Das statistische Paralleluniversum der BWL

Peter v. der Lippe (Sept. 2014)

Unter einem Paralleluniversum oder einer Parallelwelt versteht man einen nach außen abgegrenzten Bereich, in dem sich das Leben bestimmter Personen oder Gruppen unabhängig von der Außenwelt abspielt. In diesem Sinne scheint sich mir das Leben (genauer das Arbeiten mit "Statistik") von Statistikanwendern, insbesondere von solchen in der Betriebswirtschaftslehre ganz ohne störende Einflüsse von Statistikern abzuspielen. In der statistischen Parallelwelt B der BWLer ist falsch, was in der Welt S der Statistiker richtig ist und was in S falsch ist, ist in B richtig. Was hier oben ist, ist dort unten usw. Es gelten quasi andere Gesetze und es wird in den Welten eine ganz andere Sprache gesprochen (die gleichen Worte haben nicht die gleiche Bedeutung und auch grundlegende Konzepte werden inhaltlich völlig anders interpretiert). Das wäre für sich genommen nicht so dramatisch und es würde vielleicht nur etwas verwirrend und gespenstisch anmuten, etwa so wie der Besuch einer Irrenanstalt, wenn nicht noch etwas hinzukäme, dass nämlich viele BWLer ganz offensichtlich der festen Überzeugung sind, von Statistik mindestens so viel, wenn nicht mehr zu verstehen als die Statistiker.

Zumindest ist es das, was mir erneut klar wurde, als ich – leider etwas naiv und nicht ahnend, was mir blühen würde – den Text "Repräsentativität, convenience samples und Signifikanztests" (im Folgenden RCS) im März 2014 bei einer betriebswirtschaftlichen wissenschaftlichen Zeitschrift einreichte. Was ich dort erleben musste habe ich so bisher noch nie erlebt. Dass der Text abgelehnt wurde, ist dabei eigentlich noch am wenigsten wichtig. Viel interessanter scheint mir zu sein, *wie* das (in welcher Geisteshaltung einem Statistiker gegenüber) und mit welcher Begründung das geschah.

Es geht mir im Folgenden deshalb auch nicht oder nicht primär darum, mich über das Verhalten des Gutachters und des Herausgebers zu beklagen (so etwas dürfte niemand interessieren, auch wenn es vielleicht nicht ein Einzelfall war),[1] sondern darum, die Charakteristika des statistischen Paralleluniversums der BWLer, wie sie sich mir erneut durch dieses Erlebnis darstellten, herauszuarbeiten (in der Hoffnung, dass dies vielleicht einige Leser [es wäre schön, wenn auch BWLer darunter wären] interessieren könnte und ihnen vielleicht sogar zu denken geben könnte [was ich kaum zu hoffen wage]).

Um das Besondere am Paralleluniversum anschaulich zu machen zitiere ich – entgegen meiner Gewohnheit (aber ermutigt durch einen sehr bekannten Statistikerkollegen aus Dortmund) – aus einem privaten Schriftwechsel, um alles sozusagen mit "O-Tönen" untermalen zu können. Ich habe dabei an einigen Stellen diese O-Töne gekürzt um Hinweise zu vermeiden, aus denen man vielleicht erkennen könnte, um welche spezielle BWL es hier ging.

Was das Paralleluniversum betrifft, so stelle ich im Folgenden einige Punkte (insgesamt zehn) heraus, die mir charakteristisch zu sein scheinen. Als erster Punkt halte ich für sehr wichtig:

P1. Ich bin bisher immer davon ausgegangen, dass ein Professor für X mehr von X versteht als ich, wo mein Fach Y ist. Ich würde mich z.B. nie auf eine Diskussion über Topologie mit einem Mathematiker einlassen, schon allein deshalb nicht, weil ich Angst hätte, mich mit einer laienhaften Äußerung zu blamieren. Diese vielleicht etwas altmodische Geisteshaltung scheint für BWLer bei der Kommunikation mit Statistikern über Statistik nicht zu gelten. Das unerschütterliche Selbstbewusstsein mit dem man Statistiker abkanzelt und die Unempfindlichkeit gegenüber Peinlichkeiten haben mich immer wieder überrascht (und im wahrsten Sinne "befremdet"). [2]

Ich möchte ausdrücklich betonen, dass mich nicht erst die Erfahrung mit der Einreichung des Papiers RCS zu den Aussagen **P1** bis **P10** motivierten, sondern auch viele schon früher gemachte Erfahrungen die aber nur mündlich kommuniziert wurden, und die zu zitieren naturgemäß kaum möglich ist. Bevor ich auf weitere Beobachtungen (P2 usw.) eingehe, möchte ich noch kurz ausführen, welche Absicht

[1] Ich mache einen deutlichen Unterschied zwischen dem Gutachter, dessen Verhalten völlig inakzeptabel ist, und dem des Herausgebers, dem ich immerhin hoch anrechne, dass er von sich schrieb "obwohl ich ein absoluter statistischer Laie bin", was ihn allerdings auch nicht davon abhielt, meine statistischen Ausführungen für falsch zu halten und was ihn auch nicht auf die Idee brachte, sich einmal bei einem Statistiker Rat zu holen.

[2] Das wird jedoch in gewisser Weise (wie weit traut man sich mit statistischen Äußerungen konkret zu werden und in Details zu gehen?) in meinen Thesen P2 und P3 etwas relativiert.

ich mit meinem RCS Papier hatte. Es ging mir letztlich nur darum, die folgende Warnung von J. Bortz, vor "sog. 'anfallenden' oder 'ad hoc' Stichproben (z.B. die 'zufällig' in einem Seminar anwesenden Teilnehmer)[3] in der Hoffnung, auch so zu aussagefähigen Resultaten zu gelangen" zu begründen:[4]

> "Vor dieser Vorgehensweise sei nachdrücklich gewarnt. Zwar ist die Verwendung inferenzstatistischer Verfahren nicht daran gebunden, dass eine Stichprobe aus einer wirklich existierenden Population gezogen wird: letztlich lässt sich für jede 'Stichprobe' eine fiktive Population konstruieren, für die diese 'Stichprobe' repräsentativ erscheinen mag. Die Schlüsse, die aus derartigen Untersuchungen gezogen werden, beziehen sich jedoch nicht auf real existierende Populationen und können deshalb wertlos sein" (Bortz, S. 85)

und zwar ausdrücklich so zu begründen, dass es vielleicht von Anwendern leichter zu verstehen ist als wenn es in der sonst unter Statistikern üblichen mehr "formalen", mathematisch "technischen" Art erfolgt wäre. Konkret habe ich z.B. gefragt,

> "… welchen Sinn "Hypothesen" (die ja immer Vermutungen über die GG sind) haben, wenn es mehr oder weniger unklar bleibt, was genau die GG ist, für die die Stichprobe "repräsentativ" sein soll"[5]

Ich dachte – vielleicht etwas naiv – dass ich damit ganz "hehre" Absichten hatte, und dass ich deshalb auch eine andere Behandlung verdient hätte, als die die ich dann leider erleben musste.

Am 9.4.2014 erhielt ich vom Herausgeber die folgende Stellungnahme des Gutachters

> *Die Autoren befassen sich im eingereichten Manuskript mit der für Wissenschaftler zweifelsohne bedeutsamen Frage, welchen Anforderungen eine Stichprobe genügen sollte. Während ich die von den Autoren - freilich nicht explizit - zum Ausdruck gebrachte Einstellung (dass eine Vielzahl an statistisch zweifelhaften Arbeiten teilweise sogar erstklassig publiziert wird) teile, so habe ich dennoch meine Zweifel, ob und in welcher Form dieser Artikel publikationsfähig ist. Der erste Teil - die Kapitel 1 und 2 - enthalten nur Erkenntnisse, die der "gelehrte" Wissenschaftler bereits kennt und der "ungelehrte" Anwender nicht lesen bzw. verstehen mag. Nachdem die Autoren zu dem in der Wissenschaft allgemein anerkannten Schluss kommen, dass die Repräsentativität einer Stichprobe ausschließlich vom Ziehungsmechanismus abhängt bzw. über diesen definiert ist (und durch sonst nichts), verbietet sich eigentlich jede Diskussion darüber, welche anderen Assoziationen man fälschlicherweise mit Repräsentativität noch verbinden könnte, selbst wenn dies gelegentlich von einigen Anwendern/Praktikern so ins Feld geführt werden mag.*
>
> *Auch Kapitel 3 bleibt für den statistisch einigermaßen vorgebildeten Leser weitgehend Überraschungsfrei: Es dürfen eben Signifikanztests nur dann angewendet werden, wenn eine Zufallsstichprobe vorliegt und ein Stichprobenfehler zu ermitteln ist.*
>
> *Spannend wird das Manuskript meines Erachtens lediglich in Abschnitt 4, der aber leider relativ kurz ausfällt und zudem nur einige wenige Beispiele für schlechte wissenschaftliche Praxis liefert.*
>
> *Für eine eventuelle Überarbeitung habe folgende Anregungen:*
>
> *1. Die Diskussion des Repräsentativitätsbegriffs wäre m.E. eher in einem Lehrbuch (oder in einer Transferzeitschrift; es sind ja häufig Praktiker, die mit Repräsentativität falsche Assoziationen verbinden) gut aufgehoben und sollte entfernt werden. Man kann dabei durchaus auf die zwischen Tehroei und Praxis unterschiedlichen Anforderungen eingehen: Der Anwender will einen "treffsicheren" Punktschätzer, während der Statistiker ange-*

