

Walter Krämer

Verhindert die Statistikausbildung den Fortschritt der Wirtschafts- und Sozialwissenschaften?

1. Einleitung und Überblick

Sehr geehrte Damen und Herren,

ich muss mich zunächst einmal für den etwas reißerischen Titel meines kleinen Referates entschuldigen: Verhindert die Statistikausbildung den Fortschritt der Wissenschaften? Natürlich nicht. Die Frage so zu stellen hieße, die Bedeutung unseres schönen Faches doch etwas zu übertreiben. Schließlich ist die Statistik ja „nur“ eine **Hilfswissenschaft**, sie kann den Fortschritt in den Sachwissenschaften befördern oder bremsen, auf keinen Fall verhindern.

Wenn wir aber fragen: Steht die aktuelle Statistikausbildung, etwa an deutschen Wirtschaftsfakultäten, dem Erkenntnisfortschritt in der Ökonomie eher im Wege, so kann man durchaus sagen: Ja. Das werde ich im weiteren auch noch ausführlich versuchen zu begründen.

Zuvor zur Einsortierung meiner Überlegungen: Anders als die Kollegen Rentel und von der Lippe, die sich um die Bedürfnisse der nichtwissenschaftlichen Abnehmer von Statistikstudierenden gekümmert haben, befasse ich mich mit der Statistik als Hilfswissenschaft in anderen **Wissenschaften**. Auch die eher kurzfristig-organisatorischen, also sozusagen **taktischen** Probleme, die bisher diskutiert wurden – etwa die Umstellung auf Bachelor- und Masterstudiengänge, die Anrechnung von Leistungspunkten oder die verschiedenen Wahlmöglichkeiten im Studium – lasse ich mal außen vor. Stattdessen richte ich meinen Blick eher auf langfristige, strategische Zusammenhänge zwischen dem, was die Studierenden lernen und dem, was sie später als Forscher und Professoren in diversen Anwendungswissenschaften brauchen. Das ist völlig unabhängig davon, wie viele Stunden Statistik man im Studium belegt hat oder ob man einen Diplom-, Bachelor- oder Master-Studiengang durchläuft.

Nur eine kleine Nebenbemerkung kann ich mir hier nicht verkneifen: nämlich wie rückgradlos und ohne jeden Widerstand viele deutsche Hochschullehrer sich dieses in vielen Kontexten kontraproduktive Bachelor- und Master-Studium haben überstülpen lassen. Das fängt mit den amerikanischen Mickymausgraden Bachelor und Master an. In Italien sagt man auch nach Bologna weiter Laurea 1 und 2 dazu, in Frankreich heißt das Ding „maîtrise“. Warum benutzen wir in Deutschland nicht auch unsere eigenen traditionellen Grade Baccalareus und Magister?

Aber auch jenseits dieser Äußerlichkeiten ist das zweigliedrige Studiensystem in vielen Fächern absoluter Blödsinn. Wir haben deshalb am Fachbereich Statistik in Dortmund beschlossen, das bewährte und weltweit anerkannte Diplom so lange anzubieten, wie es gesetzlich möglich ist. Wie Sie wissen, gibt es ja auch eine Klage der Bochumer Kollegen am Bundesverfassungsgericht gegen das in NRW drohende Diplom-Verbot. Es ist mir ein absolutes Rätsel, wie man etwa das deutsche Diplom bei den Ingenieuren, das weltweit den allerbesten Ruf genießt, das ein internationales Markenzeichen ist, so ohne jede Gegenleistung wegwerfen kann; dafür würde jeder Marketing-Chef eines Großkonzerns auf der Stelle entlassen.

Aber das ist nicht mein Thema. Ich wende mich lieber den folgenden Fragen zu:

1. Gibt es überhaupt ein Fach Statistik?
2. Sind die dort gelehrteten Inhalte für die späteren Berufstätigkeiten relevant?
3. Sind die dort gelehrteten Inhalte für die Anwendungswissenschaften relevant?

