist nicht gerade viel. Und hier fängt die Malaise an. Denn viele Anwender interpretieren einen signifikanten Test als Beweis, dass die Alternative richtig ist (siehe Krämer 2011 für Beispiele). Oder sie deduzieren daraus zumindest noch eine Wahrscheinlichkeit von 95 %, dass die Alternative richtig ist. Selbst in diversen Lehrbüchern kommt dieser letzte Fehler vor, siehe Krämer und Gigerenzer (2005) für eine erschreckend lange Liste von Lehrbüchern der mathematischen Statistik, die genau dieses behaupten.

Krämer (2011) nennt das den Fehler 3. Art. Dieser Fehler der dritten Art, obwohl weit verbreitet, soll aber im weiteren nicht in erster Linie interessieren. Denn in leider nur allzu vielen Anwendungen trifft schon die behauptete Ausgangswahrscheinlichkeit für einen Fehler erster Art nicht zu; diese ist sehr oft sehr viel größer.

Eine erste Ursache dafür ist das, was Kerr (1998) als "HARKing" bezeichnet ("Hypothezising after the results are known"). Auf dieses Hypothesenbilden nach der Kenntnis der Daten und die daraus folgende extreme Unterschätzung eines Fehlers 1. Art hat bereits Kruskal (1969, S. 247) hingewiesen: “Almost any set of data […] will show anomalies of some kind when examined carefully, even if the underlying probabilistic structure is wholly random – that is, even if the observations stem from random variables that are independent and identically distributed. By looking carefully enough at random data, one can generally find some anomaly […] that gives statistical significance at customary levels although no real effect is present.”

Ein berühmtes Beispiel ist die wohl erste bekannte Anwendung eines Signifikanztests überhaupt, die schon lange unter Astronomen bekannte Tatsache, dass die Umlaufbahnen der Planeten alle in engen Winkeln zueinander stehen. Im Jahr 1734 haben Daniel Bernoulli und sein Sohn Johann die Wahrscheinlichkeit berechnet, dass dies auf Zufall zurückzuführen ist (unter der Voraussetzung, dass die Umlaufbahnen zufällig zustande gekommen sind). In moderner Terminologie haben sie also das empirische Signifikanzniveau (den “prob-value”) eines Tests der Nullhypothese berechnet, dass eine solche zufällige Auswahl vorliegt. Und natürlich wird diese Hypothese bei allen gängigen Signifikanzniveaus abgelehnt.

Der Punkt ist aber, dass diese Hypothese erst *nach* Ansicht der Daten zustande gekommen ist. Und da man in jedem Datensatz Anomalien findet, führt auch jeder Datensatz – bei geeigneter Auswahl der Nullhypothese – zu dem statistisch gesicherten Schluss, dass eine "signifikante" Abweichung von der "Normalität" existiert. Wie z.B. in Krämer und Arminger (2010) nachzulesen, sind fast alle Panikmeldungen der letzten Jahre zum Thema "Leukämie durch Kernkraftwerke" auf diese Art und Weise zustande gekommen. Denn natürlich ist diese Krankheit, wie alle anderen auch, nicht gleichmäßig in der Weltpopulation verteilt, es gibt Cluster allerorten, wie bei anderen Zufallsmustern auch. Und wenn man in den Mittelpunkt

eines solchen Clusters einen Punkt legt und die Hypothese aufstellt: "In dieser Umgebung gibt es signifikant mehr Krebsfälle als anderswo", muss natürlich die Hypothese eines abwesenden Effektes hochsignifikant verworfen werden.

Einen weiteren Wachstumsschub erhalten die Wahrscheinlichkeiten für einen Fehler erster Art durch die bekannte Publikations-Verzerrung (“publication bias”). Quer durch alle Wissenschaften ist eine Tendenz nicht abzustreiten, vor allem solche Ergebnisse in den einschlägigen Journalen zu publizieren, die vom Erwarteten abweichen. Eine Kapitalmarktanalyse etwa des Inhalts, dass an Freitagen, die auf einen 13. fallen, die Aktienmärkte weltweit keinen Deut anders reagieren als an anderen Wochentagen, wird man in keiner halbwegs angesehenen Fachzeitschrift gedruckt wiederfinden. Sollten dagegen die Märkte an solchen Tagen überdurchschnittlich häufig (im Sinne von: signifikant häufiger als 50 %) fallen, ist eine Veröffentlichung fast schon garantiert.