[3] Genau das ist es was oft, und so auch im Folgenden, "convenience sample" genannt wird.
[4] Siehe RCS, S. 2. Die Abkürzung GG steht für Grundgesamtheit.
[5] RCS, S.3. Ich habe dort auch ausdrücklich geschrieben "Wir mussten deshalb in Abschn. 3 versuchen, eigene Wege zu gehen", Aber wie so vieles konnte ich auch dies immer wieder und wieder sagen, es wurde mir vom Gutachter und Herausgeber einfach nicht geglaubt. Auch im Schriftwechsel mit dem Herausgeber musste ich z.B. feststellen, dass ich mindestens dreimal ausführte, dass Schätzen und Testen zwei Seiten einer Medaille sind (indem eine Gleichung einmal nach dieser und einmal nach jener Größe aufgelöst wird) und dass dies dann immer noch für falsch gehalten wurde. Ich hätte mich genauso gut mit der Wand unterhalten können.

sichts P(X=x)=0 nur ein hinreichend enges Konfidenzintervall als Gütebeweis anbieten kann.[6]

2. Das Framing des Artikels müsste komplett anders gewählt werden. Zunächst müsste eine (relativ umfassende) Bestandsaufnahme fehlerhaften Umgangs mit der Statistik in … oder anderen Wissenschaftsfeldern belegen, dass die Autoren nicht nur vereinzelte Belege für die Unachtsamkeit von Gutachtern aufgegriffen haben und daraus die Notwendigkeit eines Appells zu korrektem Arbeiten ableiten. Den Hinweis auf die – unangebrachte – Geringschätzung deskriptiver Ergebnisse finde ich sehr positiv. Man könnte sich an der Unterscheidung der Studien nach deskriptivem, explorativem und induktivem Charakter orientieren und aus meiner Sicht gerne ein Plädoyer für die aus meiner Sicht vielfach zutreffende Nützlichkeit von deskriptiven und explorativen Studien halten. Allerdings müsste die Nützlichkeit in geeigneter Weise belegt werden.

Es ist schon sehr bitter, wenn man sich über Wochen um eine ausführliche und verständliche Begründung dafür bemüht, ***warum*** man bei sog. convenience samples statistische Tests nicht anwenden sollte, und man dann von jemand, der offensichtlich wenig Ahnung von Statistik hat, gesagt bekommt, dies alles sei trivial und längst bekannt. Aber diese Behauptung ist nicht nur falsch,[7] sie steht auch im Widerspruch zur Position des Herausgebers, der sich über Monate in mehreren Schreiben (siehe unten) immer wieder darum bemühte, zu begründen, warum man in einer solchen Situation gleichwohl mit statistischen Tests rechnen darf (also etwas tun darf, was der "gelehrte" oder einigermaßen vorgebildete Wissenschaftler gar nicht täte). Ein zweites Kennzeichen des Paralleluniversums ist wohl auch:

P2. Im Paralleluniversum besteht oft kein Bedürfnis nach einer Begründung[8] *"Es dürfen eben Signifikanztests nur dann angewendet werden, wenn eine Zufallsstichprobe vorliegt … "* (basta). So etwas reicht offenbar als Begründung völlig aus.

Es ist weiter auffallend, dass das Gutachten bemerkenswert unkonkret ist. Es gibt darin (und auch in den Briefen des Herausgebers) an keiner Stelle eine Bemerkung zu einer Formel, zu den Beispielrechnungen oder auch zu inhaltlichen Feststellungen des Papiers (auch nicht zu solchen, die eher ungewöhnlich sind, wie z.B. die in Abschn. 3.6 behauptete Verwandtschaft zwischen dem Gedanken hinter dem "Strukturkonzept" und der sog. gambler's fallacy). Besonders bezeichnend ist, dass an keiner Stelle im Gutachten oder in den Schreiben des Herausgebers das Konzept der *Stichprobenverteilung* einer Schätzfunktion wie $\bar{x}$ auftritt, und das obgleich man ihm in RCS auf Schritt und Tritt begegnet und obgleich es das zentrale Element meiner Begründung war.[9] Das legt ein weiteres Charakteristikum des Paralleluniversums nahe:

P3. Die Argumentation in Sachen Statistik ist im statistischen Paralleluniversum der BWL auffallend unkonkret und ausschließlich verbal. Es wird nicht auf einzelne Gleichungen, nicht auf statistische Fachausdrücke und auch nicht auf mathematische Details eingegangen. Die Ablehnung von Äußerungen von Statistikern ist gleichwohl nicht weniger schroff und total.

Diese Feststellung P3 passt gut zu P2. Dass Begründungen oder eine Bezugnahme auf Mathematik eher selten vorkommen wenn sich ein BWLer in Sachen Statistik äußert kann zwei Gründe haben,

1. man traut sich nicht, weil man sich mit mehr Mathematik und mehr Details leichter auf Glatteis begibt, oder weil

---

[6] Diesen Satz habe ich im Schriftwechsel wiederholt kritisch kommentiert, was dem Gutachter auch mitgeteilt wurde. Der Gutachter ist nie darauf eingegangen. Es musste ihm also eigentlich klar gewesen sein, dass er sich blamiert hat, aber es hat ihm offenbar nichts ausgemacht.

[7] Dass Signifikanztests nur bei Zufallsstichprobe zulässig sind ist *nicht* allgemein bekannt. Ich habe nicht lange suchen müssen, um in Schriften von BWLern Beispiele für Signifikanztests ohne Zufallsstichprobe zu finden.

[8] Das ist insofern beachtlich, weil das Anliegen von RCS ja gerade darin bestand, eine Begrünung zu liefern.

[9] Es ist auch auffallend, dass es in der einschlägigen Literatur von Autoren aus der BWL kaum vorkommt. Ich fand nur eine Stelle, in der es vorkam, und zwar in dem zitierten Buch von A. Kuss, wo jedoch beispielhaft nur solche Stichproben aufgeführt waren, bei denen das arithmetische Mittel der Stichprobe gleich dem der Grundgesamtheit war. Ich habe das nicht weiter thematisiert (man ist ja höflich), aber das trifft natürlich nicht das Besondere einer Stichprobenverteilung als Wahrscheinlichkeitsverteilung *aller* möglichen Stichproben vom Umfang n. Die meisten von ihnen liefern ein x-quer, das gerade nicht gleich dem μ ist.

2. das Wissen in Statistik zu gering ist, um statistische Ausführungen detaillierter, und nicht nur pauschal und verbal kommentieren zu können.

Skrupel, sich zu blamieren scheinen mir weniger wahrscheinlich der Grund für dieses Verhalten zu sein. Das würde auch nicht dazu passen, dass man sich ganz ungeniert für kompetenter als Statistiker hält (siehe **P1**). Wahrscheinlicher dürfte es wohl sein, dass einem eine typisch statistische oder mathematische Art der Argumentation nicht vertraut ist, zumindest nicht so vertraut wie eine verbale.

Ich habe zu der oben zitierten Auslassung des Gutachters am 16.7. eine ausführliche (vier Seiten) Stellungnahme geschrieben, worin ich mich u.a. gegen die Vorstellungen

a. eine (relativ umfassende) Bestandsaufnahme fehlerhaften Umgangs mit der Statistik und
b. ein Plädoyer für die Nützlichkeit von deskriptiven Studien zu verfassen und dabei "die Nützlichkeit in geeigneter Weise" zu belegen.

des Gutachters wandte und dabei ausgeführt:

> **zu a:** Eine Bestandsaufnahme fehlerhaften Umgangs mit der Statistik in…. ist genau das, was ich nicht vorhabe. Ich kann und will mich nicht als Fachfremder zum Richter über mangelhafte Statistikkenntnisse von Kollegen aus … aufspielen…. Mein Fach ist Statistik und nicht …. und ich finde, es gebietet der Respekt vor Vertretern anderer Fächer, dass man Verständnis dafür hat, dass man dort manche Dinge anderes sieht als bei uns. Hinzu kommt, dass es hier auch Abgrenzungsprobleme gibt: Manches mag auch einfach nur ungeschickt und laienhaft ausgedrückt sein[10] und wo soll man hier die Grenze ziehen? Es liegt mir nicht, hier den Oberlehrer zu spielen und die Schriften anderer nach etwas zu durchforsten, womit man sie dann bloßstellen kann…. Ich kann mir auch gut vorstellen, dass der Herausgeber …. von einem Beitrag nicht begeistert wäre, in dem seine Kollegen reihenweise vorgeführt werden.
>
> **zu b**: Ganz anders als das Gutachten (und auch ganz im Unterschied zu meinem Beitrag), würde ich so etwas, für ein viel zu allgemein gehaltenes, und deshalb auch gerade *nicht* für ein veröffentlichungswürdiges Thema halten. ….Das wirft ja auch die Frage auf: wie beweist man die Nützlichkeit einer empirischen Studie? Ist die Zahl der registrierten Arbeitslosen (Totalerhebung, deskriptive Statistik) nützlicher als die Zahl der Erwerbslosen laut Mikrozensus (eine Stichprobe) … Selbst wenn man so etwas messen könnte, was sollte man daraus für Konsequenzen ziehen?