Um es gleich vorwegzunehmen: Meine Antwort heißt in allen Fällen "Nein". Kein absolutes Nein, um hier keine Panik auszulösen, aber ein qualifiziertes Nein. Dabei ist die erste Frage gleich die wichtigste, denn

hier liegt die Wurzel des ganzen Übels. Nämlich: Wenn wir hier heute von Statistik reden, wovon reden wir da überhaupt? Was meinen wir mit "Statistik". Und meine These ist: wir reden hier über völlig verschiedene Dinge, eine einheitliche Wissenschaft namens Statistik gibt es nicht. Lassen Sie mich deswegen mit einigen grundsätzlichen Gedanken zum Wesen des akademischen Fachs Statistik anfangen.

2. Gibt es überhaupt ein Fach Statistik?

Wie Sie wissen, kommt Statistik aus dem Italienischen "Statista" = "Staatsmann". Und dementsprechend war bis vor 100 Jahren mit Statistik das gemeint, was heute im allgemeinen Wirtschaftsstatistik heißt. Im Brockhaus-Konversationslektion von 1895 liest sich das so: "Statistik ist ursprünglich so viel wie Staatskunde, worunter die systematische Darstellung der Verfassung, der Organisation, der Bevölkerungsverhältnisse, der militärischen und wirtschaftlichen Hilfsquellen und der sonstigen bemerkenswerten Einmischung eines oder mehrerer Staaten zu verstehen ist."

Völlig unabhängig und parallel dazu entwickelte sich in der Mathematik die Wahrscheinlichkeitsrechnung, die wurde dann von Ronald Fisher, Jerzy Neyman und Egon Pearson auf ausgewählte Inferenz Probleme ausgewählter Anwendungswissenschaften angewandt, etwa um die Wirkung unterschiedlicher Düngemittel in der Landwirtschaft zu identifizieren oder die Existenz von Wirkungen überhaupt erst nachzuweisen. Ich betone: **ausgewählte** Probleme aus ausgewählten Anwendungswissenschaften. Hier ging es nicht in erster Linie um Datenprobleme, sondern darum, systematische und zufällige Effekte auseinanderzuhalten, sozusagen darum –wenn Sie mir diese Alliteration einmal verzeihen – dem Zufall Zügel anzulegen.

Und es ist eine Erbsünde in unserer Wissenschaft, an der wir heute noch leiden, dass, wer auch immer es gewesen ist, diese beiden Wissenschaften unter einem einzigen Dach namens Statistik zwangsvereinigt hat. Wenn ich das eine einmal „Stochastik“ oder „Inferenzstatistik“ und das andere „Datenanalyse“ nenne, so haben die beiden so viel miteinander zu tun wie Geographie und Geologie. Kein Mensch käme auf die Idee, daraus eine einzige Wissenschaft zu machen, und genauso wenig gehören Stochastik und Datenanalyse in denselben Topf. Erkenntnisziele und Erkenntnismethoden sind hier wie dort extrem verschieden; beides ist wichtig, wenn auch nicht in allen Anwendungen gleichermaßen, aber beides gleich zu bezeichnen schadet beiden Wissenschaften gleichermaßen. Ich meine daher, man sollte diese beiden Wissenschaften auch organisatorisch trennen.

Was ich hier Datenanalyse nenne, befasst sich mit Dingen wie Messfehler und Datenbereinigung, mit der adäquaten Definition von Begriffen wie Einkommen, Armut und Arbeitslosigkeit, mit Datensicherheit und Anonymisierung und Datenbankdesign ganz allgemein, mit fehlenden Daten und Antwortverweigerung, mit der Qualitätsberücksichtigung bei Preisindizes, mit der optimalen Gestaltung von Fragebögen, mit dem praktischen Ziehen von Stichproben, mit der optimalen grafischen Präsentation von Daten usw. Diese Problemkreise sind zugleich auch die in der Berufspraxis und in den Anwendungswissenschaften in aller Regel wichtigeren. Sehen Sie sich doch nur die aktuelle Debatte um die Unterschichtenproblematik an: wer oder was ist eigentlich arm? Es gibt dazu eine schöne Untersuchung im Allgemeinen Statistischen Archiv (Semrau und Stubig 1999; siehe auch Krämer 2000); demnach schwankt die deutsche Armutsquote von 0,5% bis 20 %, je nachdem, wo man die Armutsgrenze zieht. Und alle diese Grenzen sind gleichermaßen sinnvoll und haben ihre Verfechter in der aktuellen deutschen Sozialpolitik. Oder nehmen Sie die Indexproblematik. Sie kennen die Thesen von Boskin u. a. (1996), wonach die aktuellen Preisindizes den realen Preisanstieg in den entwickelten westlichen Industrienationen jährlich um bis zu einem Prozentpunkt überschätzen (Stichwort "Qualitätsbereinigung"). Oder nehmen Sie die Stichprobenproblematik. Meine drei letzten Auftragsprojekte aus Wirtschaft und Verwaltung kamen von der GEMA, von der Deutschen Post AG und vom Bundesversicherungsamt; alle hatten die Problematik von Stichproben zum Gegenstand. Ich glaube, hier liegt eines der wichtigsten Anwendungsgebiete der herkömmlich „Statistik“ genannten Wissenschaft überhaupt. Oder nehmen Sie die Fragebögen und Umfrageproblematik ganz allgemein.