Aber natürlich ist das Signifikanzniveau dieses Tests weit höher als das nominal unterstellte, wie hoch auch immer dieses sei. Auf die perversen Konsequenzen dieser Publikationsmechanismen hat schon Sterling (1959, S. 30) hingewiesen: “There is some evidence that in fields where statistical tests of significance are commonly used, research which yields nonsignificant results is not published. Such research, being unknown to other investigators may be repeated independently until eventually by chance a significant result occurs.” Konsequent zu Ende gedacht bedeutet dies: ein “signifikanter” Effekt wird auf jeden Fall gefunden, ganz gleich, was wirklich Sache ist. Beck-Bornholt und Dubben (2004) liefern eine schöne Sammlung von dergleichen Effekten in der Medizin.

Dieses Aufblähen von Wahrscheinlichkeiten für einen Fehler erster Art findet auch dann statt, wenn einzelne Forscher jeder für sich nach den Regeln der Kunst verfahren. Aber selbst das findet ebenfalls sehr oft nicht statt. Vielmehr ist es fast schon als die Regel anzusehen, dass unter den vielen in einem Aufsatz berichteten Resultaten nur die besonders "signifikanten” übrig bleiben. Und davon gibt es selbst bei Abwesenheit jedweder Effekte genug, wenn man nur oft genug testet. Denn ein Signifikanzniveau von etwa 5 % sagt ja doch nur: sollte die Nullhypothese zutreffen, d.h. sollte es keiner Effekte geben, werden dennoch in 5 % aller Fälle Werte der

jeweiligen Prüfgröße jenseits des kritischen Wertes vorkommen. Und dreimal darf man raten, welche Tests dann die Seiten der Fachjournale schmücken!

In der Fachliteratur ist dieses Phänomen seit langem als Data Mining bekannt (Lovell 1983): Testet man hundertmal zum Niveau 5%, wird man im Mittel in fünf Fällen, selbst bei Abwesenheit jeglicher Effekte, signifikante Resultate erhalten, und genau diese sind es dann, die ins Auge fallen und nach einer Publikation zu verlangen scheinen.

Diese bisher diskutierten Anwendungsfehler sind sozusagen Betriebsunfälle, bei korrekter Anwendung von Signifikanztests, bzw. bei Publikationsstrategien, die positiven wie negativen Resultaten gleiche Chancen gäben, kämen sie nicht vor. Aber die im Abschnitt 1 zitierte Fundamentalkritik wäre dadurch nicht entschärft. Denn diese zielt tiefer, sie argumentiert, dass die Herangehensweise als solche erkenntnishindernd sei. Selbst wenn man nach den Regeln der Kunst verfährt und ein Test das Signifikanzniveau einhält, so Ziliak und McCloskey (2008), die wohl rabiatesten Gegner jedweder Signifikanztesterei, ist in vielen Anwendungen mit dem Ergebnis dennoch nur wenig anzufangen, oder noch schlimmer: Die Ergebnisse führen systematisch in die Irre. Man verwechsle als erstes , so Ziliak und McCloskey, die statistische mit der ökonomischen Signifikanz. Geringfügige Abweichungen von der Nullhypothese würden durch große Stichproben leicht signifikant, auch wenn sachlich völlig unerheblich. Dagegen, und das ist der zweite und oft noch folgenreichere Fehler, blieben große Abweichungen von H_0, d.h. tatsächlich spürbare Effekte, wegen hoher Stichprobenvariabilität oft unsignifikant und würden so zu Unrecht als unerheblich abgetan.