Ich dachte eigentlich, ich hätte den Gutachter überzeugend genug widerlegt und war deshalb durchaus erstaunt als ich dann am 14. 8. 2014 eine "Finale Stellungnahme" mit der Ablehnung des Manuskripts erhielt. Ich habe dazu am 24. 8 2014 noch einmal einem Brief an den Herausgeber geschrieben aus dem (und aus der finalen Stellungnahme des Herausgebers) ich nachfolgernd zitiere:

Ihre Entscheidung vom 14.8. 2014 hat mich überrascht, zumal ich glaubte, in meiner ausführlichen Stellungnahme vom 16. 7. die Ausführungen Ihres Gutachters überzeugend widerlegt zu haben, wenn dieser meinte (ich zitiere ihn in kursiv), dass mein Text quasi ein alter Hut sei und es für *den statistisch einigermaßen vorgebildeten Leser* völlig klar ist*: Es dürfen eben Signifikanztests nur dann angewendet werden, wenn eine Zufallsstichprobe vorliegt*. Das heißt ja wohl im Umkehrschluss, dass nur jemand, der noch nicht einmal einigermaßen vorgebildet ist, auf die Idee kommen kann, Signifikanztests zu rechtfertigen wenn man keine Zufallsstichprobe hat. Das ist unhaltbar und das, was ich "arrogant" nannte und wogegen ich auch Ihre Kollegen in Schutz genommen habe als ich schrieb.

> Es ist ja nicht nur so, dass in empirischen Arbeiten (und das sind ja inzwischen die meisten) immer wieder auch dann Signifikanztests durchgeführt werden, wenn man gar keine Zufallsstichprobe hat, sondern es wird so etwas auch für ganz legitim gehalten, vor dem Hintergrund eines Begriffs von Repräsentativität, den angeblich kein vernünftiger Mensch mehr vertritt.

… Ich habe lange gezögert, ob ich Ihnen jetzt, nachdem Sie den Stab über mich gebrochen haben überhaupt noch schreiben soll. Aber es haben mich zwei Dinge veranlasst, dies trotzdem zu tun

- Sie schreiben "obwohl ich ein absoluter statistischer Laie bin" (dass man so etwas sagt finde ich gut[11], wo man doch heutzutage als Statistiker damit rechnen muss, dass sich jeder Nichtstatistiker, nur weil er mal etwas gerechnet hat, für den wahren Fachmann in Sachen Statistik hält und meint, einen diesbezüglich belehren zu müssen), und

[10] Das dürfte wohl auch für die Bemerkung mit $P(X=x)=0$ und dem *hinreichend engen Konfidenzintervall* gelten.

[11] Ich finde das ehrt Sie. Ich würde in einer solchen Situation nur noch einen Schritt weiter gehen und einen Statistiker nach seiner Meinung fragen.

- dass Sie, obgleich ich jetzt praktisch für Sie gestorben bin, sich noch die Mühe machten, mehrere Seiten zu schreiben um darin Signifikanztests bei "convenience samples" zu rechtfertigen … ich sehe darin ein Bemühen um die Sache, was ich sehr zu schätzen weiß.

Ich will die folgenden Ausführungen nummerieren um leichter Vor- und Rückverweisungen vornehmen zu können (Für das Thema Paralleluniversum interessieren vielleicht erst die Punkte 3ff oder 6ff)

1. Was den Gutachter betrifft, so hat er erneut auf mich einen sehr negativen Eindruck gemacht, mich allerdings auch in einem Punkt positiv überrascht (was Letzteres betrifft siehe unter 3). Er erkennt noch nicht einmal jetzt, wie blamabel und peinlich der Satz in seinem Gutachten (übrigens der einzige Satz überhaupt, in dem einmal konkret mit Begriffen der Statistik hantieret wurde) war:

   > Der Anwender will einen "treffsicheren" Punktschätzer, während der Statistiker angesichts P(X=x)=0 nur ein hinreichend enges Konfidenzintervall als Gütebeweis anbieten kann.[12]

   In einer E-Mail an Sie habe ich geschrieben, so etwas sei von der Art der Gleichung 5*3 = Donnerstag. Und ausgerechnet so jemand meint, beurteilen zu können, dass alles was ich geschrieben habe ein alter Hut und Kinderkram sei. Es ist interessant zu sehen, dass der Gutachter in seiner neuerlichen Auslassung sein Gutachten sogar für "sehr freundlich und keinesfalls herablassend" hält, dann aber im gleichen Stil weiter macht, wenn er meint, was in meinem Papier steht sei

   > "alles korrekt nachzulesen - wenn nicht im Wörtlichen, so zumindest durch Übertragung des dort zu erwerbenden Wissens. Die Ableitung einer bestimmten analytischen Funktion mag ja auch nicht publiziert sein, aber mit dem Wissen aus einem ordentlichen Mathematik-Buch kann man sich diese Ableitung erschließen."

   Das ist nichts anderes als ein krasses, aber verkapptes Abkanzeln eines Autors; verkappt, weil man wohl vermutet, dass nicht jeder weiß, was eine analytische Funktion ist (klingt ja auch eindrucksvoll). Mit Wikipedia kann man das aber schnell feststellen. Man bekommt dort auch Beispiele für solche Funktionen genannt. Ein Beispiel wäre die Exponentialfunktion. Ihr Gutachter meint also allen Ernstes, meine geistige Leistung sei nur von der Art eines Schülers, der e^{2x} ableiten kann, nachdem man ihm erklärt hat, wie man e^{x} ableitet. Wenn das keine Beleidigung ist! …

   Ist Ihnen auch aufgefallen, dass sich an keiner einzigen Stelle seines Gutachtens Sätze von der folgenden Art finden "die Bemerkung x in Abschn. y ist problematisch, weil…"? So etwas ist aber bei Gutachten, wie ich sie bisher kannte allgemein üblich (und ich habe mich auch selbst in meinen Gutachten immer an diesen Stil gehalten). Es ist aber bezeichnend für das Niveau des Gutachtens und lässt vermuten, dass mein Papier entweder gar nicht sorgfältig gelesen wurde oder gar nicht verstanden wurde. Ich neige nach den Kostproben, die der Gutachter von seinen statistischen Künsten gab…, eher zur zweiten Vermutung. Aber als ob das alles nicht schon schlimm genug wäre, es muss auch noch etwas darauf gesetzt werden. Auf meine Frage, wo denn angeblich alles das steht, was ich geschrieben habe, werden zwei Bücher genannt (natürlich ohne Seitenangaben). Es heißt, das alles stünde auch bei Cochran, Sampling Techniques 1972. Jeder, der einmal ein Lehrbuch der Stichprobentheorie aufgeschlagen hat, wird gesehen haben, dass sich die Stichprobentheorie nur mit random samples beschäftigt; was sonst alles auch noch "sample" genannt wird, wie "convenience samples" wird allenfalls mal am Rande behandelt und dann auch nur verbal, denn es gibt keine statistische Theorie über so etwas.[13]

   Noch schlimmer, man behauptet, alles stünde schon in einem Buch, das wir unseren Erstsemestern in "Statistik I" empfehlen, nämlich Bamberg/Baur/Krapp (Statistik, 17. Aufl., 2012). Ich kenne keine Disziplin, bei der man es zum Univ. Professor bringt, wenn man sich mit seinen Arbeiten nur auf dem Niveau einer Erstsemesterlektüre bewegt. Mehr an Beleidigung geht wohl nicht. Es sind einfach keine guten Manieren, wenn man als Gutachter den Autor eines Aufsatzes so unverblümt

[12] Hier wird wohl – wie ich schon in meiner Stellungnahme schrieb – darauf Bezug genommen dass eine Stichprobenverteilung oft *stetig* ist, was heißt, dass dann Wahrscheinlichkeiten nur für Intervalle definiert sind. Aber schon ein besserer Abiturient weiß, dass "nicht definiert" nicht gleichzusetzen ist mit "null". Der Logarithmus einer negativen Zahl ist nicht definiert, aber deswegen ist nicht ln(-3) = ln(-80,5) = 0.

[13] Ich habe das auch in RCS geschrieben und dort auch dargelegt, dass ich deshalb neue, eigene Wege gehen musste. In den wenigsten Büchern von Statistikern wird überhaupt das Problem angesprochen, das in meinem paper behandelt wird. Das macht ja auch die Behauptung des Gutachtes so abenteuerlich.

für minderbemittelt erklärt und ich denke auch, dass man als Herausgeber so etwas nicht unkommentiert stehen lassen sollte.

2. Falls es vielleicht nicht ganz nachempfunden werden kann, warum ich darin eine Beleidigung sehe mache ich hier ein paar sehr persönliche Bemerkungen.

   Den Text unter 2 habe ich gekürzt, weil er für die inhaltliche Kontroverse weniger relevant ist

   Für den Fall, dass es Ihnen oder Ihrem Gutachter übertrieben erscheint dass ich so massiv verärgert bin …, möchte ich Sie bitten sich einmal vorzustellen, was wohl los wäre, wenn das alles mit umgekehrten Vorzeichen ablaufen würde: Sie schreiben ein Papier und ich – erkennbar Laie auf Ihrem Gebiet – erkläre das alles für trivial und für etwas, was auch jedem BWLer der zweiten und dritten Liga längst bekannt ist, und überhaupt für eine Leistung nur von der Art, wie die Ableitung von e^{2x}. Wir korrespondieren dann über Grundbegriffe Ihres Fachs, ich lasse erkennen, dass ich glaube, was Sie schreiben sei falsch und empfehle Ihnen ein bekanntes einführendes BWL-Lehrbuch. Ich bin mir sicher, dass ich aus diesem Spiel (was zu spielen mir nie einfallen würde) verdientermaßen ziemlich lädiert herauskommen würde.[14]

3. Nun zur Tragweite meiner Ausführungen. Sie schreiben, meine Ausführungen stünden im "Widerspruch zum vorherrschenden Wissenschaftsmodell", womit Sie indirekt zugeben, dass (was auch mir sehr wohl bekannt ist) eine Praxis gängig ist, die Ihr Gutachter allenfalls bei einigen wenigen Leuten vermutet, die noch nicht einmal "einigermaßen vorgebildet" sind. Wenn das ein Beweis dafür sein soll, dass meine Ausführungen falsch sind, ist die dahinterstehende Logik wohl, dass nicht sein kann, was nicht sein darf. Sie folgern weiter: "dann könnte man praktisch alle bisherigen Forschungsergebnisse wegwerfen".