Was hier durch alle möglichen Fehler an Unfug produziert werden kann, stellt alle schlechten Konsequenzen durch Fehler bei der Inferenzstatistik weit in den Schatten (Krämer 2006, Kap. 10: „Wie es in den Wald hineinschallt...“). Ob z. B. ein Schätzverfahren erwartungstreu, effizient oder konsistent ist, hat auf die Qualität einer damit gewonnen empirischen Aussage weit weniger Einfluss als die Art und Weise, wie die Daten überhaupt erst gewonnen worden sind. Oder um mit Wirtschafts-Nobelpreisträger Haavelmo (1958) zu sprechen "The concrete results of our efforts at quantitative measurement often seem to get worse the more refinement of tools ... we call into play."

Mit diesem "refinement of tools" befasst sich vor allem die Stochastik. Wichtige Stichwörter sind hier Maximum-Likelihood und Bayes-Verfahren, Verallgemeinerte Momententenschätzer, multiple Testprobleme, lineare und nichtlineare Regression, Versuchsplanung, Approximationstheoreme, Markoffketten, Simultane Gleichungen, Kurvenschätzungen, Erwartungstreu, Effizienz und Konsistenz, Kernschätzer, parametrische vs. nichtparametrische Verfahren allgemein, Resampling-Verfahren (bootstrap usw.), Grenzwertsätze, stationäre vs. instationäre stochastische Prozesse, Kointegration, stochastische Differentialgleichungen usw. Das alles sind faszinierende Untersuchungsgegenstände, aber mit Datenanalyse im engeren Sinn haben sie allenfalls indirekt zu tun. Es gibt Überlappungen, keine Frage, aber die sind nicht so groß, als dass sie diese Zwangsehe von Datenanalyse und Stochastik rechtfertigen könnten. Nehmen Sie die Messfehlerproblematik – alle kennen die schönen Bücher des Kollegen Schneeweiß – der die stochastischen Konsequenzen von Messfehlern nachzeichnet, aber eben nur von stochastischen Messfehlern, und nach meiner Erfahrung sind die allermeisten Messfehler nicht stochastisch. Genauso gibt es auch eine ausgefeilte stochastische Theorie des Umgangs mit fehlenden Daten, aber wiederum vor allem für Fälle, in denen der Ausfallmechanismus bestimmten sehr restriktiven Kriterien folgt, die in der Praxis nicht notwendig gelten müssen. Dann liefert natürlich auch die Stochastik große Hilfen bei der Anonymisierung oder in der Stichprobentheorie, aber wenn Sie einmal die klassischen Lehrbücher, etwa Cochran (1953), lesen, stellen Sie fest, dass die Wahrscheinlichkeitstheorie, die man dafür braucht, in zwei Wochen zu lernen ist – alles ist diskret und endlich, man muss weder wissen, was eine Sigma-Algebra noch was Messbarkeit bedeutet, um Stichprobentheorie zu können. Und wenn dann die Stochastik wirklich zentrale Hilfen bei Stichprobenfragen liefert, wie etwa in den nobelpreisgekrönten Arbeiten von Jim Heckman zu den Verzerrungen durch Selbstselektion, so sind die dafür nötigen Randbedingungen in vielen Anwendungen nicht gegeben.

Und umgekehrt hat auch die Datenanalyse kaum Wirkungen auf die Stochastik; diese geht in aller Regel von unabhängig identisch verteilten Stichprobenvariablen, korrekt spezifizierten Modellen und fehlerfrei gemessenen sowie vollständig vorhandenen Modellvariablen aus. Mit anderen Worten, all die Probleme, mit denen sich die Datenanalytiker tagtäglich herumzuschlagen haben, werden von den Stochastikern einfach als gelöst unterstellt.

