von der Lippe, Peter

Working Paper

Die Manie, für alles Zahlen und Statistiken haben zu müssen: Über Messbarkeit, Rankingmethoden und den geistlosen Umgang mit Signifikanztests

IBES Diskussionsbeitrag, No. 219

Provided in Cooperation with:
University of Duisburg-Essen, Institute of Business and Economic Studie (IBES)

Suggested Citation: von der Lippe, Peter (2016) : Die Manie, für alles Zahlen und Statistiken haben zu müssen: Über Messbarkeit, Rankingmethoden und den geistlosen Umgang mit Signifikanztests, IBES Diskussionsbeitrag, No. 219, Universität Duisburg-Essen, Institut für Betriebswirtschaft und Volkswirtschaft (IBES), Essen

This Version is available at:
http://hdl.handle.net/10419/129732

IBES DISKUSSIONSBEITRAG

Institut für Betriebswirtschaft und Volkswirtschaft

Nr. 219

Januar 2016

Die Manie, für alles Zahlen und Statistiken haben zu müssen

Über Messbarkeit, Rankingmethoden und den geistlosen Umgang mit Signifikanztests

Prof. Dr. Peter von der Lippe

IBES

WIWI INF

FAKULTÄT WIRTSCHAFTSWISSENSCHAFTEN

UNIVERSITÄT DUISBURG ESSEN

IBES DISKUSSIONSBEITRAG
Nr. 219
Januar 2016

Die Manie, für alles Zahlen und Statistiken haben zu müssen

Über Messbarkeit, Rankingmethoden und den geistlosen Umgang mit Signifikanztests

Prof. Dr. Peter von der Lippe (plippe@vwl.uni-essen.de)

Impressum: Institut für Betriebswirtschaft und Volkswirtschaft (IBES)
Universität Duisburg-Essen
Universitätsstraße 12
45141 Essen
E-Mail: IBES-Diskussionsbeitrag@medman.uni-due.de

Inhaltsverzeichnis

Einführung

Statistik ist eine Disziplin, die es erlaubt, zahlenmäßige ("quantitative") empirische Feststellungen zu treffen, wie z.B. "das Land L hatte am Stichtag t P_t = 12,7 Mill Einwohner". Andere, eher "nur" "qualitative" Aussagen, wie z.B. das Land L hatte "sehr viele" Einwohner, oder es hatte "ähnlich viele Einwohner wie das Land Λ" sind weniger erwünscht, obgleich dies in der Anfangszeit der Statistik als "vergleichende Staatenkunde" noch nicht so war.[1] Es gab seinerzeit nur wenige quantitative Feststellungen und man beschränkte sich damals noch dazu auch auf unschwer "messbare" Größen, wie etwa die genannte Einwohnerzahl. Wer "Einwohner" ist und wer nicht verlangt eine operationale Definition (der "Wohnbevölkerung"), die zwar schwieriger sein mag, als auf dem ersten Blick vermutet, aber ein Einwohner ist immer noch eine reale, sicht- und greifbare Person.

Inzwischen greift die Statistik aber thematisch viel weiter aus und man hält längst nicht mehr eine hinreichend "operationale" (d.h. eine die notwendigen Operationen, bzw. Entscheidungen für eine Messung benennende) Definition für erforderlich. Wir sind es mehr und mehr gewohnt, von der Statistik auch empirische Feststellungen über "ideelle", "konstruierte" Größen zu verlangen, die eine gedankliche Abstraktion darstellen wie z.B. die "Wirtschaftsleistung", gemessen als Inlandsprodukt. Niemand ist der Wirtschaftsleistung auf der Straße begegnet, und man kann sie auch gar nicht so einfach in wenigen Worten "definieren"; denn hier tun sich viele Fragen auf, z.B. ob Kosten für Forschung eine Investition oder ein Vorleistungsinput sind und wovon man hier die Entscheidung abhängig machen soll. Die praktische Statistik und insbesondere die inzwischen mehr und mehr in Vergessenheit geratene "Wirtschaftsstatistik"[2] ist – im krassen Gegensatz zur "Statistik" im Studium und in der Forschung – zum Großteil mit der Frage befasst, wie man etwas "misst", bzw. "messbar" macht. Auch wenn man mit der erreichten Kunst des Messens oft nicht zufrieden ist (man ist es im Falle der Wirtschaftsleistung sicher weniger[3] als im Fall der Einwohnerzahl), wäre es heutzutage trotzdem sehr verpönt, etwas als "nicht messbar" zu erklären.

Wir sind stattdessen schon zufrieden, wenn etwas überhaupt *irgendwie* gemessen wird, auch wenn es offensichtlich ist, dass der Gegenstand der Messung sehr komplex ist und der Stand der Messkunst noch sehr unbefriedigend ist. Man denke z.B. an die "Messung" des wissenschaftlichen Werts einer Veröffentlichung mit der Häufigkeit, in der die Publikation von anderen Autoren zitiert wurde. Wir haben hier die paradoxe Situation, dass alle sehr wohl wissen, dass die Anzahl der Zitate vielleicht kein guter Indikator dafür ist, ob eine Arbeit wissenschaftlich "wertvoll" ist oder nicht, dass sich aber trotzdem niemand trauen wird, den "wissenschaftlichen Wert" für "nicht-messbar" zu erklären.[4]

Diese Beobachtung ist Anlass, im Folgenden der Frage nachzugehen, warum das so ist, zumal

---

[1] Es begann quasi mit vorwiegend verbalen Reiseberichten oder Darstellungen der Geographie und Geschichte eines Landes, die nur zum Teil auch Zahlen als illustrative Beigabe enthielten.