   Ich sehe das nicht so und will das auch unter 4 begründen. Aber selbst wenn man sie wegwerfen müsste, wäre auch das noch kein Beweis, dass meine Ausführungen falsch sind.

   Es ist interessant zu sehen, dass Ihr Gutachter hier das "vorherrschende Wissenschaftsmodell" etwas anders sieht und "Wegwerfen" wohl auch nicht so bedauern würde; wenn er schreibt:

   > Mangels ausreichender finanzieller Ressourcen und/oder Zeit werden dann eben an studentischen Stichproben und sonstigen Convenience-Samples mit fiktiven Marken verbal geäußerte Kaufbereitschaften für nicht existierende Produkte abgefragt. Zugegeben, alles Unsinn und Pseudo-Wissenschaft, aber offenbar für das berufliche Fortkommen der Protagonisten trotzdem förderlich (wenn nicht sogar notwendig), sofern das akzeptierende Journal nur attraktiv genug gerankt ist. Um es noch deutlicher zu formulieren: Es ist wie Parken im Halteverbot - man weiß schon, dass das nicht in Ordnung ist, aber weil man sich Vorteile davon verspricht, macht man es halt bisweilen trotzdem.

   Interessant, dass hier auch der Motor genannt wird, der diese "Pseudo-Wissenschaft" antreibt und am Leben hält, nämlich die Herausgeber von wissenschaftlichen Zeitschriften, also Leute wie Sie.[15]

   Auch wenn es somit mich nicht wirklich betrifft: ich halte die Feststellungen im Zitat für richtig, wenngleich etwas grob und taktlos formuliert (aber das passt ja auch zu den Umgangsformen des Gutachters, wie sie sich mir bisher darstellten). Auch wenn er mir zumindest in statistischen Dingen ziemlich inkompetent zu sein scheint, hier hat er wohl recht.

   Ich stimme aber auch nicht in jedem hier angesprochen Punkt dem Gutachter zu. Er scheint z.B. der Typ des zynischen Jungwissenschaftlers, der sehr wohl weiß, dass seine paper statistisch nicht ganz koscher sind, aber eben mit den Wölfen heult, für verbreiter zu halten als ich dies tun würde. Mir scheint der verbreitere Typ eher der des unbedarften Mitläufers zu sein, der gar nicht richtig begriffen hat, was hinter der statistischen Methode steht, die der Computer rein rechnerisch da für ihn durchgezogen hat. Er steht nicht im Halteverbot weil er es drauf ankommen lässt, sondern weil er das Halteverbotsschild nicht kennt und ein Schild dieses Aussehens vielleicht für eine Werbung hält. Ich glaube auch, dass dieser Typ mit der Zeit immer häufiger werden wird, was mir schon zu denken geben würde, wenn ich Herausgeber einer wissenschaftlichen Zeitschrift wäre.

4. Nun zum "Wegwerfen": Es hat eine Zeit gegeben, in der auch in der BWL noch nicht die Manie ausgebrochen war, dass man meint, um jeden Preis bei Daten, von welcher Qualität auch immer,

---

[14] Für das Thema "Paralleluniversum" ist es nicht nur interessant zu sehen, auf welche Punkte ich in meinen Bemerkungen eine Erwiderung erhielt, und welche Punkte (wie der obige) schlicht ignoriert wurden.

[15] Es ist wohl auch richtig, "dass nicht der Inhalt publizierter Artikel über Wissenschaftler-Karrieren entscheidet, sondern die Journals, in denen sie unterkommen konnten".

den ganzen Apparat der induktiven Statistik auffahren zu müssen um es publizierbar zu machen. Man kann wohl auch nicht sagen, dass seinerzeit die BWL weniger wissenschaftlich war. Man musste damals eben noch dort überzeugende Argumente bringen, wo man heutzutage einfach nur bequem rechnet und "asterisk econometrics" betreibt: * steht für 10%, ** für 5% und *** für 1%.

Es gibt auch heutzutage noch Wissenschaften, in denen man sich zum größten Teil mit der altmodischen Deskriptiven Statistik begnügen muss. Die "Empirische Wirtschaftsforschung" gehört z.B. dazu. Auch wenn man das, was man in Forschungsinstituten und in der amtlichen Statistik betreibt vielleicht nicht so wissenschaftlich finden mag wie " asterisk economics", praktisch relevant ist es allemal und das ist ja auch ein Kriterium, das …(in der BWL) … nicht völlig verpönt ist.[16]

Meine Empfehlung ist ja auch nicht Wegwerfen und keine Studenten mehr befragen, sondern: nur dann Testen, wenn es auch Sinn macht und wenn auch die Voraussetzungen dafür stimmen.

5. Bevor ich zu den in Ihrer "finalen Stellungnahme" angesprochenen Themen komme, noch kurz eine Mutmaßung, wie es vielleicht weiter gehen könnte mit der "Pseudo-Wissenschaft". So wie es eine Zeit vor der Signifikanztesterei, vor dem "cult of statistical significance" (Walter Krämer) in der BWL gab, so könnte es auch eine Zeit danach geben. *Wann* sie kommen wird, ist schwer zu sagen, aber *wenn* sie kommt, dann wohl nicht deshalb weil in Deutschland einigen statistisch versierten BWLern Zweifel am "vorherrschenden Wissenschaftsmodell" kommen …, sondern wohl eher deshalb, weil dieses Modell jenseits des Atlantiks aus der Mode gekommen ist.

   Es ist aber auch denkbar, dass sich das Modell noch sehr lange hält und die Produktion von "Pseudo-Wissenschaft" noch munter zunehmen wird, einfach weil paper mit Signifikanztests bei "convenience samples" einen positiven Nettonutzen haben. Die Kosten sind gering, worauf ja schon das Zitat des Gutachters unter 3 hinweist (nicht umsonst spricht man ja auch von "convenience") und der Nutzen ist groß, so lange " asterisk economics" ein Muss ist beim Publizieren (was selbst ja auch viel mehr als früher ein Muss ist), also kein Sinneswandel bei den editors eintritt.

   Hinzu kommt, dass es mehr und mehr benutzerfreundliche Statistik-Software geben wird und auch immer mehr Statistiknutzer ihr Statistikwissen aus zweiter und dritter Hand beziehen werden. Das gilt nicht nur für Lehrveranstaltungen, sondern auch für Bücher. Ich muss selbst zugeben, dass ich oft Schwierigkeiten habe, Originaltexte von Statistikern zu verstehen. Sich in sie hinein zu vertiefen ist oft wegen der vielen Mathematik sehr mühsam und es setzt nicht nur viel Zeit, die man nicht hat, sondern auch oft Vorkenntnisse, die man auch nicht hat, voraus. Aber es gibt ja auch Alternativen. Über Google findet man schnell einfache Texte (die aber oft mit Vorsicht zu genießen sind)[17] und man kann es niemandem verdenken, wenn er sein Statistikwissen lieber daraus bezieht. Es ist auch kein Wunder, dass Lehrveranstaltungen von Statistik-Anwendern beliebter sind als solche von Statistikern (wenn solche Vorlesungen – und mit ihnen das dazu nötige Personal – nicht ohnehin schon in den Studienordnungen und Stellenplänen weitgehend gestrichen oder reduziert wurden).

   Es spricht also viel dafür, dass in den journals mit immer komplizierteren statistischen Methoden gerechnet werden wird, die immer weniger verstanden werden, also auch immer – ganz im Sinne Ihres Gutachters – mehr "Unsinn und Pseudo-Wissenschaft" produziert wird.

   Wenn ich Herausgeber eines journals wäre, würde mir das schon zu denken geben. Es ist auch mit Händen zu greifen, dass sich die Welt der Statistiker und die der Statistik-Anwender immer mehr auseinander dividieren werden. Wir sprechen mit der Zeit praktisch andere Sprachen und verbinden mit gleichen Worten andere Inhalte. Das sieht man ja auch schon allein an den enormen Verständigungsschwierigkeiten in unserer Korrespondenz.

Die folgenden Punkte 6 – 9 betreffen Ihre Ausführungen zu statistisch methodischen Fragen

[16] Es ist ja bemerkenswert, dass dieser Hinweis auf die Empirische Wirtschaftsforschung und Wirtschaftsforschungsinstitute gar nicht verstanden wurde. Wir gehen darauf später noch (unter P9) ein.

[17] Mit dem Stichwort "Repräsentativität" kommt man z.B. im Internet zu einen Text von Jan Seifert …, in dem allen Ernstes die Begriffe geschichtete Stichprobe und Quotenauswahl als Synonyme behandelt werden und wo es dann auch konsequent heißt "Wenn man es sehr genau nimmt, dürften die üblichen inferenzstatistischen Verfahren bei einer geschichteten Stichprobe gar nicht verwendet werden" (was ja für die Quotenauswahl gilt, aber nicht für die geschichtete Stichprobe; hier wird also einfach etwas behauptet was schlicht falsch ist).