Sie merken schon, ich übertreibe hier. Sicherlich gibt es auch zahlreiche Teilgebiete der mathematischen Statistik – etwa Clustermethoden und die explorative Datenanalyse ganz allgemein, auch sogenannte „naive“ Zeitreihenverfahren und die komplette axiomatische Theorie der Indexzahlen - die ohne Stochastik auskommen und eher datengetrieben operieren. Und das riesige Gebiet der robusten Statistik kann man sogar als eine Antwort der Stochastik auf die hier angesprochenen Probleme sehen: Wie kann man Methoden finden, die auch dann funktionieren, wenn die Standardannahmen versagen? Oder anders ausgedrückt: Zwischen diesen beiden hier skizzierten Polen der völlig stochastikfremden Datenerfassungs- und Definitionsproblematik auf der einen und der völlig datenfreien abstrakten mathematischen Wahrscheinlichkeitstheorie auf der anderen Seite tummelt sich ein Kontinuum von Modellen und Methoden, die einen mehr dem einen, die anderen mehr dem anderen Extrem verbunden. Aber dieses Kontinuum ist nicht gleichmäßig; es hat vielmehr zwei Modalwerte, die hinreichend voneinander entfernt sind, um ein Existenzrecht für zwei verschiedene Wissenschaften zu begründen.

Meine These ist nun: Die aktuelle Zwangsheirat dieser beiden verschiedenen Wissenschaften hat verschiedene unerwünschte Konsequenzen. Die erste ist: Wenn jeder, der eine der beiden Disziplinen

betreibt, sich Statistiker nennen kann und an einem Fachbereich nur ein Statistiker-Posten existiert, dann bestimmt der Zufall, welche dieser Disziplinen an der jeweiligen Fakultät gelehrt wird. Ist die Person ein Stochastiker, wird die Datenanalyse nur stiefmütterlich behandelt, ist sie ein Datenanalytiker, kommt die Stochastik nur ansatzweise vor. Hätten die beiden Disziplinen dagegen separate Namen, wie etwa Geologie und Geographie, dann gäbe es auch an jeder halbwegs anständigen Wirtschaftsfakultät einen Lehrstuhl für Datenanalyse und einen für Inferenzstatistik. Und das ganze Problem, über das die Statistischen Ämter und andere Abnehmer von statistisch ausgebildeten Universitätsabsolventen aus Wirtschaftsfakultäten heute klagen, wäre verschwunden.

3. Vom Nutzen der Inferenzstatistik

Wir sind uns also einig: Die datenanalytische Seite des Modell- und Methodenkontinuums namens "Statistik" ist für die spätere Berufspraxis wie für die meisten Anwendungswissenschaften gleichermaßen wichtig, bleibt aber in vielen Ausbildungsgängen etwas unterbelichtet. Wenden wir uns jetzt der anderen Seite dieses Kontinuums, der Stochastik alias Inferenzstatistik zu. Das sind alles schöne Forschungsgebiete; ich selbst habe mich Jahrzehnte meines Lebens mit Begeisterung darin getummelt. Dennoch plädiere ich hier für mehr Bescheidenheit, denn die Wissenschaften bzw. Forschungsfelder, in denen stochastische Modelle und inferenzstatistische Methoden an zentraler Stelle in die Ergebnisse einfließen, lassen sich an den Fingern einer Hand abzählen. Ich denke etwa an Optionsbewertungen in der Finanzwirtschaft, wo in der Tat eine fundierte Kenntnis von stochastischen Differentialgleichungen unentbehrlich ist. Ohne diese stochastischen Modelle und Methoden wäre eine rationale, theoretisch fundierte Optionsbewertung auch nicht möglich; hier ist also die Stochastik für das Untersuchungsziel ganz zentral.