[2] Das geht sogar so weit, dass kaum noch jemand etwas mit dieser Disziplin anfangen kann und inzwischen selbst das Wort "Wirtschaftsstatistik" auch ganz anders gebraucht wird, nämlich als "Statistik für WiWi-Studenten" (was sich dann meist als ganz "normale" Einführung in die üblichen statistischen Methoden entpuppt). Wohin das führt konnte ich neulich bei der Beurteilung einer zur Veröffentlichung eingereichten Arbeit sehen: der Verfasser glaubte, Neues beitragen zu können zur Deflationierung des Inlandsprodukts und glaubte bei der Deflationierung ginge es darum, die den Aggregaten (z.B. Privater Verbrauch) zugrundeliegenden *Mengen* einzelner Güter zu bestimmen (etwa x Tonnen Äpfel usw.). Er hätte das auch selbst leicht als unsinnig erkennen können, wo doch solche Aggregate auch Dienstleistungen enthalten.

[3] Man denke nur an die in fast regelmäßigen Zeitabständen immer wiederkehrenden Diskussionen über "Sozialprodukt vs. Wohlstand" und ob man nicht viel besser die Menschen einfach danach fragen sollten, wie glücklich sie sind. Dabei ist die Messung (oder auch schon nur die "operationale Definition") des "Glücks" ohne Zweifel eine große Herausforderung und viele meinen, man habe sie schon zufriedenstellend gemeistert (oder das Problem erfolgreich umschifft), indem man die Menschen bittet, auf einer fünf-Punkte Skala von 1 = sehr unglücklich bis 5 = sehr glücklich eine Zahl anzukreuzen.

[4] Es ist jetzt naheliegend, sich zu fragen: warum verlangen wir von der Statistik "Messbarkeit" auf der ganzen Linie und kann sie dieser Erwartung überhaupt gerecht werden? Der Beitrag geht dieser Frage nach.

- wir sonst, bei unseren Gedanken, Argumentationen und alltäglichen Eindrücken durchaus damit zufrieden sind, mit Größenvorstellungen zu operieren die weit entfernt davon sind, mit Messwerten, also Zahlen unterlegt werden zu können, und uns auch
- mit Sinneseindrücken zufrieden geben, die wir noch nicht einmal mit Worten – geschweige denn mit Zahlen – exakt beschreiben können, wie z.B. Farben, bei denen wir deshalb oft mit Vergleichen, wie "tannengrün" oder "moosgrün" arbeiten,[5] und
- im Zweifel auch oft geneigt sind, anekdotischen Berichten von Betroffenen mehr Glauben zu schenken, als den "nackten" Zahlen, wenn sie nur anschaulich und engagiert von jemand vorgetragen werden, der uns erfahren und kompetent genug erscheint.[6]

Man kann sich nun fragen, was für einen Nutzen einige ins Philosophische gehende Überlegungen zur Statistik (wie die folgenden) haben sollen. Ihr Nutzen könnte darin liegen, dass wir mit ihnen Grenzen der Statistik aufzeigen, die – und das ist unser Problem – in unserer Zeit immer weniger beachtet werden. Unsere These ist, dass

1. Messungen und Zahlen (als ihr Ergebnis) eine Faszination ausüben und dass uns der modisch gewordene Gebrauch der Statistik in der Politik und bei der Entscheidung über "Modelle" und "Theorien" dazu drängt, alles messen zu wollen, obgleich
2. was unsere Wahrnehmungs- und Ausdrucksmöglichkeiten betrifft zu einer Bevorzugung quantitativer und Geringschätzung nicht-quantitativer (qualitativer) empirischer Feststellungen eigentlich kein Anlass besteht,
3. dass man es aber trotzdem heutzutage allgemein sehr schwer hat mit Ausführungen, dass x nicht "messbar", oder x* als Maß für x nicht "aussagefähig" sei zu überzeugen;
4. nicht zuletzt weil wir mit einer äußerst beliebten und einfachen, aber kaum durchdachten Methode, die wir "Punktsummenmethode" nennen wollen, glauben etwas in der Hand zu haben, womit buchstäblich alles leicht "messbar" ist und wir
5. darüber vergessen, dass die oft für harmlos gehaltene Obsession, alles messen zu wollen auch Schaden anrichtet, nämlich in Gestalt einer Geringschätzung von wirtschaftsstatistischem Wissen und von verbalen Interpretationen von Methoden und Ergebnissen der Statistik, sowie von unzureichend begründeten Forderungen, möglichst detailliert und aktuell amtliche Zahlen bereitzustellen und eines geistlosen (weil es nur ein unverstandenes Ritual ist) Einsatzes von Signifikanztests.

Entsprechend gliedern sich die folgenden Ausführungen in fünf Abschnitte. Die Geisteshaltung, wonach buchstäblich alles "messbar" ist führt, zusammen mit dem Zwang, in Veröffentlichungen mit Mathematik und der Anwendung neuester komplizierter statistischer Methoden brillieren zu müssen, um in der Wissenschaft Karriere machen zu können zu

- einer Fließbandproduktion von (inhaltlich) irrelevanten und unoriginellen Fleißübungen in Mathematik und Statistik zur angeblichen empirischen Überprüfung (?) von "Theorien", wobei noch hinzukommt, dass für die Autoren *allein die Menge* solcher in der Wissenschaft notwendiger Pflichtübungen *zählt* und
- es schon deshalb tabu ist, sich über deren Fragwürdigkeit in verständlichen Worten zu äußern weil sie sich alle methodisch im Rahmen des "allgemein Üblichen" bewegen.

[5] Man kann nicht eine Skala definieren, nach der z.B. tannengrün und moosgrün die Werte 14 und 19 bekämen. Aber wir kommen trotzdem mit vagen Vergleichen zurecht, weil wir mit Sinnesorganen nicht besser ausgestattet sind.
[6] Ich nenne dies die "impressionistische" vs. der "statistischen" Methode der Erkenntnisgewinnung.