6. Ich will nicht alle Punkte in Ihrer finalen Stellungnahme aufgreifen und mich auch etwas kürzer als bisher fassen, weil mir ja die Erfahrung zeigte, dass ich offenbar wenig Chancen habe (und wohl auch mit mehr Worten nicht mehr Chancen hätte), Sie zu überzeugen.

   Sie gliedern ihre "Überlegungen in den Theorientest und in Vorgehensweisen zur Schätzung von Kenngrößen in einer Grundgesamtheit". Ich habe schon einmal gesagt, dass das Schätzen keine höheren Anforderungen an eine Stichprobe stellt als das Testen. Aus Sicht der Statistik sind Schätzen und Testen zwei Seiten der gleichen Medaille. Beides wird auch unter dem Wort "Schließen" oder "Inferenz" zusammengefasst und wenn man "statistisch" dazu sagt, also "statistische Inferenz" dann ist damit gemeint, dass die Wahrscheinlichkeitsrechnung involviert ist. Wenn Sie schreiben

   > Sie zitieren *Bortz*, der (dies?) ausdrücklich kritisiert und gezogene Schlüsse als möglicherweise wertlos erachtet. Damit kann ich nichts anfangen (*welche* Schlüsse?).

   dann ist zu bedenken, dass damit genau diese *auf der Wahrscheinlichkeitsrechnung basierten* Schlüsse gemeint sind und dass diese von anderen "Schlüssen", z.B. logischen Folgerungen im Rahmen einer verbalen Argumentation zu unterscheiden sind. Wenn Sie schreiben

   > Das Kriterium für eine Stichprobe ist in diesem Anwendungskontext weder, dass sie „zufällig gezogen worden ist“ noch dass sie „repräsentativ“ (in jedweder Variante der Bedeutung dieses Worts) ist, sondern einzig, dass sie innerhalb der Reichweite der Theorie liegt.

   so ist dazu zu sagen, dass es Ihnen zwar unbenommen bleibt, was Sie alles als "Stichprobe" bezeichnen wollen und was nicht, dass aber "in der Reichweite einer Theorie" sein und "zufällig gezogen" sein zwei ganz verschiedene Kategorien sind. Es gibt hier auch keinen trade off, mehr von dem gleicht weniger von jenem aus. Für die Anwendbarkeit der Wahrscheinlichkeitsrechnung – und nur darum ging es ja in meinem Papier – ist allein "zufällig gezogen" relevant.

P4. Im statistischen Paralleluniversum der BWL werden exakt definierte Begriffe, wie "Zufallsauswahl" durch wenig operationale Begriffe, wie "Reichweite einer Theorie" ersetzt und auf die gleiche Stufe gestellt (ohne zu fragen, wie man z.B. die "Reichweite" messen könnte oder gar wie formelmäßig Beziehungen zwischen hier auf die gleiche Ebene gestellte Konzepte herzustellen wären). Es wird gleich auch noch gezeigt, dass "Zufallsauswahl" mit Randomisierung verwechselt wird.

   Dass eine statistische Methode Anwendbarkeitsvoraussetzungen hat, ist gar nicht so selten. Das ist nicht nur beim Schätzen und Testen so, sondern auch bei so einfachen Sachen, wie z.B. die Berechnung eines arithmetischen Mittels. Hat A die Steuerklasse III und B Steuerklasse V, so macht es keinen Sinn, zu sagen, dass sie im Durchschnitt Steuerklasse IV haben. Hier setzt eine Methode (Berechnung eines arithmetischen Mittels) einen Skalentyp (metrische Skala) voraus, der im Fall der Steuerklasse nicht gegeben ist. Deshalb ist die Rechnung zwar rechnerisch richtig, sie verdient es aber trotzdem weggeworfen zu werden. Ganz in diesem Sinne setzt die Anwendbarkeit der "statistischen Inferenz" auch etwas voraus, nämlich das Vorhandensein einer Zufallsauswahl.

7. Aber wir haben ja schon einen Dissens bei der Frage, <u>was überhaupt eine Zufallsauswahl ist</u>. In ihren Ausführungen kommt immer wieder der Gedanke auf, dass eine nichtzufällige Auswahl quasi dann ihren nichtzufälligen Charakter verliert und zu einer Zufallsauswahl wird, wenn die ausgewählten Personen anschließend nach dem Zufallsprinzip in Gruppen aufgeteilt werden.

   > Ich verstehe nicht einmal, was bspw. daran falsch daran ist, „die zufällig in einem Seminar anwesenden Teilnehmer“ *nach Zufall* in Experimentalbedingungen einzuteilen, diese pro Bedingung einem Treatment zu unterziehen und dann den Effekt des Treatment z.B. so zu testen, dass man die Wirkungen über die Experimentbedingungen vergleicht, um *einen Beleg* für die Hypothese (= durch diese Daten *gestützt/nicht gestützt*) zur Wirkung des Treatments zu erhalten.

   Daran ist nichts falsch, *wenn* der Ausgangspunkt dieser sog. *Randomisierung* – d.h. Zuordnung (assignment) von Einheiten (z.B. Personen) zu Gruppen nach dem Zufallsprinzip – die Grundgesamtheit oder eine *Zufalls*-Stichprobe wäre. Falsch ist nur, zu meinen, dies würde etwas daran ändern, dass die Personen, die hier z.B. auf Experiment- und Kontrollgruppe aufgeteilt werden, selbst keine Zufallsauswahl darstellen. Was hier verwechselt wird ist *Randomisierung*, eine *restlose Zuordnung aller* n Einheiten zu Gruppen nach dem Zufallsprinzip und eine Zufallsauswahl also eine

Auswahl von n aus N > n Personen nach dem Zufallsprinzip.[18] Nehmen wir das Beispiel von *Müller/Voigt/Erichson* (2010) auf das Sie sich ja auch im Folgenden beziehen

> Diese Autoren haben ein convenience sample (ca. 500 Magdeburger Studenten) und teilen dieses – *hoffentlich nach Zufall* – in vier Experimentalbedingungen auf. Die Messwerte sind *Indikatoren* für Zahlungsbereitschaften. Die Autoren testen Hypothesen, die sie – hoffentlich ordentlich – aus einer Theorie ableiten. Ich vermag daran *per se* methodisch nicht Falsches zu erkennen.

Nehmen wir an, die n = 500 Studenten werden in vier Gruppen zu jeweils 125 Studenten aufgeteilt. Wenn die n Studenten nicht zufällig ausgewählt waren, und das ist ja unbestritten, dann bleiben sie auch nach dieser Operation *nicht*-zufällig ausgewählt, und das gilt auch jeweils für die 125 Studenten in jeder Gruppe, und zwar egal ob sie zufällig oder nicht zufällig in vier Gruppen a 125 Studenten aufgeteilt wurden; denn keine der 500 Personen ist aus der Grundgesamtheit zufällig ausgewählt worden. Es werden ja bei diesem Vorgang der Randomisierung weder Personen aus dem Kreis der 500 entfernt noch neue, andere Personen aus der Grundgesamtheit hinzugefügt.

Sie spielen den Zufall bei der Auswahl gegen den Zufall bei der randomization aus, indem Sie schreiben

> Der Zufall kommt erst dann zwingend ins Spiel, wenn diese Stichprobe auf Versuchsbedingungen aufgeteilt wird, aber das thematisieren Sie nicht.

und Sie bedenken nicht, dass der Zufall in beiden Fällen eine ganz andere Rolle spielt: bei der Auswahl bringt er Stichprobe und Grundgesamtheit in die Beziehung zueinander, die dem alles bestimmenden Modell einer Ziehung von Kugeln aus einer Urne in der Wahrscheinlichkeitsrechnung entspricht (was ja auch schon voraussetzt, dass Grundgesamtheit [Urne] und Stichprobe [gezogene Kugeln] klar definierte verschiedene Dinge sind)[19] und bei der Randomisierung dient der Zufall dazu, evtl. sehr subtile systematische Unterschiede zwischen Teilgesamtheiten, die verschiedene Treatments erfahren, zu vermeiden.[20]

Wie sich später (S. 13f) zeigt wurde dies alles gar nicht zur Kenntnis genommen oder nicht verstanden.

8. Danach sollte klar sein, wie wichtig nicht nur der Zufall, sondern auch die eindeutige Definition der Grundgesamtheit bei der statistischen Inferenz ist. Und genau das (was ist hier überhaupt die Grundgesamtheit?) ist bei einem convenience sample der nächste kritische Punkt. Einen Augenblick lang dachte ich, Sie könnten vielleicht recht haben mit Ihrer Überlegung unter 7, *wenn* man bei Müller/Voigt/Erichson die n = 500 Studenten als die Grundgesamtheit und nicht als Stichprobe auffassen würde. Aber das ist nach dem zur Zufallsauswahl und Randomisierung Gesagten schon formal abwegig, es würde Ihnen auch inhaltlich nichts nützen. Denn Sie wollen ja keine Aussagen über die Magdeburger Studenten machen, sondern *Theorien* prüfen, also Aussagen treffen, die für etliche Millionen Menschen gelten sollen.[21]

---

[18] Bei einer Auswahl gibt es immer Einheiten, die nicht ausgewählt wurden (i.d.R. sind das mehr als die Ausgewählten). Bei der Zuordnung bleibt aber niemand übrig der nicht ausgewählt wurde. *Jede* Person wird einer und nur einer Gruppe zugeordnet.

