Inferenzstatistische Sprachspiele in den Humanwissenschaften: Eine kleine Fallstudie

von *Raphael Diepgen,* Bochum

Zusammenfassung: Es werden an einem konkreten Fall interessante Irrungen und Wirrungen inferenzstatistischer Argumentation in den Humanwissenschaften nachgezeichnet.

1. Zur Eröffnung

Nachdem mein kritischer Beitrag (Diepgen, 1992) über die Diskrepanz zwischen statistischem Überbau und statistischer Alltagspraxis in vielen Humanwissenschaften auf heftigste, in ihrer Erregung kaum mehr kontrollierte Empörung bei zumindest einem Interessensvertreter der mathematisch-statistischen Profession gestoßen ist (Schmitz, 1993), erscheint es mir angezeigt, meine Kritik zumindest exemplarisch etwas mehr zu belegen.

Grundsätzlich hat die Stochastikdidaktik kritisch zu prüfen, ob und inwieweit sich die statistischen Theorien in der Praxis widerspiegeln und sich erst dadurch als tatsächlich anwendungs- und dann vielleicht sogar bildungsrelevant erweisen. Dabei dürfen sich die als Beamte dem Gemeinwohl verpflichteten "Studienräte" nicht von noch privilegierteren beamteten, augenscheinlich aber dennoch besonders dem partikularen Interesse ihrer Profession verpflichteten Statistikprofessoren hindern lassen.

Forschungsmethodisch ist dies nicht ganz einfach, denn es geht ja - ähnlich wie etwa bei der kriminologischen Dunkelfeldforschung - um die heikle Frage, inwieweit bestimmte Normen und Regeln auch tatsächlich im praktischen Verhalten eingehalten werden, wobei sich dieses Verhalten nicht im Scheinwerferlicht der öffentlichen Bühne, sondern nur im Dämmerlicht der professionellen Hinterbühne zeigt. Den numerischen p-Werten in den publizierten Forschungsberichten etwa läßt sich schließlich überhaupt nicht ansehen, wie sie entstanden sind, insbesondere nicht, ob es tatsächlich um Wahrscheinlichkeiten beispielsweise im Sinne von Fisher oder Neyman und Pearson handelt.

Nichtsdestoweniger gibt es inzwischen - von mathematischen Statistikern selbstverständlich kaum wahrgenommen - recht umfangreiche und auch empirisch untermauerte Untersuchungen zur problematischen Rolle der Inferenzstatistik in den Humanwissenschaften, auf die hier wiederum der Kürze we-

Stochastik in der Schule **14** (1994), Nr. 1, S. 10-22.

gen nur verwiesen werden kann (vgl. etwa Gigerenzer und Murray, 1987). Diese Forschung zeigt beispielsweise auf, wie die Etablierung inferenzstatistischer Prozeduren in den Humanwissenschaften durch eine gezielte Vermischung eigentlich unvereinbarer Konzepte von Fisher erstens, Neyman und Pearson zweitens und Bayes drittens in den entsprechenden Statistiklehrbüchern vorbereitet wurde.

Verweisen kann diese Forschung weiterhin beispielsweise darauf, daß die inferenzstatistische Alltagspraxis in den Humanwissenschaften heutzutage reguliert wird durch - kommerziell vertriebene und daher die Bedürfnisse der Anwender sicherlich treffende - Statistikprogrammpakete wie SPSS, BMDP oder SAS, in denen die überaus komplizierte und sensible Handlungslogik des Hypothesentestens schrumpft zur schlichten Automatik der Berechnung von p-Werten für alles mögliche und damit die "Tests" gänzlich von vorab formulierten "Hypothesen" entkoppelt werden.

Daß trotz des Neyman-Pearsonschen Überbaus in all diesen Statistikprogrammpaketen vom Fehler zweiter Art nirgends die Rede ist und insbesondere keine Möglichkeiten für die Berechnung oder Kontrolle seiner Wahrscheinlichkeit geboten werden, paßt ebenso in dieses Bild wie die seltsame Tatsache, daß ein gewisser Jacob Cohen (1969) Jahrzehnte nach Neyman und Pearson mit einem Buch berühmt werden konnte, das lediglich dem augenscheinlich verblüfften Publikum der Verhaltenswissenschafter die Augen dafür öffnete, daß es da beim Hypothesentesten auch noch einen Fehler zweiter Art gebe, der sogar in seiner Wahrscheinlichkeit berechenbar sei. Leider mußte Cohen (1990) zwanzig Jahre später zugestehen, daß es ihm nicht gelungen sei, die Augen des Publikums dafür auf Dauer offen zu halten.