1. Was heißt "messen" und warum faszinieren uns Zahlen?

1.1. Messen

Eine Messung ist die Abbildung eines empirischen Relativs (Menge, für deren Einheiten Relationen definiert sind, etwa die Personen A, B,…) in ein numerisches Relativ (also in eine Zahlenmenge, wie die der natürlichen oder [umfassender] der reellen Zahlen): wenn A größer ist als B, verlangen wir von Messwerten der Körpergröße x, dass auch $x_A > x_B$ ist. Wenn A blond, B aber schwarzharig ist, soll für ein Maß y der Haarfarbe gelten $y_A \neq y_B$. Man könnte statt $y_A = 0$ und $y_B = 1$ auch $z_A = 5$ und $z_B = 3$ wählen, denn $0 \rightarrow 5$ und $1 \rightarrow 3$ wäre eine zulässig Transformation $y \rightarrow z$ (wenn $y_i \neq y_j$ muss auch $z_i \neq z_j$ sein). Denn bei der Haarfarbe ist nur die Äquivalenzrelation (=, ≠) definiert; d.h. blond ist nur ungleich, nicht besser oder schlechter als schwarzharig, bei der Körpergröße ist dagegen zusätzlich auch noch die Ordnungsrelation definiert. Danach, was definiert ist und welche Transformation der Zahlen zulässig ist unterscheiden wir verschiedene Skalen (Messniveaus; vgl. Tab. 1).

Tab. 1: Messniveaus

	Name der Skala	Eigenschaft: zusätzlich definiert ist	zulässige Transformation [1)]
1	Nominalskala [2)]	Äquivalenzrelation (=, ≠)	eins-zu-eins Transformation
2	Ordinalskala	Ordnungsrelation (>, <) [3)]	monoton steigend
3	Intervallskala	Einheit und Nullpunkt (beides willkürlich)	linear $y_i = a+bx_i$
4	Ratioskala [4)]	Nullpunkt nicht mehr willkürlich	proportional $y_i = bx_i$ (a = 0)
5	Absolutskala	auch Einheit nicht mehr willkürlich	keine Transformation mögl.

1) Die Zahlenangaben sind von Skala 1 bis 5 zunehmend aussagefähiger und deshalb dürfen sie auch immer weniger durch eine Transformation ($x \rightarrow y$) verändert werden.
2) man würde hier im Alltag meist nicht von "Messung" sprechen
3) man beachte, dass in Tab. 1 keine Relation wie "ist ähnlich wie" oder "ungefähr so wie" auftaucht
4) engl. ratio = Verhältnis (weil bei $y = bx$ auch $y_i/y_j = x_i/x_j$ gilt, sich die Relationen also nicht verändern)

Es gibt sicher Eigenschaften, bei denen die Vorstellung von einem Mehr oder Weniger näher liegt (wie etwa Körpergröße oder Temperatur) und solche, bei denen sie uns fernerliegt. Das Beispiel Temperatur zeigt dass das Skalenniveau eine Frage des Stands der Messkunst ist und nicht in der Natur der Sache liegt: früher konnte man nur *kalt* < *lauwarm* < *warm* < *heiß* unterscheiden und man war noch weit entfernt von dem, was man heutzutage meist allein als "Messung" bezeichnen würde, nämlich die Bestimmung einer *metrischen* Skala (ab Skalentyp 3) für die Eigenschaft "Temperatur". Das Beispiel aus der Geschichte mag dazu beitragen, für die Zukunft auch ähnliche Fortschritte für den zweiten Fall zu erwarten, d.h. bei Gegenständen, auf die sich erst in neuerer Zeit die Statistik erstreckt, nämlich Dinge, die nach Popper der "Welt 3" angehören,[7] wie z.B. eine wissenschaftliche Arbeit als ein Produkt des menschlichen Geistes, das zweifellos von mehr oder weniger Wert sein kann oder auch Sprachen, von denen gleich noch die Rede sein wird.

Will man die Messbarkeit einer Eigenschaft auf einem bestimmten Skalenniveau behaupten, d.h. eine solche Messung verteidigen, so ist

1. zumindest die Geltung der Charakteristika der Skala (was bei dem Skalentyp definiert ist) für die jeweilig zur Diskussion stehende Eigenschaft nachzuweisen und
2. besser (fundamentaler) noch die Messung mit dem *Begriff* (Konzept) der zu messenden Eigenschaft zu vergleichen.

Punkt 1 ist weniger anspruchsvoll als Punkt 2, auf den wir erst in Abschn. 3 eingehen werden. Wird der wissenschaftliche Wert (W) einer Veröffentlichung durch die *Anzahl* (Ratioskala!) der Zitate (Z)

[7] Nach Popper umfasst Welt 1 die gegenständliche (physische) Welt und Welt 2 psychische Zustände, Dispositionen usw. Welt 3 umfasst Erzeugnisse des menschlichen Geistes (Erzählungen, Mythen, Kunstwerke, Theorien etc.). Es sind unkörperliche, aber gleichwohl wirkliche (weil Wirkungen entfaltende) Dinge. Für Popper ist z.B. die Existenz ungelöster Probleme (z.B. unbewiesener Vermutungen in der Mathematik) ein Zeichen dafür, dass Welt 3 real ist. Vgl. Popper u. Eccles, S. 63 ff.

gemessen, so ist zumindest darzulegen, dass der W_i der Veröffentlichung i auch dreimal so groß ist, wie der von Veröffentlichung j wenn $Z_i = 3Z_j$ ist; denn das ist ja bei einer Ratioskala impliziert. Wenn es Gründe gibt anzunehmen, dass der Wert W_i nicht drei- sondern vielleicht nur zweimal so groß ist wie der Wert W_j (obgleich $Z_i = 3Z_j$), oder auch sogar geringer sein könnte als W_j (also $W_i < W_j$) ist eine Messung von W durch Z nicht möglich und man müsste sich beim Versuch, W zu messen etwas Anderes einfallen lassen. Ist $W_i = ½W_j$ obgleich $Z_i = 3Z_j$ dann wäre Z noch nicht einmal im Sinne einer Ordinalskala zu interpretieren. Z kann kein Maß für W sein und wir brauchen uns noch nicht einmal mit so problematischen Dingen wie der "Sinnhaftigkeit" der Anzahl der Zitate oder dem "Wesen" des wissenschaftlichen Werts einer Veröffentlichung auseinanderzusetzen.