Aber sonst fällt es mir recht schwer, irgendein Sachproblem zu finden, dessen Beantwortung entscheidend vom Ausgang eines statistischen Tests abhängt. Am nächsten kommt hier noch die Hypothese effizienter Märkte, ebenfalls wieder in der Finanzwirtschaft, wo in der Tat die mathematische Statistik bei der Antwort gefordert ist. So folgt etwa ein bereinigter und logarithmierter Aktienkurs in einem informationseffizienten Markt exakt einem Random Walk (mit Drift), und wenn er das nicht tut, ist der Markt nicht effizient. Ferner darf es nicht möglich sein, mehrere Aktienkurse linear so zu kombinieren, dass dabei etwas Stationäres herauskommt – Aktienkurse sind in effizienten Märkten niemals kointegriert. Und auch die Gültigkeit der Kaufkraftparitätentheorie oder der Erwartungstheorie der Zinsstruktur hängt entscheidend vom Ausgang geeigneter Kointegrationstests ab. Aber nennen Sie mir sonst irgendeine Substanzfrage in der Soziologie, Psychologie oder Ökonomie, wo man gespannt auf das Ergebnis eines statistischen Schätz- oder Testverfahrens wartet, um irgendeine Sachdebatte zu entscheiden. Mir fällt da nicht viel ein. Ich bemerke hier ganz im Gegenteil eine Tendenz, die Ergebnisse solcher statistischer Analysen, sofern sie denn in Sachdebatten eingreifen, komplett zu ignorieren.

Dann gibt es noch einen weiteren Grund für die zunehmende Skepsis gegenüber inferenzstatistischen Methoden in den Anwendungswissenschaften: Auch wenn man den gesamten Nutzen, den kumulativen Beitrag der mathematischen Statistik durchaus schätzt, so kann man doch den zusätzlichen Nutzen, den Grenznutzen, eher skeptisch betrachten. Nehmen Sie den Sprung von der gewöhnlichen Kleinst-Quadrat- auf die zweistufige und dann dreistufige Kleinst-Quadrat-Methode. Es ist wohl unbestritten, dass C. F. Gauss mit seiner KQ-Methode einen genialen und für viele Wissenschaften äußerst nützlichen Einfall hatte (wobei weniger die recht simple Methode an sich als vielmehr deren Motivation und Herleitung der eigentliche Geniestreich waren). Dagegen hat die zweistufige KQ-Methode von Henry Theil im Wesentlichen nur noch die Ökonometriker bewegt, und die weitere Verfeinerung zur dreistufigen KQ-Methode durch Zellner und Theil hat vielleicht sogar mehr Nach- als Vorteile gebracht. Ich selbst habe mich in meiner Dissertation (siehe Krämer 1980) sehr lange mit dem Nutzen komplizierterer Schätzverfahren in fehlspezifizierten Modellen befasst. Das Ergebnis war, dass bei nicht ganz korrekt spezifizierten Modellen kompliziertere Verfahren die Ergebnisse eher verschlechtern.

Viele Substanzwissenschaftler scheinen aber zu glauben, ohne eine Verbeugung vor der aktuellen Methodenmode keinen anspruchsvollen Aufsatz bei anspruchsvollen Fachzeitschriften einreichen zu dürfen. Diese Mode war zur Zeit meiner eigenen Promotion die Simultane Gleichungsproblematik – davon redet heute keiner mehr - heute ist es Integration und Kointegration. Nach dem bekannten Maslow-Prinzip (Maslow 1966: „... it is tempting, if the only tool you have is a hammer, to treat everything as if it were a nail.“) werden die gerade aktuellen Modelle und Verfahren ohne Rücksicht auf Verluste selbst Datensätzen übergestülpt, bei denen sie absolut nichts verloren haben, und die dabei aufgewandte Zeit und Energie geht der eigentlichen Substanzforschung verloren.

Ein letzter Grund zur Skepsis gegenüber der Bedeutung der Aussagekraft moderner inferenzstatistischer Verfahren – hier speziell Signifikanztests - kommt noch hinzu: Was wissen wir, wenn ein solcher Test eine Nullhypothese ablehnt? Über die vielfach missverstandene Bedeutung des Wortes "signifikant" nachher noch etwas mehr. Hier kommt es mir auf etwas anderes an: Eine abgelehnte Nullhypothese kann nämlich – neben dem immer drohenden Fehler 1. Art - auch andere Ursachen haben als dass die Alternative zutrifft. Die meisten statistischen Tests überprüfen immer eine „Sammelhypothese“, einmal die offizielle Nullhypothese und dann noch eine ganze Latte weiterer Unterstellungen, die das jeweilige Modell betreffen. Nur wenn diese weiteren Hypothesen alle zutreffen und allein die offizielle Nullhypothese falsch ist, kann ein signifikantes Testergebnis die übliche Interpretation erlauben. Mein Lieblingsbeispiel sind signifikante Regressionskoeffizienten im Sinne eines signifikanten t-Tests. Diese signifikanten Koeffizienten können sehr einfach auch durch korrelierte Störgrößen zustande kommen, welche die geschätzte Störgrößenvarianz verkleinern und damit den t-Test, in dessen Nenner die geschätzte Störgrößenvarianz auftaucht, so aufblähen, dass auch völlig insignifikante Regressoren plötzlich signifikant erscheinen.