Kurzum: Ohne einen kritischen und allemal vage Indizien angemessen würdigenden Blick hinter die Kulissen tatsächlich praktizierter Statistik bleibt die statistikdidaktische Diskussion naiv. Daß manche professionelle Mathematiker an dieser Naivität ein Interesse haben mögen, ist tragbar - und beispielsweise aus der didaktischen Diskussion über die "mathematikfreiere" Explorative Datenanalyse vertraut, die sich ja gerade aus der Unzufriedenheit der Praktiker mit den hochmathematisierten, aber wenig problemangemessenen Verfahren traditioneller Statistik speiste (vgl. Biehler, 1982). Insofern schlägt der jüngste Praxisbericht von Buth (1993) in dieser Zeitschrift grundsätzlich den richtigen Weg ein, auch wenn sich Buth - trotz seiner grundsätzlichen Skepsis ob des "nur stark eingeschränkten und für den Laien enttäuschenden Programms" der mathematischen Statistik (wieso eigentlich nur für den Laien?) -

noch als recht braver und höflicher Praktikant erweist, insofern er die erlebte Praxis kaum kritisch auf ihre Begründung hinterfragt. (Warum z.B. benutzen die von Buth besuchten Medizinstatistiker einen konventionellen Test mit umständlicher β-Fehlerkontrolle und unnötig groß fixierter Stichprobe, wo doch dasselbe ein sparsamer und einfacherer Sequentialtest leistete?)

Den von Buth eingeschlagenen Weg fortsetzend will ich hier nun exemplarisch von einem jüngst erlebten Fall inferenzstatistischer Argumentation in den Humanwissenschaften berichten, der einen recht guten Eindruck vermitteln mag, wie von manchen Humanwissenschaftern mit Statistik umgegangen wird.

2. Der Fall

Vor einiger Zeit brachte die Psychologische Rundschau, als Organ der Deutschen Gesellschaft für Psychologie gleichsam das offizielle Zentralblatt deutschsprachiger Hochschulpsychologie ein Sonderheft zum Thema Psychotherapieforschung heraus. Dieses Thema hat für die Profession der Psychologen selbstverständlich größte Bedeutung, geht es doch letztlich dabei um die Frage, welche der vielen miteinander konkurrierenden Formen der Psychotherapie sich "wissenschaftlich" als besonders wirksam erweisen und damit den entsprechend ausgebildeten Psychotherapeuten den Zugang zu den bislang nur vom Ärztestand besetzten Töpfen der gesetzlichen Krankenkassen verheißen.

Eröffnet wurde dieses Sonderheft mit einem Beitrag - gleichsam einem Leitartikel - aus der Feder des jung-dynamischen, sehr renommierten und in vielen hochkarätigen Gremien sehr einflußreichen Professors Klaus Grawe (1992), der sich gemeinsam mit seinen Mitarbeitern seit Jahren meta-analytisch mit der Psychotherapieforschung beschäftigt. Viele der von Grawe gesichteten Studien zeigen zwar im Vergleich zwischen - wie auch immer - psychotherapeutisch behandelten Gruppen und unbehandelten Kontrollgruppen signifikante Unterschiede in relevanten Variablen; es zeigen sich aber in entsprechenden Vergleichsstudien in der Regel keine signifikanten Unterschiede zwischen Gruppen, die mit verschiedenen Formen von Psychotherapie behandelt wurden. Da die Krankenkassen kaum dazu bereit sein dürften, unterschiedslos alle Formen der Psychotherapie zu finanzieren, kann Grawe diesen Befund aber nicht einfach so stehen lassen: es muß trotz der fehlenden signifikanten Unterschiede zwischen den Therapieformen ein "wissenschaftlich" belegbarer Unterschied her!

Zu dessen Konstruktion bedient sich nun Grawe folgender inferenzstatistischer Argumentation: Üblicherweise seien die in den entsprechenden Studien verglichenen Gruppen viel zu klein, die Power der entsprechenden Signifikanztests also viel zu gering, um bei den realistischerweise überhaupt zu erwartenden, durchaus aber noch praktisch relevanten Unterschieden zwischen den Therapieformen signifikante Ergebnisse zu liefern. Dennoch könne man trotz dieser fehlenden Signifikanzen der Einzelvergleiche doch auf einen signifikanten Unterschied beispielsweise dann schließen, wenn in einer Studie, in der zwei unterschiedlich therapierte Gruppen zu jeweils mehreren Meßzeitpunkten in mehreren Variablen verglichen würden, "überzufällig" viele dieser Einzelvergleiche, obgleich nichtsignifikant, zugunsten der einen Therapieform ausfielen.