Was als Argumente für eine Messung ins Feld geführt und als überzeugend akzeptiert wird kann sich im Laufe der Zeit ändern. So machten z.B. Johann Gottfried Herder und Wilhelm v. Humboldt die Auffassung populär, dass Sprachen die Qualität des Denkens prägen und weil die meisten Völker lieber zu den gut denkenden als zu den schlecht denkenden Völkern gehören wollen, förderte dies auch die Vorstellung von unterschiedlich hohen Entwicklungsstufen (E), also einer Ordnungsrelation zwischen Sprachen. Die Skalenunterscheidung von Tab. 1 ist noch zu differenzieren. So wird z.B. unterschieden

- ob für *alle* Ausprägungen eines Merkmals eine Relation (etwa >) definiert ist oder ob dies nur für eine Teilmenge der Ausprägungen gilt, und
- das Konzept einer *Typologie* im Unterschied zur Messung auf einer metrischen Skala.

Aus Platzgründen soll auf diese beiden Punkte hier nicht näher eingegangen werden. Eine Typologie gilt als unbefriedigende oder rudimentäre Sonderform der Messung: Typen werden meist durch Synthese vieler (auch messbarer) Merkmale definiert, aber die Typen T_1, T_2, … selber gelten nur als unterschiedlich, d.h. als auf einer Nominalskala abgebildet.

Wir sind bisher davon ausgegangen, dass Messung mit Angabe von nur einer Zahl verbunden ist, also eindimensional ist. Eine Eigenschaft x ist eindimensional wenn man jeweils eine ihrer Ausprägungen mit einer und nur einer Zahl beschreiben kann. Wenn *eine Zahl* x nicht ausreicht, sondern $m \geq 2$ Zahlen nötig sind, man also die Angabe eines *Vektors* **x** braucht,[8] ist die Eigenschaft mehrdimensional, etwa m = 2 im Beispiel der Einheiten A, B, und C mit den Spaltenvektoren $\mathbf{x}_A = \begin{bmatrix} 3 \\ 5 \end{bmatrix}$, $\mathbf{x}_B = \begin{bmatrix} 8 \\ 5 \end{bmatrix}$ und $\mathbf{x}_C = \begin{bmatrix} 6 \\ 4 \end{bmatrix}$. Man kann B hinsichtlich der zweidimensionalen Eigenschaft als "höher" einstufen als A, weil $x_{1B} = 8 > x_{1A} = 3$ aber $x_{2B} = x_{1B} = 5$. Aber A und C sind nicht so einfach vergleichbar, ohne eine "Reduktion" auf nur eine Dimension y durchzuführen. Eine von unendlich vielen Möglichkeiten wäre eine gewogene Linearkombination, etwa $y_1 = 0{,}2x_1 + 0{,}8x_2$ mit dem Ergebnis $y_{1A} = 4{,}6$, $y_{1B} = 5{,}6$ und $y_{1C} = 4{,}4$ eine andere wäre $y_2 = 0{,}8x_1 + 0{,}2x_2$ mit $y_{2A} = 3{,}4$, $y_{2B} = 7{,}4$ und $y_{2C} = 5{,}6$ so dass $y_{1C} < y_{1A}$, aber $y_{2C} > y_{2A}$.

1.2. Faszination von Zahlen

Zahlen haben seit jeher eine Faszination ausgeübt. Pythagoras wird eine geradezu mystische Beziehung zu Zahlen nachgesagt. Er muss sehr fasziniert gewesen sein von der Entdeckung des Zusammenhangs zwischen Zahlenverhältnissen und der Harmonie in der Musik, wonach das Verhältnis der Saitenlänge bei einer Oktave 1:2, Quint 2:3 oder Quart 3:4 (also 3/7 zu 4/7) ist. Auch die berühmte Gleichung $a^2+b^2=c^2$ ist im Falle der Katheten a = 3, b= 4 und der Hypotenuse c = 5 ein Verhältnis natürlicher Zahlen untereinander. In Zahlen (und damit auch in einer Gleichung) ausdrückbare feste empirische Beziehungen sind vor allem bei Himmelsbeobachtungen schon früh entdeckt und bestaunt worden, weshalb die Astronomie auch zu den ältesten Wissenschaften gehört. Sehr faszinierend

[8] Das ist zwar exakt, aber trotzdem unbefriedigend; denn es wird nicht gesagt, wie man entscheiden soll, ob man zwei oder mehr Zahlen braucht und warum eine Zahl nicht ausreicht.

dürfte auch der als drittes Keplersches Gesetz bekannte Zusammenhang $x^3 = cz^2$ sein, wobei x die Entfernung des Planeten zur Sonne, z seine Umlaufzeit und c eine Proportionalitätskonstante ist.