Und dann natürlich das Phänomen des „Data-Mining“. Wenn man 100 völlig korrekte Nullhypothesen bei einem 5% Signifikanzniveau testet, wird man im Mittel 5 Ablehnungen erhalten. Dreimal dürfen Sie raten, welche fünf dieser 100 Studien dann in der Fachpresse erscheinen. Oder technisch ausgedrückt: Das nominelle und das tatsächliche Signifikanzniveau bei statistischen Signifikanztests stimmen oft nicht überein. Beträgt dieses tatsächlich etwa nur 1%, wie in strengeren Tests oft unterstellt, so darf man es keinem Journalisten verübeln, wenn er etwa eine „signifikante“ Häufung von Leukämiefällen in der Nähe eines Kernkraftwerks zum Anlass eines kernkraftkritischen Artikels nimmt. In 15 km Umkreis um das Kernkraftwerk Krümmel z. B. erkrankten in den 90er Jahren acht Kinder an Leukämie, ein Vielfaches des bundesweiten Durchschnitts für vergleichbare Regionen. Diese Häufung schlug Riesenwellen in den deutschen Medien und wurde als Beweis für die Gefährlichkeit des Kraftwerks angesehen. Denn falls das Kraftwerk nicht als Auslöser in Frage käme, so die Logik der Kernkraftgegner, wäre die Wahrscheinlichkeit für diese – dann rein zufällige - Häufung weit weniger als 1%. Ergo sei diese Häufung statistisch signifikant.

In Wahrheit ist sie alles andere als signifikant. Mit der gleichen Logik könnten wir auch „beweisen“, dass Fluglärm AIDS erzeugt oder dass die Anwohner von Keksfabriken häufiger als andere unter Darmkrebs leiden. Denn was hier interessiert, ist doch die Wahrscheinlichkeit, dass in der Nähe **irgendeines** Kernkraftwerks rein durch Zufall eine Häufung auftritt. In Deutschland gibt es 17, weltweit 442 aktive Kernkraftwerke. Schlägt man um jedes einen Kreis von 15 km, wird man auch dann, wenn Kernkraftwerke damit nichts zu tun haben sollten, in einigen der Kreise weniger, in anderen mehr Leukämiefälle als im Durchschnitt finden, so wie es der Zufall eben will. Die Wahrscheinlichkeit, dass in einem davon eine bei isolierter Sicht zu einem Niveau von 1% signifikante Abweichung auftritt, beträgt, wie man leicht ausrechnet, 98,8%. Mit anderen Worten, man wird fast sicher in der Nähe irgendeines Kraftwerks eine „signifikante“ Häufung von Blutkrebsfällen finden (ebenso wie signifikante Häufungen von mathematischen Wunderkindern, Lottomillionären oder Fehlgeburten).

Mit der gleichen Methode ist es mir selbst zusammen mit meinem Mitarbeiter Ralf Runde gelungen, an der Deutschen Börse einen "geteilt-durch-fünf-Rest-eins-Effekt" nachzuweisen (Krämer und Runde 1992):

An Börsentagen, die geteilt durch fünf den Rest eins ergeben, also am 1., 6., 11., 16., 21., 26. und 31. Tag des Monats, hat der deutsche Aktienindex DAX eine signifikant höhere Rendite als sonst.

Wie kam dieses Ergebnis zustande? Getestet wurden auch noch der geteilt durch fünf -Rest-zwei, Rest drei, Rest vier-Effekt, der geteilt durch sechs Rest-eins, Rest zwei, Rest drei, Rest vier, Rest fünf-Effekt, der geteilt durch sieben -Rest-eins, Rest zwei, Rest drei, Rest vier, Rest fünf-Effekt usw. – kein Wunder, dass in einem dieser Fälle allein der Zufall eine vermeintlich „signifikante“ Abweichung erzeugt.