Genau dies sei aber typischerweise der Fall. So seien etwa von den 113 Mittelwertsvergleichen zwischen einer psychoanalytisch und einer gesprächstherapeutisch behandelten Gruppe in einer Studie von Meyer (1981) zwar nur wenige signifikant geworden; dennoch lasse sich trotz der fehlenden Signifikanzen die Überlegenheit der Gesprächstherapie als "hochsignifikant" belegen mit dem Hinweis darauf, daß in der untersuchten Stichprobe bei 90 Vergleichen die Gesprächstherapie und nur in 23 Vergleichen die Psychoanalyse (und sei es auch nur einen Deut) besser abschnitt, was unter der Nullhypothese gleicher Wirksamkeit beider Therapieformen in der Population im Binomialtest eine Überschreitungswahrscheinlichkeit von weniger als 0,0001 habe.

3. Die Kritik

Obwohl noch in demselben Sonderheft mehrere renommierte Forscher - insbesondere der bei Grawe schlecht abschneidenden psychoanalytischen Provenience - kritisch Stellung nahmen, moniert von diesen niemand die Fraglichkeit der von Grawe vorgebrachten inferenzstatistischen Argumentation, wie augenscheinlich zuvor auch niemand von den Herausgebern und Gutachtern der Psychologischen Rundschau, die sich dann erst einige Hefte später gezwungen sahen, unter dem ironischen Titel "Münchhausen-Statistik" einen kurzen Kommentar (Diepgen, 1993) etwa folgenden Wortlautes zu veröffentlichen:

Diese von Grawe an ganz zentraler Stelle benutzte Argumentationsfigur wäre aber nur dann stichhaltig, wenn die Einzelvergleiche stochastisch unabhängig voneinander wären. Davon kann aber keine Rede sein: Es wurden nämlich

augenscheinlich immer wieder dieselben Patienten verglichen, zu mehreren Meßzeitpunkten und in einer Fülle von vermutlich mehr oder minder stark korrelierenden Kriterien für den Therapieerfolg. Was Grawe hier betreibt, ist eine Spielart von Münchhausen- oder Bootstrap-Statistik, der es gelingt, sich aus dem Sumpf nichtsignifikanter Ergebnisse am eigenen Schopf bzw. Stiefelriemchen herauszuziehen.

Der Trick dabei: Hat man zu wenig Probanden, dann nutzt man diese zum Ausgleich halt mehrfach, indem man ihnen mehr oder minder ähnliche Fragen mehrere Male stellt. Konkret könnte eine vergleichende Therapieerfolgsstudie in Zukunft dann so aussehen: 3 (in Worten: drei) Patienten werden mit der vier Monate dauernden Psychotherapie A behandelt, 3 andere mit der ebenfalls vier Monate dauernden Therapie B. Gemessen werden nach jedem Behandlungsmonat vier, miteinander vermutlich hochkorrelierende 10-stufige Indikatoren für Befindlichkeit, etwa SA: soziale Aktivität, ES: Entspanntheit, EZ: Erfolgszuversicht und KB: körperliche Befindlichkeit. Dabei mögen sich etwa folgende Mittelwerte ergeben:

	1. Monat				2. Monat				3. Monat				4. Monat			
	SA	ES	EZ	KB	SA	ES	EZ	KB	SA	ES	EZ	KB	SA	ES	EZ	KB
Therapie A	5,5	7,3	4,2	6,5	6,0	8,1	6,2	6,3	5,9	8,3	6,9	5,8	6,1	8,2	7,1	5,7
Therapie B	5,2	6,5	4,1	6,0	5,3	6,7	4,5	6,2	5,5	7,2	6,6	5,7	8,8	9,3	9,0	8,5
Besser	A	A	A	A	A	A	A	A	A	A	A	A	B	B	B	B

Kein Einzelvergleich ist hier aufgrund der Miniaturstichproben signifikant, aber in 12 von 16 Einzelvergleichen gewinnt die Therapie A, was im Binomialtest eine Überschreitungswahrscheinlichkeit von weniger als 0,05 hätte, nach der Logik von Grawe also als ein signifikanter Beleg für die Überlegenheit von Therapie A zu interpretieren wäre! Diese - vom Stichprobenumfang gänzlich unabhängige - Überschreitungswahrscheinlichkeit ließe sich übrigens durch Aufnahme immer weiterer - von den schon gemessenen kaum unabhängigen - Befindlichkeitsindikatoren beliebig verringern. Und das Zwischenschieben weiterer Meßzeitpunkte dürfte ähnliches bringen: Denn wer, wie hier die Gruppe A, - und sei es rein zufällig - einen guten Start erwischt hat, wird seinen Vorsprung erst allmählich verlieren.