[19] Schon das haben wohl diejenigen übersehen, die mit Daten statistische Tests durchführen, die eigentlich die Grundgesamtheit darstellen. Ohne die oben genannte Relation gibt es auch keine Formeln. Bei jedem Test geht es darum ob eine Stichprobe zufällig oder überzufällig von der hypothetisch angenommenen Grundgesamtheit abweicht. So etwas setzt voraus, dass klar ist, was hier Stichprobe und was Grundgesamtheit ist

[20] Es ist nicht ganz verständlich, warum sie die Zufälligkeit bei dem assignment so sehr betonen ("*hoffentlich nach Zufall*") und zum Dreh- und Angelpunkt machen, wenn die Gesamtheit, die in Gruppen aufgeteilt wird einigermaßen homogen ist (was bei Studenten einer Vorlesung ja vielleicht der Fall sein könnte. was übrigens auch nicht sehr für "Repräsentativität" spricht). Es ist ja nicht so zwingend, dass sich Studenten, wenn sie nicht nach Zufall, sondern nach dem Alphabet (Namen) aufgeteilt werden massiv hinsichtlich Produktkenntnis usw. unterscheiden (Studenten von A bis M kennen das Produkt, die von N bis Z kennen es nicht). Der Dreh- und Angelpunkt ist doch wohl eher, was für eine Art von Auswahl die n = 500 Studenten darstellen.

[21] Über ganze *Theorien* durch Befragung von Hörern einer Vorlesung oder von Besuchern eines Einkaufszentrums entscheiden zu wollen (nach dem Motto *** = die "H_0 Theorie" ist falsch) scheint mir ein ziemlich kühnes Unterfangen zu sein. Es wäre auch schon mit echten Stichproben etwas kühn. Ich wusste gar nicht dass das in Ihrem Fach *so* verbreitet ist. Es ist auch weit entfernt vom Testen in der Statistik, wo man ja immer noch *per Vollerhebung* eindeutig und *ohne jeden Zweifel* sehen *könnte*, ob $\mu = \mu_0 = 35$ stimmt oder nicht. Man könnte dann sogar viel mehr als nur $\mu \neq \mu_0$ feststellen, z.B. dass $\mu = 36{,}7$ ist. Aber könnte man jemals (auch

Sie schreiben als Erwiderung zu meiner These, dass man sich beim convenience sample nachträglich die passende Grundgesamtheit zurecht strickt

> Der Punkt, dass es „mehr oder minder unklar bleibt, was genau die Grundgesamtheit ist, für die die Stichprobe repräsentativ sein soll“, stimmt … so nicht.

Wenn das nicht stimmt, was ich geschrieben habe, es also klar ist, was bei einem convenience sample die Grundgesamtheit ist, dann können Sie mir wohl auch sagen, wie groß die Grundgesamtheit im Fall der Magdeburger Studenten ist. Umfasst sie N = 500.000 Personen oder ist sie vielleicht sehr viel größer, weil die Theorie ja nicht nur für jüngere Konsumenten in Sachsen Anhalt gelten soll? Ist N vielleicht 280 Millionen? Gilt die Theorie für Europa oder auch für ganz Asien? Und wer gehört zur Grundgesamtheit? Ziemlich sicher wohl ein Student der Uni Marburg, aber gehört auch die Hausfrau aus Kasachstan dazu, zu der es vielleicht kein passendes Pendant unter den Magdeburger Studentinnen gibt?

Natürlich habe ich auf diese Fragen keine Antwort bekommen. Stattdessen wurden die 96.322 Praxen im Bundesarztregister, von denen gleich die Rede ist, problematisiert.

P5. Im Paralleluniversum der BWL spielt eine ganz anders (und ohne Bezug auf das Urnenmodell) definierte Grundgesamtheit (GG) eine Rolle, die zugleich wenig klar ist. Man ist noch nicht einmal bereit, zu sagen, was in einem konkreten Fall der Umfang N der GG ist und wie man überhaupt zu einer konkreten Zahl für N gelangt. Man überlässt stattdessen die Definition der GG einer nachträglichen (nach Abschluss der Befragung und ihrer Auswertung) Diskussion über eine "Theorie". Die GG *ist* das was (mehrheitlich) zur GG *erklärt* wird.

Um zu zeigen, wie grundlegend anders die Situation bei einer echten Stichprobe ist bringe ich das Beispiel des Panels von Arztpraxen der Kassenärztlichen Bundesvereinigung (KBV), das mir gut vertraut ist, weil ich dort als Mitglied eines wiss. Beirats und Berater in Stichprobenangelegenheiten involviert war. Dort wurde aus den 96.322 Praxen im Bundesarztregister (BAR) 8.717 zufällig ausgewählt. Es ist klar, dass hier N = 96.322 ist und es ist auch klar, ob eine konkrete Praxis in die Grundgesamtheit fällt oder nicht. Die Praxis von Dr. A, in der nur Privatpatienten behandelt werden können, oder die von Dr. B in Frankreich gehören z.B. ganz eindeutig nicht dazu, weil sie nicht zur kassenärztlichen Versorgung in der Bundesrepublik zugelassen sind und deswegen auch nicht im BAR erfasst sind.

Wenn man tatsächlich aus einer Urne (hier dem BAR) gezogen hat sind solche Fragen völlig klar. Aber sind solche Abgrenzungsfragen auch nur annähernd so klar und eindeutig, wenn man gar nicht aus einer Urne gezogen hat? Sie können ja noch nicht einmal sagen, wie groß N ist. Sie können auch dann noch nicht einmal exakt N angeben wenn, wie in der folgenden Aussage, die Grundgesamtheit das ist, was man zur Grundgesamtheit erklärt:

> Es ist der Job des Forschers, im Theorie-Abschnitt des Journal-Aufsatzes und/oder im Abschnitt Limitations zu präzisieren, wie er/sie die Reichweite dieser Theorie sieht. Alle Personen oder andere Untersuchungseinheiten, für die er/sie Gültigkeit postuliert, *sind* die Grundgesamtheit.

So etwas hat nichts mehr mit der Situation zu tun, von der man ausging bei der Herleitung der Formeln für die statistische Inferenz, also von Formeln, die anzuwenden Sie ja nach wie vor auch bei einem convenience sample für legitim halten. Diese Formeln sind hergeleitet worden mit der Modellvorstellung, dass aus einer gegebenen Urne Kugeln gezogen wurden und zwar so, dass *jede* Kugel eine bekannte Wahrscheinlichkeit (die nicht 0 oder 1 ist) hat, gezogen zu werden. Sie passen nicht zu einer Situation, wo man keine Kugeln zieht, sondern Kugeln vorfindet und wo man sich erst noch eine zu den Kugeln passende Urne hinzudenken muss.[22]

9. An einigen Stellen fühle ich mich, um dies abschließend zu sagen, auch falsch verstanden. Ich habe z.B. nie in Zweifel gezogen, dass eine Klumpenstichprobe eine Stichprobe im Sinn der Stichpro-

---

wenn es hier so etwas wie ein Vollerhebung gäbe) so sicher und exakt über eine "Theorie" urteilen, wo es doch bei einer Theorie um viel mehr als nur um eine konkrete Zahl (wie etwa 35 oder 36) geht?

[22] Es gibt keine mathematische Theorie über eine Auswahl aus einer Gesamtheit, die ihre Existenz erst der "Erklärung" eines Forschers verdankt. Ich glaube auch, dass es eine solche Theorie nie geben kann. Es kann auch keine Geometrie geben, wenn man es offen lässt, was Forscher zu einer Geraden erklären und was nicht.

bentheorie ist (also eine Zufallsstichprobe). Es gibt auch Stellen in Ihrer finalen Stellungnahme, die ich einfach nicht verstehe, wie z.B. diese

> Wenn Sie sogar sagen wollen, dass „Strukturähnlichkeit“ der Stichprobe mit der Grundgesamtheit hinsichtlich der Merkmale, die stark mit dem interessierenden Merkmal korrelieren, und Qualität der Induktion bei interessierenden Merkmalen nicht zusammenhängen, hätte ich etwas gelernt, auch wenn ich es – intuitiv – nicht verstehe.

Ich weiß nicht, worauf Sie hier Bezug nehmen, aber vielleicht können wir uns ja auch darauf einigen, dass wir in Sachen Statistik offenbar verschiedene Sprachen sprechen.[23]

Abschließend möchte ich noch sagen, dass ich mir schon vorstellen kann, dass alles dies nicht angenehm sein mag, wenn man es mit der Praxis empirischer Arbeiten in manchen Disziplinen (nicht nur in der BWL) vergleicht. Ich kann auch verstehen, dass man, wenn hier die Dinge, nicht zusammenpassen, gefühlsmäßig dazu neigt, als Übeltäter eher an den Statistiker als den Statistikanwender zu denken. Was ich aber nicht verstehen kann ist, dass man nicht sehen will, *dass* die Dinge nicht zusammenpassen und dann auch nicht darüber reden will, *warum* sie nicht zusammenpassen.

Ich habe mir mit dieser Stellungnahme, wie schon mit der vom 17.7. sehr viel Mühe gegeben und ich hoffe, dass es nicht wieder ganz umsonst war.

Ich bekam vom Herausgeber eine Antwort (1.9.2014), mit der ich eigentlich gar nicht mehr gerechnet hatte. Sie hat mir endgültig gezeigt, dass es überhaupt keinen Sinn macht, sich als Statistikprofessor mit BWL-Professoren über Statistik zu unterhalten. Andererseits sind die folgenden Ausführungen durchaus wichtig, weil sie zeigen, warum eine Kommunikation gar nicht mehr möglich ist und wie dafür einige "tiefer liegende" Divergenzen zwischen den Universen (siehe P6 bis P10) verantwortlich sind.