Ziliak und McCloskey (2008) nennen das auch den "sizeless stare": Man starrt nur auf die zwei oder drei Sterne über dem Regressionskoeffizienten, der Koeffizient als solcher wird im Wesentlichen ignoriert. Mit, wie Ziliak und McCloskey argumentieren, katastrophalen Folgen für den Fortschritt in den Wirtschaftswissenschaften: "Sizeless economic research has produced mistaken findings about purchasing power parity, unemployment programs, monetary policy, rational addiction, and the minimum wage. In truth, it has vitiated most econometric findings since the1920s and virtually all of them since the significance error was

institutionalized in the 1940s. The conclusions of Fisherian studies might occasionally be correct. but only by accicent." Oder anders ausgedrückt: Die Jagd nach Signifikanzen hat die Jagd nach Realitäten abgelöst.

Wie in Krämer (2011) an vielen Beispielen dargelegt, trifft dieser Vorwurf auf viele empirische Arbeiten zumindest in den Wirtschaftswissenschaften auch tatsächlich zu. Bei Durchsicht aller jemals im German Economic Review erschienenen empirischen Aufsätze stellt Krämer fest, dass in über 500 Testanwendungen nur berichtet wird, eine Variable oder ein Effekt sei signifikant, ohne die Größe des Einflusses überhaupt nur zu erwähnen. Aber genau darauf kommt es doch in den Anwendungen hauptsächlich an. Solange aber empirische Wirtschaftsforscher vor allem auf die Sterne hinter ihren Regressionskoeffizienten und nicht auf deren sachliche Bedeutung achten, lenken statistische Signifikanztests von den wahren Fakten ab, und sind damit keine Hilfe, sondern eine Bremse für die Wissenschaft.

4. Theorieerhärtende versus theorieattackierende Tests

Die in Abschnitt 3 ausgeführten Fehlentwicklungen betreffen durchweg sog. konfirmatorische oder theorieerhärtende Tests. Damit ist gemeint, dass eine Ablehnung der Nullhypothese einer vorgefassten Theorie entgegenkommt. Man vermutet einen Effekt, etwa des Inhalts "Kernkraftwerke erzeugen Leukämie" und hofft, der Signifikanztest lehnt die Nullhypothese der Abwesenheit eines solchen Effektes ab. Und wie in Abschnitt 3 gezeigt, ist das Erzeugen einer solchen Ablehnung eine der leichtesten statistischen Übungen überhaupt. Diese mit der leichten Verfügbarkeit einschlägiger Sofwarepakete überhandnehmende konfirmatorische Testerei ist damit ein Hemmschuh für den Fortschritt in allen Wissenschaften, in denen dieser Unfug Platz gegriffen hat.

Die Anhänger dieser Generalkritik übersehen dabei aber eine viel wichtigere Aufgabe von statistischen Signifikanztests, nämlich etablierte Theorien einer Prüfung zu unterwerfen. Ganz im Sinne des Popperschen Wissenschaftsparadigmas hält man so lange an herkömmlichen Modellen fest, bis die Daten eine Abkehr davon erzwingen, und kommt so Schritt für Schritt der Wahrheit näher. Und hier können statistische

Signifikanztests eine große Hilfe sein. So fordert etwa die etablierte Kapitalmarkttheorie, dass sukzessive Renditen risikobehafteter Wertpapiere keine Korrelationen aufweisen. Sind diese dennoch vorhanden, ist der Kapitalmarkt nicht mehr effizient, im Widerspruch zur etablierten Theorie. Diese wäre damit widerlegt.