4. Die Abwehr der Kritik

Wie sieht nun Grawes abschließende Stellungnahme zu dieser Kritik aus? Ihrer für die Humanwissenschaften so typischen Vermischung inhaltlicher und mathematisierter Sprachebenen wegen sei wörtlich zitiert:

"Viele statistische Verfahren setzen Unabhängigkeit zwischen den einzelnen Ereignissen oder Messungen voraus. Die Annahme stochastischer Unabhängigkeit kann nun auf die verschiedenste Weise mehr oder weniger stark verletzt werden. Die Entscheidung darüber, ob die Voraussetzung der Unabhängigkeit als erfüllt angesehen werden kann, ist daher immer im Hinblick auf die spezifische inhaltliche Aussage zu prüfen, deren Validität durch die Verletzung der Unabhängigkeitsannahme beeinträchtigt werden könnte, und sie ist nur selten im Sinne eines 'entweder-oder', sondern meist im Sinne eines 'mehr oder weniger' zu beantworten.

Ich will dies an einem auf Diepgens Kritik bezogenen Beispiel erläutern: Wenn zwischen zwei Behandlungsgruppen im Mittelwert in Merkmal 1 ein Unterschied zugunsten von Behandlung A gegenüber Behandlung B besteht, dieser Unterschied im Verhältnis zur Binnenvarianz jedoch nicht groß genug ist, um die statistische Signifikanz zu überschreiten, dann ziehen wir nach den Regeln der Statistik die Schlußfolgerung, daß der beobachtete Unterschied im Zufallsbereich liegt. Kommen wir für ein zweites gemessenes Merkmal zu genau demselben Ergebnis, dann werden wir immer noch bei der Schlußfolgerung bleiben, daß sich die beiden Behandlungsgruppen nicht unterscheiden. Was aber ist, wenn sich für ein drittes, viertes, fünftes .. ntes Merkmal immer dasselbe Ergebnis zeigte, daß jedesmal A einen größeren Wert hat als B? Würden wir eine solche Häufung auf der einen Seite immer noch als zufällig betrachten können? Natürlich, wenn die Merkmale 1, 2, ... n immer wieder dasselbe erfaßten, so wäre der Eindruck des nicht mehr mit dem Zufall Vereinbaren ein bloßes Artefakt redundanter Information. Wie aber wäre es, wenn die Merkmale etwas Unterschiedliches erfaßten? Müßten wir dann nicht schlußfolgern, A und B unterschieden sich zwar nicht in ihren Einzelmerkmalen, wohl aber in ihrer Gesamtheit?

Die Berechtigung für diese Schlußfolgerung ist offensichtlich abhängig von dem Ausmaß, in dem die Merkmale 1 ... n etwas Unterschiedliches oder etwas Gleiches messen. Dies ist nun eine inhaltliche und empirische Frage. Inhaltlich nehmen Untersucher, die viele Aspekte der Wirkungen von Therapien zu erfassen versuchen, offensichtlich an, daß sie unterschiedliche Merkmale

messen, denn was würde es für einen Sinn machen, 30 oder 50mal dasselbe zu messen? Empirisch läßt sich die Frage der Abhängigkeit-Unabhängigkeit auf der Basis der Interkorrelationen zwischen den Merkmalen beantworten. Korrelieren die einzelnen Merkmale untereinander sehr hoch, dann ist ein Vergleich über alle Merkmale offensichtlich redundant und die Voraussetzung für einen Test, der Unabhängigkeit zwischen den einzelnen Messungen voraussetzt, wäre grob verletzt. In dem Ausmaß, in dem sich die Korrelationen zwischen den Einzelmerkmalen jedoch einer Nullkorrelation annähern, wären die Voraussetzungen für einen Test wie etwa den Binomialtest erfüllt.

In den Therapiestudien, die ich selbst bisher durchgeführt habe und in denen regelmäßig über 50 Merkmale gemessen wurden, betrug die durchschnittliche gemeinsame Varianz zwischen den zur Veränderungsmessung eingesetzten Merkmale zwischen 10 und 20% ... . Soweit dies aus den einschlägigen Veröffentlichungen ersichtlich ist, sind auch in anderen Therapiestudien die Interkorrelationen zwischen den Effektmaßen durchschnittlich relativ niedrig. Es besteht demnach zwar keine völlige Unabhängigkeit zwischen den Merkmalen, wie es für einen Binomialtest wünschenswert wäre, aber die einzelnen Variablen messen doch in einem so hohen Ausmaß Verschiedenes, daß die Information, die in der Gesamtheit der Merkmale vorliegt, weit über die Information hinausgeht, die mit jedem Einzelmerkmal gegeben ist.