Dass Gleichungen dieser Art Messungen im engeren Sinne (also Messwerte auf einer metrischen Skala) verlangen hat das Streben, alles messen zu wollen enorm beflügelt. Die Nutzung von Zahlenangaben der Statistik für "Modelle" (die sich stets als Gleichungen oder Gleichungssysteme darstellen) ist eine Art Quantensprung im Verhältnis zur früher vorherrschenden Statistik als illustrative Beigabe[9] und hat zu einer wahren Massenproduktion von Regressionsrechnungen in der empirischen Forschung geführt, wobei die Originalität (als evtl. eine Dimension des besagten " Werts" einer Arbeit) oft nur darin besteht, dass man die gleiche Methode auf einen neuen, noch nicht untersuchten Datensatz angewendet hat.

Was bei solchen Anwendungen der Regressionsrechnung allerdings meist im Vordergrund steht ist nicht der Zahlenwert $\hat{\beta}$ für eine Schätzung des Regressionskoeffizient β, sondern allein ob es dieser erlaubt die Nullhypothese H_0: $\beta = \beta_0$ auf einem Signifikanzniveau von 5% oder 1% zu verwerfen und so die Zahlenangabe quasi zu einer Ja-Nein-Entscheidung zu "verdichten". Viele interessiert an der Statistik primär diese Leistung der "Verdichtung", weil sie quasi automatisch *durch eine Berechnung* erfolgt. Diese sog. "Asterisk Ökonometrie" (* bedeutet 10% und ** 5% und *** 1% Signifikanzniveau oder auch "hochsignifikant") ist deshalb bei Statistikanwendern hoch im Kurs, während die "Deskriptive Statistik" nur viele Zahlen liefert, die dann aber leider ihrerseits noch wieder (verbal) interpretiert werden müssen, weniger Ansehen genießt.[10]

Wir wollen hier nicht auf das verbreitete Missverständnis hinaus, mit "signifikant" sei bewiesen, dass die Hypothese H_0 – und damit die zur Diskussion stehende Theorie[11] – falsch sei (es besagt nur, dass die Beobachtung wenig *wahrscheinlich* wäre, wenn H_0 richtig *wäre*), sondern nur die Erwartung kritisieren, der Computer (die Statistik-Software) könne quasi automatisch ohne Mitwirkung des zu Fehlern neigenden Menschen ein Urteil über eine Theorie fällen und so die verbale, argumentative Würdigung einer Theorie durch Berechnung entbehrlich machen. Wir kommen darauf in Abschn. 5 zurück. So etwas zu erwarten ist reizvoll; denn der Drang, möglichst alles auf dem Niveau einer *metrischen* Skala zu *messen* und durch den Computer zu entscheiden geht auch Hand in Hand mit unserer meist sehr dürftigen Fähigkeit als "normale" Statistiknutzer, Zahlen gedanklich weiter zu verarbeiten und zu einer Integration oder "Gesamtschau" von Zahlen zu kommen.[12] Man sieht dies auch daran, dass bei der in mehr oder weniger regelmäßigen Zeitabständen wieder aufkommenden Diskussion über "Sozialprodukt versus Wohlfahrt" bzw. Qualität des Lebens immer wieder die eine, alles zusammenfassende Zahl, etwa das "Ökosozialprodukt" (y) gefordert wird, in dem Geldbeträge für Gutes (x_G) zum Sozialprodukt x hinzuaddiert und Geldbeträge für Schlechtes (x_S) vom Sozialprodukt abgezogen werden. Man gelangt dann meist nach einiger Zeit wieder zur Einsicht, dass die Größe $y= x+x_G-x_S$ (auch wenn alle x-Größen in der gleichen Einheit, nämlich in Geld gemessen werden) wenig

[9] Man könnte quasi von "Statistik 2.0" im Unterschied zur (nur deskriptiven) "Statistik 1.0" sprechen.

[10] Vielleicht ist das deshalb so, weil sie *viele* Zahlen und nicht die eine, alles zusammenfassende Zahl liefert. Bei Diskussionen mit Statistikanwendern in der BWL habe ich erfahren müssen (was mir neu war), dass offenbar unter ihnen der Umstand weitgehend unbekannt ist, dass man es in der "empirischen Wirtschaftsforschung", also dem, womit sich die meisten *volks*wirtschaftlichen Analysen der Wirtschaftsforschungsinstitute beschäftigen, meist "nur" mit Deskriptiver Statistik zu tun hat, die bloße Deskription also keineswegs eine Statistik minderen Ranges ist.

[11] Es ist schon abenteuerlich, dass man glaubt, eine ganze "Theorie" so "verdichten" zu können, dass sie auf einen bestimmten Zahlenwert β_0 für β hinausläuft. Hinzu kommt, dass die H_0 als Punkthypothese praktisch nie angenommen wird (also immer "signifikant" ist), wenn nur der Stichprobenumfang n hinreichend groß ist.

[12] Vor über zwanzig Jahren habe ich versucht, dies in einer eher etwas scherzhaften Fußnote über den Vergleich der Typenerkennung durch numerische Taxonomie einerseits und das Auge andererseits deutlich zu machen: "Man könnte versuchen, durch eine lange Liste von Messwerten etwa einen Hund zu charakterisieren. Angesichts der sehr unterschiedlichen Formen, Größen und Farben von Hunden dürfte das zwar nicht einfach sein, aber es ist denkbar. Die Zusammenschau der Zahlenfülle wird allerdings sehr schwierig sein, und so ist es auch nicht sicher, daß man in jedem Fall einen Hund von einer Katze oder gar einem Wolf unterscheiden kann, was dem Auge aber offenbar mit Leichtigkeit gelingt" (v. d. Lippe, 1995, S. 65).

Sinn macht [13]und man statt dessen nur ein "dashboard" von einzelnen Indikatoren anbieten kann. Unbefriedigend erscheint es auch, dass Armuts- und Reichtumsberichte viele Zahlen, nicht nur eine Zahl für die Anzahl der Armen liefern, je nachdem, wo man die Armutsgrenze legt (in Prozent vom arithmetischen Mittel, Median usw.).