Das einzige Problem in diesem Kontext ist, zwischen statistisch und ökonomisch signifikanten Abweichungen zu unterscheiden. Denn natürlich werden bei hinreichend großen Stichproben immer Signifikanzen vorkommen, aber die so aufgedeckten minimalen Abweichungen lassen sich nicht durch Handelsstrategien für Gewinne ausnutzen, sind also ökonomisch irrelevant. Von dergleichen Feinheiten abgesehen, sind aber solche nichtkonfirmatorischen Tests, das heißt Signifikanztests, bei denen eine Ablehnung einer vorgefassten Meinung widerspricht, in vielen Fächern unentbehrlich für das Aussortieren falscher Theorien.[3]

Der Begriff der nichtkonfirmatorischen Tests im hier verwendeten Sinn überlappt sich stark mit dem, was in der Ökonometrie als Spezifikationstests bekannt ist. Der Terminus geht wohl auf Haussmann (1978) zurück und meint in aller Regel Tests einer Hypothese, dass das jeweils verwendete statistische Modell als solches überhaupt die Daten gut beschreibt (also in diesem Sinne "korrekt spezifiziert" ist). Eine spezifische Alternativhypothese ist in aller Regel nicht vorhanden. Krämer und Sonnberger (1986) stellen im Kontext des Linearen Regressionsmodells eine große Auswahl solcher Spezifikationstests vor. Denn vor allem Lineare Modelle werden leider allzu oft ohne weitere Motivation und Begründung an alle möglichen Daten angepasst, und längst nicht immer treffen die für eine sinnvolle Interpretation der geschätzten Koeffizienten nötigen Voraussetzungen tatsächlich zu. Sind wirklich alle relevanten Regressoren aufgeführt? Ist die Beziehung tatsächlich linear? Sind Wechselwirkungen auszuschließen? Sind die Regressionskoeffizienten über die ganze Stichprobe identisch? Die Konsequenzen eines Nein sind hier geradezu dramatisch und lassen jede auf ein solches Modell aufbauende Analyse zu statistischen Schrott verkommen. Und genau darum, um nichts anderes statistischen Schrott, handelt es

[3] Diese Unterscheidung ist in der Literatur nicht ganz klar. Im vorliegenden Kontext hängt es vom Anwender ab, ob ein Test konfirmatorisch ist oder nicht. Ist ihm oder ihr eine Ablehnung angenehm, da einer liebgewordenen These entsprechend, ist der Test konfirmatorisch. Ist ihm oder ihr eine Ablehnung unangenehm, ist der Test nichtkonfirmatorisch.

sich nach Ansicht des Verfassers bei rund 90 %, wenn nicht mehr, aller empirischen Papiere in den Wirtschaftswissenschaften der letzten drei Jahrzehnte.

Verglichen mit einer Fehlspezifikation der ersten Momente ist es weit weniger wichtig, ob die Störgrößen eines solchen Regressionsmodells wirklich alle die gleiche Varianz oder eine Normalverteilung besitzen. Und auch eine mögliche stochastische Abhängigkeit der Störgrößen ruiniert noch nicht die auf einem solchen Modell aufbauenden Analysen. In diesem Sinne kann man es also nur als eine weitere Sackgasse der Forschung bezeichnen, dass genau diesen Fragen in der Vergangenheit erheblich mehr Aufmerksamkeit gewidmet worden ist als einer korrekten Spezifikation der ersten Momente des Modells. Tests auf Unkorreliertheit oder Homoskedastie der Störgrößen ökonometrischer Modelle füllen ganze Schrankwände in statistischen Bibliotheken, die wirklich wichtigen Spezifikationstests werden dagegen in Forschung und Lehre recht stiefmütterlich behandelt.

Und es ist genau an dieser Stelle, dass ein Umschwenken in der ökonometrischen Methodenforschung noch großen Nutzen zu versprechen scheint. Indem die Statistiker den Ökonomen helfen, bessere Modelle für die Realität zu finden, wird auch das Schätzen der Koeffizienten dieser Modelle wieder zu einem sinnvollen Unterfangen, was es leider in den letzten Jahrzehnten nur in Ausnahmefällen gewesen ist.