Die tatsächliche Sachlage ist also weit entfernt von dem Szenario, mit dem Diepgen mein Vorgehen ad absurdum zu führen versucht. Ich halte es durchaus nicht für absurd, ein statistisches Kriterium dafür zu entwickeln, ab wann die Gesamtinformation eines Datensatzes weit genug über die in den einzelnen Merkmalen enthaltene Information hinausgeht, um einen statistischen Test durchführen zu können, der diese Gesamtinformation ausschöpft, ohne all die Voraussetzungen erfüllen zu müssen, die für die Anwendung multivariater parametrischer Verfahren gegeben sein müssen, der aber doch die in den Daten enthaltene Redundanz für den Signifikanztest berücksichtigt. Ich gebe gerne zu, daß ich diese Redundanz für den mit den Daten der Meyer-Studie von mir durchgeführten Binomialtest nicht berücksichtigt habe, weil ein elaboriertes Kriterium dafür bisher fehlt. Die einseitige Häufung der Überlegenheiten zugunsten der Gesprächspsychotherapie in der Meyer-Studie ist aber so eindeutig, daß sich an der inhaltlichen Ergebnisaussage nicht das Geringste ändern würde.

Als methodische Anmerkung ist Diepgens Kommentar sicherlich berechtigt. Allerdings bleiben seine Überlegungen sehr an der Oberfläche des angeschnittenen Problems, das eine gründliche Bearbeitung durchaus lohnte. Meine inhaltliche Argumentation ist von dieser methodischen Frage allerdings überhaupt nicht betroffen. Man könnte den von mir durchgeführten Signifikanztest einfach fortfallen lassen und sich für die inhaltliche Argumentation auf die rein deskriptive Information beschränken." (Grawe, 1993, S. 184ff).

5. Kommentar

Geschrieben wurde diese Stellungnahme - wie gesagt - für eine große und ob des brisanten Themas besonders interessierte Fachöffentlichkeit von einem erfolgreichen, also mit der Denkweise seiner Kollegen vermutlich bestens vertrauten Psychologieprofessor, der sich überdies seit Jahren mit Meta-Analysen befaßt und als Leiter größerer Forschungsprojekte jederzeit dazu in der Lage ist, sich Beratung durch mathematisch-statistische Experten einzukaufen. Es dürften sich daher in dieser - selbstverständlich auf die Zustimmung der Kollegen zielenden - Stellungnahme ganz typische Denk- und Argumentationsweisen von Humanwissenschaftern dokumentieren. Was läßt sich darüber diesem interessanten Dokument entnehmen? Unter anderem dies:

1. Obwohl es letztlich um die peinliche meta-analytische Problematik geht, was denn nun eigentlich aus einer durch eine Vielzahl von teils signifikanten, teils nichtsignifikanten Befunden, also aus einer durch eine Vielzahl unterschiedlicher "Entscheidungen" für oder gegen Hypothesen charakterisierten Empirie zu folgern ist, wird das von Neyman und Pearson ausdrücklich nur für eine Einzelentscheidung konzipierte Konzept des Signifikanztestes überhaupt nicht grundsätzlich in Frage gestellt, sondern trotz seines offensichtlichen Ungenügens in dieser Situation nach wie vor als methodisches Muß für "Wissenschaftlichkeit" akzeptiert.

2. Dieses Muß ist sogar so drängend, daß nun versucht wird, auch noch die - im Unterschied etwa zur Bayes-Statistik - von Neyman und Pearson grundsätzlich nicht vorgesehene Integration vieler in einzelnen Hypothesentests durchgeführten "Entscheidungen" auf der Metaebene wiederum als "Signifikanztest" zu organisieren.

3. Dieser meta-analytische "Signifikanztest" kann dann freilich ausschließlich in Fisherschen Begrifflichkeiten konzipiert werden: Es wird lediglich ex post danach gefragt, wie wahrscheinlich denn die längst bereits vorliegende oder eine extremere Empirie wäre, wenn denn die ex post formulierte Nullhypothese gälte.