Viele Menschen haben nicht nur Schwierigkeiten beim gleichzeitigen Betrachten vieler Zahlen, sondern auch dann, wenn es in einer Statistik nur um eine Zahl geht. Es fällt ihnen nicht viel mehr zu sagen ein, als "interessant" oder "hätte ich nicht gedacht" und das "Verstehen" von Zahlen geht oft nicht viel weiter als die (meist noch durch Graphiken erleichterte) Feststellung, dass 5 mehr ist als 4. Bei jeder Berechnung erhält man notwendig eine Zahl als Ergebnis, aber das heißt nicht, dass diese Zahl auch "sinnvoll" ist, wobei es – wie in Abschn. 3 gezeigt wird – sehr schwer ist, zu sagen, was "sinnvoll" macht, d.h. ihr "Sinn" verleiht oder warum eine Zahl "sinnlos" ist.[14].

Aber trotz solcher Probleme gibt es viele Vorteile von Zahlen – auch wenn es sehr viele sind, aus denen man sich erst einen Reim machen muss – gegenüber Worten. Man könnte im Wesentlichen an drei Vorteile der zahlenmäßigen gegenüber der verbalen Darstellung denken,

1. Die in Gleichungen, wie $x^3 = cz^2$ zum Ausdruck kommenden *Faszination der Mathematisierung der Empirie* wäre ohne Zahlen gar nicht möglich
2. Zahlen sind "*objektiver*" (7 wird von jedem als 7 verstanden und die Farbe F eines Objekts i beträgt $F_i = 18{,}3$ ist besser als F_i = "tannengrün", weil letzteres unterschiedliche Assoziationen erlaubt), d.h. Zahlen erlauben keine persönliche Färbungen und Umdeutungen, wie es bei Worten möglich ist, so dass mit ihnen auch eher ein Konsens zu erzielen ist als mit Reden (allerdings nur solange es um die Zahl selber geht) und
3. es ist auch ein "*mechanistisches Politikverständnis*", d.h. das Denken in Instrumenten (Hebeln) und deren Wirkung, was den Drang, alles zu messen beflügelt.

Wer ein Gaspedal und eine Bremse hat, wird sich auch einen Tachometer wünschen. Man verspricht sich in der Politik von Zahlen als "Zielwerte" auch, dass sie eine Überwachung der Zielerreichung erleichtern. Aber das ist in der Praxis nicht selten eine Illusion, weil das, was mit der Zahl gemessen werden soll, wie z.B. die "Defizitquote" beim Maastricht-Kriterium von 3% ein Wort ist, das der näheren Bestimmung bedarf. Zwar weiß jeder, dass 5% nicht 3% sind, aber die Defizitquote selber gibt Anlass zu einem endlosen Spiel, in dem abwechselnd eine Fraktion bemüht ist, die Definition zu "verwässern" und gerade deshalb die andere Fraktion umgehend bemüht ist, dem durch weitere "Härtung" der Kriterien und Verschärfung der Überwachung von deren Einhaltung gegenzusteuern.[15]

Die "Objektivität" von Zahlen besagt also wenig, was die daraus zu ziehenden Schlüsse betrifft. Dem Spiel "verwässern-härten-verwässern …" usw. entspricht ein Spiel, das oft bei Vorträgen zu beobach-

---

[13] Es ist, wie in von der Lippe 1995 gezeigt, sogar Produkt einer missverstandenen Volkswirtschaftlichen Gesamtrechnung (VGR), denn die VGR beruht auf einem Gleichungssystem (auf Basis der Kreislauftheorie) und verlangt deshalb eine doppelte Buchführung. Einfach x_G addieren oder x_S abziehen wäre eine Buchung ohne Gegenbuchung. Und wenn man auf das "Naturvermögen" "abschreibt" ist zu fragen, wer denn dieses Vermögen durch Investition geschaffen hat. Aber niemand stellt solche Fragen.

[14] Ein berühmtes Beispiel für eine sinnlose Zahl ist die Schätzung des Gesamtgewichts der Stadt Brüssel mit allen Häusern, Menschen etc. durch Adolphe Quetelet. Die geschätzte Zahl ist *sinnlos, weil es keinen Zusammenhang* gibt, *in dem die Zahl relevant wird* (denn niemand will Brüssel wegtragen). Einer Zahl "Sinn" verleihen könnte neben der Nützlichkeit für bestimmte Zwecke auch bedeuten: *Anschaulichkeit* und Vervollständigung eines Bildes, das man von einer Sache hat, *Vergleichbarkeit* mit Bekanntem, Erkennen bisher unbekannter Aspekte, sowie die Beseitigung von Lücken und Widersprüchen in unserem Wissen.

[15] Bei in Zahlen ausgedrückten Klimazielen erleben wir das gleiche Spiel. Jede von der Politik gesetzte Zahl kann eine Diskussion auslösen, in der dann umgehend diese Zahl in Frage gestellt wird. So wurde z.B. von Russland geltend gemacht, dass man dort wegen der vielen Wälder ein anderes Ziel bei der CO_2 Reduktion ansetzen sollte. Jede Zahl verlangt quasi als notwendige "Zugabe" Worte darüber, *was* mit ihr gemessen werden soll. Und hier wird klar, wie eng "zählen" und "erzählen" zusammenhängen. Die Grenze zwischen beiden ist nicht immer klar. In manchen Sprachen wird der Unterschied nur in der Orthographie (also graphemisch) gemacht: conter und compter oder im Geschlecht el cuento und la cuenta. Im Mittelhochdeutschen soll "zeln" allgemein der Akt des Zählens und Informierens und generell eine autorisierte Mitteilung bedeutet haben.

ten ist, wenn Zahlen aus Statistiken zur Untermauerung eines Standpunkts präsentiert werden: es gibt keine Statistik, die man sich zu Herzen nehmen *muss*, weil man immer genügend methodische und sachliche Einwände gegen sie anführen kann.[16] Man kann gegen jede Statistik argumentieren und die subjektiven, unklaren Worte der Sprache tragen dabei meistens den Sieg über die nüchternen, objektiven Zahlen davon. Warum ist das so? Wir behaupten, dass der Grund dafür ist, dass wir mehr Fähigkeiten haben im Umgang mit nicht-numerischen Wahrnehmungen und Begriffen, als mit numerischen Informationen.