> … Leider hat sich das Thema von der Verwendung von Begriffen wie Repräsentativität auf ein grundsätzliches verlagert, nämlich wie empirische Wissenschaft betrieben wird.
> Vielleicht nehmen Sie meine nochmaligen Anmerkungen nur als ein Zeichen meiner Wertschätzung Ihrer Person, denn manche Ihrer Passagen klingen etwas gekränkt.[24] Es war keinesfalls meine Absicht, Sie zu kränken. Ich wollte nur anmerken, wie ich die Themen sehe, über die Sie sehr grundsätzlich schreiben. Sie dürfen sich dann bitte auch nicht darüber ärgern, wenn ich sehr grundsätzlich antworte …

Es ist klar, dass ich mich nicht darüber, sondern über die Beleidigungen des Gutachters geärgert habe.

Schätzen und Testen sind zwei Seiten einer Medaille?

Es ist völlig unbestritten, dass aus einer aktuellen Grundgesamtheit von 96.322 Arztpraxen nach dem Urnenmodell eine zufällige Stichprobe gezogen und statische Kenngrößen bestimmt werden können, die diese Grundgesamtheit beschreiben. Für mich ist dies die Welt der Konfidenzintervalle, und es wird Zielgruppen geben, für die solche Kenngrößen Information enthalten. Genauso gibt es aber auch Zielgruppen, die Theorien entwickeln und diese testen und daraus Erkenntnisse gewinnen. Ich weiß nicht, ob es irgendeine Theorie gibt, die als Reichweite *exakt* diese 96.322 Arztpraxen hat, traue mich aber einfach einmal zu behaupten, dass es solche Theorien nicht gibt. Insofern wird man an den n = 8.717 auch nichts testen. Wenn jemand die zufällige Stichprobe in überflüssiger Weise auch noch als repräsentativ bezeichnet, ändert dies den Ergebnissen nichts.

---

[23] Ich bin mir gar nicht sicher, ob ich so etwas überhaupt gesagt habe. Was ich vielleicht gesagt habe ist, dass mir keine Formeln bekannt sind, mit denen man vom Ausmaß (!) der Strukturähnlichkeit auf die Genauigkeit und Sicherheit einer Schätzung von Parametern der Grundgesamtheit schließen kann. Das ist ja auch ein Argument gegen die Betonung der Strukturähnlichkeit. Es gibt ja noch nicht einmal ein Maß dafür, um wie viel eine Auswahl strukturähnlicher ist als eine andere, oder um wie viel besser eine Schätzung wird, wenn bei gleicher (!) Strukturähnlichkeit der Umfang n der nicht-zufälligen "Stichprobe" größer gemacht wird, also z.B. 2000 statt 500 Studenten befragt werden. Die entsprechenden Formeln für das mindestens erforderliche n gelten auch hier wieder nur bei einer *Zufalls*auswahl aus einer klar definierten Grundgesamtheit.

[24] Es wäre vielleicht schön gewesen, wenn man sich zumindest über Benimm-Regeln von Gutachtern einigen könnte, wenn man sich schon nicht über Grundbegriffe der Statistik verständigen kann. Das Verhalten des Gutachters, ist weit mehr als etwas, was einen vielleicht "etwas kränkt".

Ich möchte die Person sehen, die eine Methode hat, mit der man feststellen kann, dass die Reichweite einer Ärztetheorie exakt N = 97.537 oder N = 183.389 beträgt.

Die akademische Forschung, die ich kenne, beschäftigt sich mit der Entwicklung von Theorien und dem Theorientest (sprachlich genauer: der Prüfung von Teilaussagen der Theorie, die als *Hypothesen* bezeichnet werden). Eine *Theorie*, aus der man eine Hypothese in Form einer statistischer Vermutung zu einem Linearitätsausmaß wie z.B. $\rho = 0.7$ oder $\rho > 0.7$ ableiten könnte, ist mir niemals begegnet (und ich kann mir eine solche Theorie und daraus abgeleitet eine derartige Hypothese nicht vorstellen), gleichwohl jemand, der sich – warum auch immer – für den Grad an Linearität eines Zusammenhangs interessiert, auch einem Konfidenzintervall für ρ Erkenntnis abgewinnen kann, wenn er/sie eine Zufallsstichprobe zur Verfügung hat.

Aus Sicht der Formeln mögen Schätzen und Testen zwei Seiten einer Medaille sein, aber die Begriffe werden doch in den empirischen Wissenschaften in aller Regel mit systematisch unterschiedlichen Absichten verbunden.

Es ist interessant, dass etwas zwar aus "Sicht der Formeln" richtig sein kann, aber trotzdem falsch sein soll (im statistischen Universum der Statistiker ist mir so etwas noch nicht begegnet).

P6. Es wird hartnäckig eine Unterscheidung gemacht zwischen der "Welt der Konfidenzintervalle" und dem (die empirische Arbeit dominierenden) Interesse von Forschern, "die Theorien entwickeln und diese testen". Dabei werden Hypothesen, die sich auf eine Zahl reduzieren (H_0: ρ = 0,7) – nur um solche geht es in der Statistik –als zu eng abgelehnt. Man kann sich offenbar auch gar nicht vorstellen, warum es Sinn macht, eine Zufallsstichprobe von Arztpraxen zu ziehen, obwohl man gar nicht vor hat, eine Ärztetheorie "testen" zu wollen, die auch für Ärzte zur Zeit von Ramses II gelten soll.* Man verspricht sich von statistischen Tests auch Aussagen (über "wahr", bzw. "falsch" einer "Theorie"), die diese Tests gar nicht liefern können. Mehr zu diesem Punkt (das Missverständnis, ein Test zeige, welche Hypothese richtig ***ist***, während er nur zeigt, wie *wahrscheinlich* etwas ist, *wenn* eine Hypothese richtig ***wäre*****) siehe auch unter P9.

* Damit hängt wohl auch zusammen, dass meine Hinweise auf deskriptive statistische Arbeiten im Rahmen der amtlichen Statistik oder der empirischer Wirtschaftsforschung (oben S. 7 dort Punkt 4) offenbar überhaupt gar nicht verstanden worden sind.

** Dieser Punkt spricht auch eine alte Kontroverse an zwischen R. A. Fisher (p-value als "Bestätigungsgrad" einer Hypothese –ohne dass eine konkrete Alternativhypothese im Spiel ist) und der jetzt herrschenden Theorie von Pearson und Neyman (Test als Entscheidung zwischen konkurrierenden Hypothesen).

Theorientest

Was Sie ganz fundamental angreifen, ist die Vorgehensweise in der akademischen empirischen Forschung, die sich mit Theorien und ihrem Test beschäftigt. Akademische empirische Forschung ist zum überwiegenden Teil der Test von Theorien und nicht die Hochrechnung aus Stichproben (ich setze einfach einmal die Zahl 99% an, ohne sie jetzt belegen zu können). Im Kern sagen Sie, man darf nur testen, wenn man eine Zufallsstichprobe hat, anderenfalls muss man sich mit Deskription begnügen. Was ich – gerne als Laie – einwende, ist, dass Sie sich – als Experte auf dem Gebiet – im Widerspruch zum gängigen Wissenschaftsmodell (das seit etwa 1930 praktiziert wird – ich bin so mutig, diese Jahreszahl zu verwenden) befinden und dass Sie, wenn Sie so vehement kritisieren, gute Gegenentwürfe haben müssen.

P7. Die wiederholte Berufung auf das "gängige Wissenschaftsmodell" läuft auf einen "Beweis" auf Basis des Grundsatzes "Was alle machen kann nicht falsch sein" hinaus. Dieses Prinzip ist beliebt, wenn man keine Gründe für richtig und falsch angeben kann. Wenn man das aber kann, braucht man keinen "Gegenentwurf": man muss nur alle auffordern, es richtig zu machen.

(1) Ich kenne keine Theorie, für nur Gültigkeit speziell für den Zeitpunkt der Datenerhebung beanspruchen würde. Die Grundgesamtheit von Stichproben zur Prüfung von Theorien sind die Objekte/Personen/Einheiten in der durch die Theorie bestimmten Reichweite, also auch in Vergangenheit und Zukunft. Insofern ist es *prinzipiell nicht möglich*, für einen Theorientest eine Zufallsstichprobe aus dieser Grundgesamtheit zu verwenden.

(2) Ich kenne nicht *den* Test einer Hypothese. Die gängige Wissenschaftspraxis ist es, dass es einzelnen Forschern gelingt, eine Theorie in einem Top-Journal zu platzieren. Diese Autoren erheben in einigen Convenience Samples und beantworten die Frage nach der Gültigkeit der Theorie bzw. der daraus abgeleiteten Hypothesen mit dem Instrumentarium der statistischen Tests (nach dem Urnenmodell). Das mag einerseits eine Veröffentlichung sein, die weiter keine Resonanz findet. Wenn sie andererseits aber der Scientific Community interessant erscheint, werden viele weitere Forscher an anderen aus dem Universum der Reichweite der Theorie ausgewählten Convenience Samples[25] eine Replikation versuchen, um weitere Belege zu sammeln. Insofern entspricht es nicht der Realität der wissenschaftlichen Praxis, Wissen auf Grundlage eines Tests einer Hypothese zu gewinnen, sondern es entsteht zunehmendes Ver*trauen in die Gültigkeit der Hypothese*, wenn sie sich vielfach bewährte (wobei einige Negativ-Befunde eher die Forschung eher aktivieren).