Dieses systematische Fischen nach Testergebnissen, denen in Wahrheit keinerlei systematische Effekte zugrunde liegen, ist in vielen Wissenschaften weit verbreitet und hat dazu geführt, dass wahre Kenner angesichts „signifikanter“ Testergebnisse kaum noch mit den Schultern zucken. Denn auch seriöse Journale veröffentlichen lieber Aufsätze, die Abweichungen von hergebrachten Theorien zeigen, nichtsignifikante Resultate bleiben häufiger als signifikante ungedruckt. Als Konsequenz sind daher unter den tatsächlich publizierten wissenschaftlichen Arbeiten aller Fächer weit mehr Fehler erster Art vertreten als dem offiziellen Signifikanzniveau entspräche, und sind vermeintliche Effekte, in welcher Angelegenheit auch immer, in Wahrheit nur das Trugbild eines Fehlers 1. Art.

4. Das Signifikanztest-Ritual

Zum Abschluss ein ganz trauriges Kapitel. Selbst jenseits der Data-Mining-Problematik und selbst bei völlig korrekt spezifizierten Modellen werden nämlich die Ergebnisse statistischer Signifikanztests oft falsch interpretiert und damit als Zeugen für Aussagen benutzt, die auf falsche Fährten führen und so den Fortschritt in den Sachwissenschaften bremsen. Denn einmal heißt "signifikant" ja nur, dass, **falls die Nullhypothese zuträfe**, ein solches Ergebnis mit weniger als 5% Wahrscheinlichkeit zustande käme. Das gilt aber für viele andere Nullhypothesen auch. Daraus ist keinesfalls zu schließen, die Nullhypothese wäre wahr.

Trotzdem geschieht das aber oft.

Genauso irreführend ist auch die Interpretation eines nichtsignifikanten Tests. Die automatische Reaktion bei den meisten Anwendern ist in diesem Fall: H_0 ist falsch. Die Wahrheit ist: H_0 kann nicht verworfen werden, und das kann viele Gründe haben. Der wichtigste ist: Die Stichprobe ist zu klein. Da reichen oft eklatante Abweichungen von H_0 nicht für ein signifikantes Testergebnis aus. Bei großen Stichproben dagegen führen schon kleinste Abweichungen von H_0 zu signifikanten Prüfgrößen.

Und dann ist vielen Anwendern überhaupt nicht bewusst, was genau "signifikant" bedeutet. Es gibt hier eine schöne Studie von Haller und Kraus (2002); sie haben mehreren Dutzend deutschen Psychologiestudenten, die gerade einen Kurs in induktiver Statistik hinter sich hatten, sowie 30 ihrer Lehrer und 40 weiteren Wissenschaftlern den folgenden Fragebogen zur Bedeutung eines t-Tests vorgelegt, mit den Antwortalternativen ja oder nein. Es waren zwei Stichproben vom Umfang 20 auf identische Erwartungswerte zu testen, die t-Statistik war zu einem Niveau von 1% signifikant. Welche der folgenden Aussagen trifft dann zu?

- 1) Die Hypothese, es gäbe keine Unterschiede, ist falsch.
- 2) Die Wahrscheinlichkeit, dass die Nullhypothese zutrifft, liegt unter 1%.
- 3) Die Vermutung, es könnte Unterschiede geben, ist richtig.
- 4) Man kann immerhin die Wahrscheinlichkeit dafür angeben, dass diese Vermutung richtig ist.
- 5) Die Wahrscheinlichkeit, bei Ablehnung der Nullhypothese einen Fehler zu machen, ist kleiner als 1%.
- 6) Wenn das Experiment sehr oft wiederholt würde, käme in 99% der Fälle eine signifikante Prüfgröße zustande.

Alle diese Behauptungen sind falsch. Aber nur 20% der Statistik-Lehrer, 10% der übrigen Wissenschaftler und keiner der Studenten sah das so. Die meisten kreuzten eine oder mehrere der obigen Falschaussagen als richtig an. Selbst die Dozenten hielten im Mittel zwei der oben aufgeführten Falschaussagen für wahr. "Teaching statistics to psychology students in German Universities does not include effective enlightenment about the meaning of significance: four out of every five methodology instructions show misconception about this concept with their students" (Haller und Kraus 2002, S.8).