2. Wir kommen mit nicht-zahlenmäßigen Sinneswahrnehmungen und Begriffen gut aus

Wenn man sagt, Statistik ist ein Instrument der *empirischen* Forschung, dann ist damit gemeint, dass die Zahlen der Statistik auf "Beobachtungen" beruhen, seien es direkt menschliche Wahrnehmungen oder von Maschinen registrierte Spuren von "beobachtbaren" Vorgängen, also indirekt menschliche Wahrnehmungen. "Beobachtbar" nimmt Bezug auf unsere Sinnesorgane, die uns mit überlebenswichtigen Sinneswahrnehmungen (SW) versorgen, uns aber keine Zahlenangaben liefern. Betrachten wir, nur als Beispiel für einen der fünf Sinne, den Geschmack. Er vermittelt uns über mehr oder weniger spezialisierte Rezeptorzellen eine SW, wobei wir allerdings schon große Schwierigkeiten haben, sie auch nur in Worten zu beschreiben, geschweige denn sie sinnvoll durch eine Zahl zu repräsentieren, also zu "messen". Es ist üblich, fünf Grundmuster (Geschmacks*richtungen* oder Grund*qualitäten*) wie süß, salzig, sauer, bitter und (neuerdings) umami[17] zu unterscheiden, aber es ist schwer

- diese Grundmuster zu definieren und sie exakt gegeneinander abzugrenzen (was unterscheidet sauer von bitter?),[18] oder
- auch nur Worte zu finden für unterschiedliche Intensitätsstufen in jeweils einem der fünf Grundmuster (gleichwohl können aber Probanden in entsprechenden Versuchen, Bewertungen auf einer Punkteskala etwa von 1 bis 5 angeben) und
- Geschmacksempfindungen von anderen Wahrnehmungen zu unterscheiden.[19]

Unsere Sinnesorgane liefern uns keine Zahlen, etwa durch Einblendung in ein Auge (selbst dort nicht, wo so etwas theoretisch denkbar wäre, wie z.B. bei der Wahrnehmung der Temperatur oder der Einschätzung einer Entfernung), sondern "nur" Eindrücke, die kaum mit Worten angemessen zu beschreiben sind. Man kann also festhalten:

- Die Natur kennt keine Überlegenheit von "Messbarem" gegenüber Eindrücken, die allenfalls mit sonst eher verpönten Assoziationen und Analogien ("tannengrüne" Farbe, "metallischer" Geschmack) zwischenmenschlich kommuniziert werden können und
- der Mensch als Individuum hat trotzdem in seiner natürlichen Umwelt kein Unbehagen am "Nicht-Messbaren" und kein Bedürfnis nach Messungen und
- dass es ihm gleichwohl möglich zu sein scheint, für Unterschiede in der Intensität von Wahrnehmungen, auch wenn er diese nicht "definieren" kann, Zahlen auf einer Punkteskala anzugeben (d.h. ein "rating" vorzunehmen),[20] spricht dafür dass das nicht zu gelten scheint, was stets

[16] Auf meiner Website findet man eine Satire unter dem Titel "Was tun, wenn einem eine Statistik nicht passt?"

[17] Nach einer neueren Auffassung ist "umami" (was aus dem Japanischen kommt und "Fleischgeschmack" bedeuten soll) eine fünfte Grundrichtung des Geschmacks, während die frühere Grundrichtung "herzhaft-würzig" nicht mehr als eine solche Grundrichtung gilt. Wie unterscheidet man, ab wann "Geschmäcker" eine *Grundrichtung* bilden? Es sind auch schon weitere Qualitäten, wie fettig, metallisch oder wässrig als zusätzliche Grundrichtungen diskutiert worden.

[18] Das gilt auch wenn ein Muster in einer (allerdings nicht eindeutigen) Relation zur Anwesenheit bestimmter Stoffe (wie z.B. Glutamat, Salze oder Wasserstoffionen H+) stehen mag. Was hier über den Geschmackssinn gesagt wurde, gilt auch für andere Sinne, z.B. für den Geruchssinn. Es heißt, der Mensch könne bis zu 10.000 unterschiedliche Duftstoffe erkennen, sie aber nicht, oder nur sehr stümperhaft beschreiben.

[19] so ist z.B. (neuerdings) "scharf" keine Geschmacksrichtung, sondern ein Schmerzsignal und bekanntlich ist der Geschmack auch sehr abhängig vom Geruch und von der Temperatur. Die Zungenoberfläche hat auch Sinneszellen für Temperaturunterschiede.