P8. Es ist klar, dass man nicht eine Stichprobe ziehen kann aus einer Grundgesamtheit (GG), die es noch gar nicht gibt, sondern die es erst in der (näheren? weiteren?) Zukunft geben könnte oder aus einer GG, die es gar nicht mehr gibt. So etwas wird aber gefordert, wenn es um die auch Vergangenheit und Zukunft umfassende "Reichweite" geht. Es fragt es sich dann natürlich, wie mit convenience samples etwas möglich sein kann, was mit Zufallsstichproben nicht möglich ist. Man kann auch mit convenience samples nicht Ungeborene befragen, meint aber darüber entscheiden zu können, welche Theorie später für sie gelten wird.

Meine These (und die von Bortz und vielen anderen) war, dass man **nicht** statistische Tests durchführen sollte, wenn man nur ein convenience sample hat. Hier kommt jetzt aber die These, dass man nie Zufallsstichproben haben kann, und deshalb statistische Tests **gerade dann** durchführen muss, wenn man ein convenience sample hat. Hier wird nicht nur um jeden Preis etwas verteidigt, was nach Meinung des Gutachters keinem vernünftigen Menschen in den Sinn kommt, sondern es wird dabei auch mit Gründen operiert, die ins Absurde reichen.

(3) Ich kenne keine Akzeptanz von empirischer Forschung in Journals zu Theorien, wenn das Ergebnis des Tests *zur Gültigkeit* einer Hypothese nicht binär ist (gestützt/nicht gestützt). Hier hilft die von Ihnen als Alternative vorgeschlagene Deskription durch Mittelwerte etc., mit der sich diese Forschung begnügen soll, nicht weiter. Die akademische Praxis verwendet ein binäres „Significance is all that matters"-Kriterium. Die Sternchendiskussion ist hier nicht weiterführend, die könnte man genauso gut für Konfidenzintervalle (siehe oben) führen.

(4) Ich erkenne aus meiner Erfahrung nicht, wieso die Thematik (empirische Forschung kann aller Regel keine Zufallsstichproben nutzen, will aber dichotome Resultate für die postulierten Theorien haben) etwas im Speziellen mit …, BWL, VWL oder empirischer Wirtschaftsforschung zu tun hat.

P9. Das ausschließliche Interesse an dichotomen (binären) Resultaten im Sinne von 👍 oder 👎 (für den Fall, dass man *** nicht mag) wertet nicht nur große Teile der Statistik ab (insbes. die Deskriptive Statistik), es verkennt auch, was mit einem statistischen Test eigentlich "gewonnen" wird: Signifikanz heißt doch nicht H_0 – oder gar eine "Theorie" – ist falsch und H_1 ist richtig ("gilt"), sondern nur, dass die *Wahrscheinlichkeit* der konkreten Stichprobe dafür spricht, H_0 für falsch zu *halten* (zu verwerfen) wenn H_0 gelten *würde*.* Das setzt aber voraus, dass man überhaupt von Wahrscheinlichkeiten und einer GG sprechen kann (ohne GG kann es auch keine Hypothesen über die GG geben). Wie man das alles im Falle von einem convenience sample können will, bleibt rätselhaft. Es ist auch rätselhaft, wie man zeigen kann, ob (oder gar *in welchem Maße*) ein convenience sample eine "Replikation" eines anderen convenience samples ist.

* Die Unterscheidung zwischen **richtig sein** und **für richtig halten**, fällt dann flach, wenn die Daten keine Stichprobe sind, sondern praktisch die komplette GG umfassen. Das zeigt ja gerade, warum (wie in RCS gezeigt) statistische Test in dieser Situation so unsinnig sind (obgleich solche Rechnerei durchaus verbreitet ist).

Wenn z.B. ein Forscher im agrarischen Bereich …. durch entsprechende Theorien herleiten kann, dass ein Pflanzenschutzmittel X aufgrund der implementierten Eigenschaften gegen einen bestimmten Pilz auf Weizen wirksamer ist als das Mittel Y, würde dieser Forscher Versuchsfelder mit Weizen bepflanzen und bei einem Teil X und bei einem anderen Y anwenden (convenience sample). Eine Zufallsstichprobe in dem Sinn, aus der Menge aller vergangenen, gegenwärtigen und zukünftigen Weizenhalme irgendwo auf dieser Erde bei Beachtung aller denkbaren Weizensorten eine Stichpro-

[25] Selbst wenn es ein klar definiertes "Universum der Reichweite einer Theorie" gäbe, ist es mehr als fraglich, ob man bei Convenience Samples überhaupt von "ausgewählt" sprechen kann.

be gemäß dem Urnenmodell zu ziehen und dann im zweiten Schritt daraus zufällig Weizenhalme auf zwei Gruppen (X versus Y) aufzuteilen, ist nicht vorstellbar. …. Das gleiche Beispiel lässt sich leicht auch für medizinische Behandlungsmethoden einer Krankheit konstruieren. Man wird keine Zufallsstichprobe aller in der Vergangenheit, Gegenwart und Zukunft an allen Orten dieser Erde Erkrankten für alle Varianten dieser Krankheit gewinnen können. Und man will wissen, *ob* man treatment X oder Y verwenden soll.

Es lohnt sich nicht, dies zu kommentieren, weil es keine zutreffende Beschreibung der entsprechenden Methoden bei pharmazeutischen oder biologischen Versuchen ist, sondern eine Aufgabenstellung formuliert, die solche Versuche praktisch unmöglich machen würde.

(5) Ich hatte angemerkt, dass sich die Praxis der empirischen Forschung, zumindest soweit ich sie überblicke, bspw. ersatzweise mit einer Randomisierung der Zuordnung der Testeinheiten auf die Versuchsbedingungen behilft, um wenigstens einen Zufallsprozess zu haben, der die Anwendung statistischer Tests zur Prüfung von Theorien mit binärem Erkenntnis ermöglicht. Vielleicht bräuchte man jedoch eine andere „Statistik“.

Bitte verstehen Sie, dass ich zwar vielseitige und sicherlich gerechtfertigte Kritik an der Praxis des akademischen Betriebs in Ihrem Manuskript finde, aber keine klare Vision, wie eine Statistik jenseits des Urnenmodells aussehen könnte, die sich besser für die Praxis eignet, wie ich sie in (1) bis (5) beschrieben habe. Labels wie Unsinn und Pseudowissenschaft helfen mir als Schriftleiter dieser Zeitschrift auch nicht weiter.

P10. Hier wird erneut Zufallsauswahl und Randomisierung gegeneinander ausgespielt. Man hält die Wahrscheinlichkeitsrechnung (genauer die statistische Inferenz) schon allein dann für anwendbar, wenn bei der empirischen Arbeit *irgendwo überhaupt* etwas wie ein "Zufallsprozess" auftaucht. Man fragt aber andererseits nach einer "anderen Statistik". bzw. einer "Statistik jenseits des Urnenmodells" und sieht nicht, dass

- es eine *Statistik* "jenseits des Urnenmodells" bereits gibt (sie heißt "Deskriptive Statistik"), die man aber nicht mag, weil sie nicht 👍 und 👎 liefert, sondern Zahlen und Grafiken über die man sich noch Gedanken machen muss (die man noch "interpretieren" muss),
- es eine Wahrscheinlichkeitsrechnung "jenseits des Urnenmodells" nicht gibt,
- man sich aber jenseits des Urnenmodells befindet, wenn man *irgendjemand* befragt, nur weil er bzw. sie gerade verfügbar ist.

Labels wie "Unsinn" und "Pseudowissenschaft" hat der Gutachter ins Spiel gebracht, nicht ich (siehe S. 6 oben) Ich gebe gerne zu, dass ich keine Vision geliefert habe. Mir scheint es auch wichtiger zu sein, erst einmal die bestehende Statistik diesseits des Urnenmodells richtig zu verstehen als Visionen zu liefern.

Ich habe auf diesen Brief nicht mehr geantwortet, nicht nur deshalb, weil eine Antwort ausdrücklich unerwünscht war, sondern auch weil mir vieles, was hier zu lesen ist über Theorien-, bzw. Hypothesentests, Grundgesamtheit und Stichprobe doch reichlich bizarr vorkam, so dass ich auch gar nicht wusste, was ich dazu sagen sollte. Hinzu kommt, dass wir uns schon genug im Kreise gedreht haben.

Das Tückische an einem Paralleluniversum, ist ja, dass man nach einiger Zeit nicht mehr weiß, wo oben und unten ist. Das betrifft

- nicht nur Aussagen über Statistik (man kann keine Punktschätzung machen, weil die Wahrscheinlichkeit $P(X=x)= 0$ ist und man kann deshalb leider nur ein Konfidenzintervall berechnen; oder: die GG ist bei einer gegebenen Stichprobe unterschiedlich groß, je nachdem, ob es um Schätzen oder Testen geht),
- sondern auch, dass man Personen gar nicht mehr wiedererkennt: ich war nie der Einzelkämpfer für eine abstruse Sache, der Ketzer oder Geisterfahrer, zu dem ich offenbar im Laufe der Zeit für den Herausgeber immer mehr wurde, sondern ich befinde mich bei der Frage statistische Inferenz bei convenience samples in bester Gesellschaft mit allen Fachkollegen. Alles was ich versuchte, war nur die Sicht der Statistik verständlich und einsichtig zu machen. Allerdings wohl ohne Erfolg. Man will sich eben in Sachen "Statistik" nichts sagen lassen, sondern nur andere belehren.

Ich hätte vielleicht das RCS-Papier besser bei einer theologischen oder archäologischen Fachzeitschrift einreichen sollen. Dort wäre es zwar auch nicht veröffentlicht worden, aber es wären mir vielleicht einige Demütigungen erspart geblieben.