5. Fazit

Die hier geäußerte Kritik will nicht sagen, in den empirischen Wirtschaftswissenschaften sei in den letzten Jahr nichts vernünftiges geleistet worden. Ganz im Gegenteil. Große Fortschritte wurden etwa bei der Berücksichtigung von Qualitätsänderungen bei Preisindices oder bei der Erfassung der Schattenwirtschaft gemacht (siehe etwa Feld und Schneider 2010). Und nicht zu vergessen die vor einigen Jahrzehnten noch kaum zu erträumenden Möglichkeiten des Datenzugangs, auf die Rendtel (2011) so verdienstvoller Weise hinweist. Und in den ökonomischen Fachzeitschriften findet man vermehrt auch Versuche, Signifikanztests im Popperschen Sinne zur Falsifizierung von Theorien einzusetzen (siehe etwa

Zarzoso u. a. 2010). Aber genauso wie ein mit beladener Supertanker nicht mit einer einzigen Handbewegung sofort in eine andere Richtung umzulenken ist, sondern durch seine Massenträgheit noch meilenweit in die alte Richtung weitertreibt, steht auch in der empirischen Wirtschaftsforschung ein mentales Trägheitsmoment einer dringend nötigen Kurskorrektur sehr störend im Wege.

Literatur

Beck-Bornholdt, H.-P. and Dubben, H.-H. (2004): *Unausgewogene Berichterstattung in der medizinischen Wissenschaft - publication bias-*, Hamburg (Institut für Allgemeinmedizin des Universitätsklinikums Hamburg-Eppendorf).

Egeler, R., Wöll, Th. und Zwick, M. (2011): "Perspektiven für die amtliche Statistik." *Wirtschafts- und Sozialstatistisches Archiv,* diese Ausgabe.

Feld, L.P. and Schneider, F. (2010): „Survey on the shadow economy and undeclared earnings in OECD countries“, *German Economic Review* 11, 109-149.

Haavelmo, Trygve, (1958): The role of the econometrician in the advancement of economic theory. *Econometrica* 26, 351 – 357.

Hausmann, J. A. (1978): "Specification Tests in Econometrics". *Econometrica* 46**,** 1251–1271.

Kerr, N. L. (1998). “HARKing (Hypothesizing After the Results are Known)”. *Personality and Social Psychology Review*, 2, 196-217.

Krämer, W. (2004): "Statistik: Vom Geburtshelfer zum Bremser der Erkenntnis in den Sozialwissenschaften?" *Kölner Zeitschrift für Soziologie und Sozialpsychologie* 44, 51-60.

Krämer, W. (2008): "Verhindert die Statistikausbildung den Fortschritt der Wirtschafts- und Sozialwissenschaften?", *Wirtschafts- und Sozialstatistisches Archiv,* 41-50.

Krämer, W. (2011): "The cult of statistical significance." Erscheint in *Schmollers Jahrbuch.*

Krämer, W. and Sonnberger, H. (1986): *The Linear Regression Model under Test,* Heidelberg (Physica-Verlag).

Krämer, W. and Gigerenzer, G. (2005): „How to confuse with statistics. The use and misuse of conditional probabilities,” *Statistical Science* 20, 223-230.

Krämer, W. and Arminger, G. (2010): “True believers, or numerical terrorism at the nuclear power plant,” erscheint in *Jahrbücher für Nationalökonomie und Statistik.*

Kruskal, W. (1968): “Tests of statistical significance.” In: David Sills et al. (ed.): *International Encyclopdedia of the Social Sciences*, New York (McMillan), 238-250.

Lovell, M.C. (1983): “Data Mining”, *Review of Economics and Statistics* 65, 1 – 12.

Abraham H. Maslow (1966). *The Psychology of Science*, New York (Harper & Row).

McCloskey, D. (2002): *The Secret Sins of Economics*, New York (Wiley).

Rendtel, U. (2011): "Die Zukunft der Statistik: Eine persönliche Betrachtung." *Wirtschafts- und Sozialstatistisches Archiv,* diese Ausgabe.

Sterling, T.R. (1959): „Publication decisions and their possible effects on inferences drawn from tests of significance – or vice versa, “, *Journal of the American Statistical Association,* 30-34.

Zarzoso; I., Nowak-Lehmann, F., Klasen, S. and Larch, M. (2009): “Does German development aid promote German exports?” *German Economic Review* 10, 317-338.

Ziliak, S. and McCloskey, D. (2008): *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice and Lives*, University of Michigan Press.