4. Trotz dieser rein Fisherschen Begrifflichkeit auf der Metaebene wird auf der Ebene der rezipierten Einzeluntersuchungen durchaus mit dem Fisher ganz fremden Neyman-Pearson-Konzept der Teststärke argumentiert. Hierbei fällt folgendes auf: Es wird zwar einerseits auf die eigentlich zu geringe Teststärke der einzelnen Signifikanztests verwiesen, andererseits wird die Durchführung dieser Signifikanztests selbst überhaupt nicht kritisiert. Kurzum: Es wird - zumindest für den um "Wissenschaftlichkeit" besonders ringenden Bereich der Psychotherapieforschung - selbstverständlich an der Norm festgehalten. Untersuchungen mittels Signifikanztests auszuwerten, obwohl diese Tests zugestandenermaßen angesichts der akzeptiert kleinen Stichproben überhaupt nicht dazu in der Lage sein können, für relevant gehaltene Abweichungen von der Nullhypothese mit befriedigender Wahrscheinlichkeit zu entdecken.

5. Die Konstruktion des "Signifikanztests" auf der Metaebene wird - angesichts der öffentlich vorgetragenen Kritik aus der Feder eines Mathematikers - zu begründen versucht durch sehr vage formulierte, teilweise in mathematischer Begrifflichkeit nicht mehr rekonstruierbare, allenfalls intuitive Vorstellungen, beispielsweise der suggestiven Interpretation des mathematischen Begriffs der "Korrelation" zweier Variablen mit deren "Informationsgleichheit".

Besonderen Aufwand muß diese Konstruktion darauf verwenden, die angemahnte stochastische Unabhängigkeit der Einzelversuche plausibel zu machen. Denn während die stochastische Unabhängigkeit bei Einzelexperimenten relativ problemlos durch den die Stichprobenziehung organisierten Zufallsgenerator gesichert werden kann, wird sie für ein solches "Metaexperiment" selbstverständlich zum (tatsächlich unüberwindlichen) Problem. Deshalb geht es hier nicht mehr ohne mathematisch unhaltbare Behauptungen. Die stochastische Abhängigkeit zweier Variablen schlicht mit ihrer Korrelation gleichzusetzen, ist augenscheinlich ein bei Humanwissenschaftern weitverbreiteter Irrtum, denn anders wäre die Bemerkung nicht zu verstehen: "In dem Ausmaß, in dem sich die Korrelationen zwischen den Einzelmerkmalen jedoch einer Nullkorrelation annähern, wären die Voraussetzungen für einen Test wie etwa den Binomialtest erfüllt."

Aus einer Nullkorrelation der Einzelmerkmale zu einem Zeitpunkt folgt aber tatsächlich - vom Sonderfall einer gemeinsamen Normalverteilung abgesehen - allgemein weder deren stochastische Unabhängigkeit, noch die Unabhängigkeit von Mittelwertsvergleichen zwischen zwei Gruppen im Hinblick

auf diese Merkmale zu einem Zeitpunkt. Und überhaupt nicht folgt aus der Nullkorrelation von verschiedenen Merkmalen zu einem Zeitpunkt, daß die Mittelwertsvergleiche zweier Gruppen hinsichtlich eines Merkmales über mehrere Zeitpunkte hinweg stochastisch unabhängig wären - eine Problemdimension, die schlicht unterschlagen wird. Folgen diese stochastischen Unabhängigkeiten schon nicht aus Nullkorrelationen, so folgen sie schon ganz und gar nicht aus den zugestandenen Korrelationen in einer Höhe bis zu etwa $\sqrt{0,2} \approx 0{,}45$.

Kurzum: Der Hinweis auf diese "niedrigen" Korrelationen ist tatsächlich überhaupt kein den durchgeführten "Binomialtest" legitimierender Beleg für die stochastische Unabhängigkeit der vielen Mittelwertsvergleiche; die zugestandenen Korrelationen in dieser Größenordnung sind vielmehr mit sehr hohen stochastischen Abhängigkeiten zwischen diesen Mittelwertsvergleichen vereinbar. Die "berechnete" Überschreitungswahrscheinlichkeit im Metaexperiment ist und bleibt pure Erfindung.

6. Während üblicherweise in den Humanwissenschaften die Notwendigkeit von Signifikanztests legitimiert wird gerade aus der Sorge davor, nur zufällige Abweichungen einer untersuchten Zufallsstichprobe von der eigentlich interessierenden Population fälschlicherweise für repräsentativ für die Population zu halten, wird genau diese Sorge von der Konstruktion des Signifikanztestes auf der Metaebene wieder völlig vergessen: Denn durch nichts wird hier ersichtlich die Wahrscheinlichkeit kontrolliert, daß die "signifikante" Überlegenheit der einen über die andere Therapie zufälligerweise nur für die - beliebig kleine (!) - Stichprobe der therapierten Probanden gilt. Man kontrolliert hier auf der Metaebene also mit großem Aufwand und vermeintlicher Präzision Fehlerwahrscheinlichkeiten in der einen Dimension, während man sich über die möglicherweise viel größeren Fehlerwahrscheinlichkeiten in der anderen Dimension überhaupt keine Gedanken mehr macht.