Selbst in Lehrbüchern findet man Aussagen wie die folgende: "eine Ablehnung zum Niveau 5% bedeutet: Mit Wahrscheinlichkeit 95% trifft die Nullhypothese zu." Ich will hier verschweigen, aus welchem Werk diese Stelle stammt. Es ist auf jeden Fall eines der meistverkauften Statistik-Lehrbücher auf dem gesamten deutschen Markt.

Nach Haller und Kraus richtet das "Signifikanztest-Ritual" also netto mehr Schaden als Nutzen an (siehe dazu auch Sedlmeir 1996, Gigerenzer u. a. 2004 oder Krämer und Gigerenzer 2005). Es trägt kaum zur weiteren Erkenntnis, mehr zur Vernebelung des Erkenntnisfortschritts bei. Statt den Blick für die in den Daten enthaltene Botschaft zu schärfen, lenkt es sogar eher davon ab. Das routinemäßige Durchführen von Signifikanztests ohne Rücksicht auf die Bedingungen, unter denen solche Tests überhaupt sinnvoll sind, unter gleichzeitiger Fehlinterpretation der erzielten Resultate, lenkt auf falsche Fährten, macht blind für wichtige Signale, engt die Sicht auf die soziale Umwelt ein und ist netto eine Bremse des Erkennens in so mancher Wissenschaft.

5. Fazit

Durch die Betonung klassischer Schätz- und Testverfahren, die zudem auch oft noch falsch verstanden werden, und durch die Kanalisierung von Lehr- und Forschungsaufwand in deren Effizienzverbesserung hat sich die konventionelle, etwa an deutschen Wirtschaftsfakultäten gelehrt Statistik von den Bedürfnissen der Anwender zusehends entfernt. Parallel dazu findet eine Vernachlässigung vieler für die Anwendungswissenschaften wichtiger Verfahren der Datenerhebung und – aufbereitung statt, so dass sich die deutsche Universitätsstatistik über den Abbau ihrer Stellen nicht zu wundern braucht.

Literatur

Boskin, Michael J., Ellen R. Dulberger, Robert J. Gordon, Zvi Grilliches und Dale W. Jorgenson, 1996: Toward a more accurate measure of the cost of living. Abschlußbericht für das Senate Finance Committee, 4. Dezember.

Cochran, William G., 1953, Sampling Techniques, New York: Wiley.

Gigerenzer, Gerd, Stefan Krauss und Oliver Vitouch, 2004: "The null ritual: what you always wanted to know about significance testing but were afraid to ask." In D. Kaplan (Hrsg.): Handbook of Quantitative Methods in the Social Sciences, S. 391-408. London: Sage.

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

Haller, Heiko und Stefan Krauss, 2002: "Misinterpretation of significance: A problem students share with their teachers?" *Methods of Psychological Research – Online* 7, S. 1 – 20.

Krämer, Walter, 1980: Eine Rehabilitation der Gewöhnlichen Kleinst-Quadrate-Methode als Schätzverfahren in der Ökonometrie. Frankfurt: Haag und Herchen.

Krämer, Walter, 2000: Armut in der Bundesrepublik – Zur Theorie und Praxis eines überforderten Begriffs. Frankfurt: Campus.

Krämer, Walter, 2006: So lügt man mit Statistik (8. Taschenbuchauflage). München: Piper.

Krämer, Walter und Runde, Ralf, 1992: „The holiday effect: Yet another capital market anomaly?“ In: S. Schach und G. Trenkler (Hrsg): Data Analysis and Statistical Inference – Festschrift in Honour of Friedhelm Eicker, Bergisch Gladbach: Eul-Verlag, S. 453-462.

Krämer, Walter und Gerd Gigerenzer, 2005: “How to Confuse with Statistics.” Statistical Science 20, S. 223-230.

Maslow, Abraham H., 1966: The psychology of science. New York: Harper & Row.

Sedlmeier, Peter, 1996: “Jenseits des Signifikanztest-Rituals: Ergänzungen und Alternativen.“ Methods of Psychological Research – Online 3, S. 39 – 42.

Semrau, Peter und Stubig, Hans, 1999: „Armut im Lichte unterschiedlicher Messkonzepte“, Allgemeines Statistisches Archiv 83, S. 324-337.