7. Die Mathematisierung der Unabhängigkeitsproblematik in diesem Metaexperiment führt etwa zu den unten im "Nachtrag" präzisierten Fragen, die, wie man alsbald bemerken wird, aus dem "Gegebenen" ganz und gar nicht zu beantworten sind. Nur eine Fülle weiterer überaus problematischer und kaum belegbarer Zusatzannahmen würde es hier überhaupt erst erlauben, das Modell so weit zu präzisieren, daß die Fragen beantwortbar würden - und dies zumeist auch nur mittels Simulation. Es ist verblüffend, mit welcher Selbstverständlichkeit Grawe als Humanwissenschafter ohne jede Begründung auf diese hochkomplexen Fragen bestimmte Antworten vorauszusetzen scheint.

8. Und noch verblüffender schließlich ist, wie sich dieser Humanwissenschafter dann dennoch wieder in seinen letzten Sätzen vom eigenen - augenscheinlich völlig entfremdeten - statistischen Tun distanziert: "Meine inhaltliche Argumentation ist von dieser methodischen Frage allerdings überhaupt nicht betroffen. Man könnte den von mir durchgeführten Signifikanztest einfach fortfallen lassen und sich für die inhaltliche Argumentation auf die rein deskriptive Information beschränken." So erfrischend deutlich gibt selten ein Forscher die Überflüssigkeit und den Unernst der von ihm augenscheinlich dann nur noch als Rhetorik inszenierten Inferenzstatistik zu. Die Halböffentlichkeit dieses Geständnisses verweist auf ein augenzwinkerndes und solidarisches Einverständnis unter den Forscherkollegen: Wir wissen doch alle, daß es uns mit unserer Statistik nicht Ernst ist, sondern daß wir sie lediglich als "wissenschaftliches" Ritual nach außen hin inszenieren!

6. Und?

Ich denke, man könnte aus dieser Episode durchaus ein kleines Projekt für den Stochastikunterricht gewinnen: Mit mathematischen und nichtmathematischen Aufsätzen ausgehend von einer "wirklichen Problemstellung", Übungen zur mathematischen Modellierung nur vage formulierter Vorstellungen, Analysen von Interessen, Computersimulationen zur Abschätzung von Abhängigkeiten und vor allem erlebnisreichen Spaziergängen an den seriös kaum noch zu überwindenden engen Grenzen mathematisch-stochastischer Argumentation. Im Reisegepäck benötigt man: den Binomialtest samt seiner Teststärke, den Begriff der stochastischen Unabhängigkeit, den Begriff der Produkt-Moment-Korrelation und einen vom professionellen Interesse mancher Mathematiker unverstellten Blick auf eine bestimmte soziale Wirklichkeit.

7. Nachtrag

Der - ganz und gar nicht eindeutige und zwingende - Versuch der Mathematisierung der fraglichen Unabhängigkeitsproblematik könnte plausiblerweise etwa so aussehen:

Gegeben: Gegeben seien in einer Population zwei metrische Zufallsvariablen X und Y mit unbekannter Verteilung, aber mäßig positiver Korrelation ρ. In der Population existieren zwei Subpopulationen a und b, in denen die Variablen X und Y jeweils die spezifischen unbekannten Erwartungswerte μ_{X_a} und μ_{X_b} bzw. μ_{Y_a} und μ_{Y_b} haben. Gezogen wird nun aus beiden Subpopulationen

jeweils eine Zufallsstichprobe vom Umfang N, und für jede Stichprobe werden jeweils berechnet die Stichprobenmittelwerte m_{X_a}, m_{X_b}, m_{Y_a}, bzw. m_{Y_b}. Es sei nun E_X das Ereignis, daß die Variable X in der Stichprobe aus der Subpopulation a einen größeren Mittelwert hat als in der Stichprobe aus der Subpopulation b, formal $m_{X_a} > m_{X_b}$, und analog E_Y das Ereignis, daß die Variable Y in der Stichprobe aus der Subpopulation a einen größeren Mittelwert hat als in der Stichprobe aus der Subpopulation b, formal $m_{Y_a} > m_{Y_b}$.

Gefragt: Sind die beiden Ereignisse E_X und E_Y stochastisch unabhängig? Wenn nicht: Welchen Zusammenhang mag es geben zwischen der stochastischen Abhängigkeit zwischen diesen beiden Ereignissen einerseits, dem unbekannten Korrelationskoeffizient ρ und dem bekannten Stichprobenumfang N andererseits?

Das Gegebene könnte sicherlich im Unterricht der Sekundarstufe II in der Analyse der Originaltexte erarbeitet werden als eine zumindest naheliegende mathematische Präzisierung ungefähr dessen, was Grawe zugesteht, ebenso wie das Gefragte als eine Fragestellung, auf die Grawe eine ganz bestimmte Antwort unterstellt. Aber anders, als Grawe unterstellt, ist tatsächlich aus diesem Gegebenen das Gefragte leider überhaupt nicht zu beantworten. Im Unterricht mag man zumindest das völlige Fehlen einer Begründung für die von Grawe unterstellte Unabhängigkeit analytisch herausarbeiten oder doch zumindest intuitiv nachempfinden. Der Wunsch schließlich, das Ausmaß der stochastischen Abhängigkeit in Beziehung zu ρ und N abzuschätzen, läßt sich dann wohl nur um den Preis vieler zusätzlicher und überaus fraglicher Modellannahmen mittels Computersimulation befriedigen - und dies auch kaum abschließend.

Erlebt werden kann dabei sicherlich auf jeden Fall, wie überaus eng die Grenzen stochastischer Modellierung angesichts 'realer' Probleme sind, wie überaus hypothetisch alle Annahmen und wie überaus groß schließlich die dunklen Flecken seriös kaum noch überwindbarer Unwissenheit.

Literatur

Biehler, R. (1982): *Explorative Datenanalyse. Eine Untersuchung aus der Perspektive einer deskriptiv-empirischen Wissenschaftstheorie,* IDM Materialien und Studien Bd. 24, Bielefeld: Universität Bielefeld.

Buth, M. (1993): Zum Thema "Testen von Hypothesen": Was man aus der Forschungspraxis für die Schule lernen kann, *Stochastik in der Schule* **13**, Nr. 2, 35-46.

Cohen, J. (1969): *Statistical Power Analysis for the Behavioral Sciences*, New York: Academic Press.

Cohen, J. (1990): Things I have learned (so far), *American Psychologist* **45**, 1304-1312.

Diepgen, R. (1992): Objektivistische oder subjektivistische Statistik? Zur Überfälligkeit einer Grundsatzdiskussion, *Stochastik in der Schule* **12**, Nr. 3, 48-54.

Diepgen, R. (1993): Münchhausen-Statistik. Eine Randbemerkung zu einer Argumentationsfigur von Grawe (1992), *Psychologische Rundschau* **44**, 176-177.

Gigerenzer, G. und Murray, D.J. (1987): *Cognition as Intuitive Statistics*, Hillsdale, N.J.: Lawrence.

Grawe, K. (1992): Psychotherapieforschung zu Beginn der neunziger Jahre, *Psychologische Rundschau* **43**, 132-162.

Grawe, K. (1993): Über Voraussetzungen eines gemeinsamen Erkenntnisprozesses in der Psychotherapie. Eine Erwiderung auf Eysenck und Diepgen, *Psychologische Rundschau* **44**, 181-186.

Meyer, A.-E. (Hrsg.) (1981): The Hamburg Short Psychotherapy Comparison Experiment, *Psychotherapy and Psychosomatics* **35**, 81-207.

Schmitz, N. (1993): Leserbrief, *Stochastik in der Schule* **13**, Nr. 2, 49-50.

Dr. Raphael Diepgen, Fak. f. Psychologie, Ruhruniversität Bochum, Universitätsstraße 150, D-44780 Bochum

Statistische Glaubensbekenntnisse

Es gibt in der Statistik einige Konfessionen und es ist gut, wenn man weiß, welcher Konfession die Person, mit der man spricht, angehört. Da sind die Protestanten, das sind die klassischen Statistiker, dann die Katholiken, das sind die Bayesianer. Darüberhinaus gibt es Agnostiker, das ist eine sehr kleine Sekte. Sie glauben an keine Verteilung, weder an eine für die Daten noch an eine für die Parameter a priori. Man nennt sie auch Permutationisten, denn für sie gibt es Wahrscheinlichkeit nur, wenn sie durch Randomisierung (d.h. durch zufällige Auswahl) entstanden ist, das ist der nicht-parametrische Ansatz. In der Fachwissenschaft sind intensive Gespräche im Gang, um einen ökumenischen Ansatz zu finden.

* * * * *

In der Didaktik scheint man ein wenig nachzuhinken und trägt erst jetzt den Glaubenskrieg so richtig aus. Wie mit den Religionen ist es auch mit der Statistik: Es steckt in allen ein zutreffender Kern, wenn er aber mit Macht verbunden ist, wird er verzerrt.