[20] In diesem Sinne wird heutzutage auch Befragungen der Menschen danach, wie glücklich sie sich fühlen, ein Erkenntniswert beigemessen.

in der Wirtschaftsstatistik gefordert wurde, dass nämlich erst einmal etwas zu "definieren" ist, bevor man es erhebt oder gar "messen" will.[21]

3. Schwierigkeiten, etwas für "nicht-messbar" zu erklären

Die Kritik daran etwas messen zu wollen oder an einer konkreten Art der Messung geht meist von eben dieser Vorstellung aus, dass erst einmal etwas zu "definieren" ist, bevor man es erheben oder gar "messen" kann. Wenn man z.B. die Messung des wissenschaftlichen Werts (W) einer Veröffentlichung durch die Anzahl der Zitate (Z) kritisiert, geht man von einem *Begriff* (Konzept) des W aus und hält Z für eine unangemessene Operationalisierung von W. *Das Dilemma ist, dass die berechtigte Kritik einer Messung meist nicht ohne Bezug auf das heutzutage allgemein nicht mehr akzeptierte "Wesen" einer zu messenden Eigenschaft auskommt und dass sie deshalb – auch wenn sie noch so naheliegend und berechtigt ist – oft unterbleibt.* Die im deutschen Idealismus gepflegte und auf Platon zurückgehende Vorstellung von einem unwandelbaren zeitlosen Wesen (im Unterschied zur wandelbaren "Erscheinung"), was auch "Formen" oder "Ideen" (etwa das Gute, Schöne, Gerechte) genannt wird, ist verpönt. Dieser sog. "Essentialismus" (Popper), d.h. die Spekulation über die wahre Natur, das "Wesen" einer Sache wird heute fast einhellig abgelehnt.[22] Der naheliegende Grund ist, dass es unklar und höchst umstritten ist, ob und wie man das "Wesen" einer Sache erfassen kann. Platon hielt hier Art eine Art "Vision", ein intuitives Erfassen, eine "*Wesensschau*" (Husserl)[23] für möglich. Aber das überzeugt heutzutage nicht mehr, zumal es bei jedem Messungsversuch den Einwand erlauben würde, man habe noch nicht genug, oder nicht tief genug "geschaut".

Wenn aber andererseits allenthalben Messungsversuche auch bei noch so abstrakten und komplexen Konzepten große Mode sind, fragt es sich, was die Nachfolge der umstrittenen Wesensschau angetreten hat. Es scheint die im folgenden Abschnitt 4 behandelte Punktsummenmethode (PSM) zu sein, eine Methode zu deren Begründung in der Literatur nichts Substanzielles zu finden ist. Wie wir zu zeigen versuchen, hat aber aus unserer Sicht die Wesensschau mit der PSM keinen würdigen Nachfolger gefunden; denn so wie man sich nicht mehr traut, viel über das "Wesen" (eine verpönte Vokabel) des Gegenstands einer Messung zu sagen, kann man auch bei der PSM nur wenig zur Begründung von Entscheidungen sagen, die man bei den Verfahrensschritten der PSM Methode getroffen hat.[24]

Eine Richtung in der Statistik, die einen sehr großen Wert auf die Interpretation (Plausibilität) von Verfahrensschritten einer Methode und auf Aussagefähigkeit von statistischen Zahlenergebnissen legte war die "Frankfurter Schule" in der Statistik. Sie war in Deutschland von den 1920er Jahren bis zu den 1980er oder 1990er Jahren sehr einflussreich. Ihr Anliegen war sehr zu begrüßen, aber sie gründete es auf eine Philosophie, die heutzutage mit Recht umstritten ist.[25] Kennzeichnend für sie war die Unterscheidung zwischen "Sachlogik" und "Zahlenlogik" und die Ansicht, dass viele Gegenstände (Konzepte) der Sozialwissenschaft inhärent nicht messbar sind, bzw. erst durch Bildung von dem Gegenstand adäquaten operationalen Begriffen messbar gemacht werden können. Das ist das sog. Adäquations- oder *Operationalisierungsproblem*, wonach eine operationale Größe, wie z.B. die Anzahl der Zitate (Z) zu finden ist für eine komplexe Erscheinung wie dem wissenschaftlichen Wert (W), um diese überhaupt messen zu können. Meine Kritik hieran ist,

- dass diese an der begrifflichen Arbeit an einer "adäquaten" Definition ansetzende Betrachtungsweise nicht auf alle Messprobleme anwendbar ist, dass
- es keine lehr- und lernbare Anweisung für das "Adäquat-machen" gibt und dass es

[21] Man denke nur an die enorme Arbeit, die weltweit in die Aufstellung und Harmonisierung von Klassifikationen und Systematiken sowie in die Abgrenzung von Aggregaten der Volkswirtschaftlichen Gesamtrechnung gesteckt wurde und wird.
[22] Vgl. Popper u. Eccles, S. 69 f.
[23] Popper u. Eccles, S. 216.
[24] Bei einer wenig durchdachten Methode kann man eben auch nicht viel zur Rechtfertigung einer Anwendung anführen.
[25] Ich habe mich mit ihr in v. d. Lippe 2012 und v. d. Lippe 2013 auseinandergesetzt, was zu meinem Bedauern einigen Lesern missfiel. Es ist mir deshalb wichtig, klarzustellen, dass ich die *Intention*, einer nur an mathematisch-formalen Aspekten interessierten Statistik etwas entgegenzusetzen sehr wohl begrüße und mich auch der Frankfurter Schule in meiner geistigen Entwicklung sehr verbunden fühle, aber die der Begründung ihrer Thesen dienende *Philosophie* für kritikwürdig halte.

- auch kein Maß dafür gibt, wie nah man mit einem Messkonzept (Z) an das intendierten Konzept (W) oder gar dessen "Wesen" durch Adäquation herangekommen ist.

Es fällt auf, dass man sich in der Frankfurter Schule neben dem "Wesen" ähnlich oft auf den "Sinn" einer Messung beruft. Aber so, wie es kein Maß für die Distanz zwischen Z und W gibt, dürfte es auch schwer sein, zu begründen, warum man z.B. die Messung von W durch Z für "sinnlos" oder trotzdem für "sinnvoll" hält. Der "Sinn" ist, ähnlich wie das "Wesen" eine Kategorie, auf die man sich heutzutage nur noch schwer berufen kann.[26] Es ist keine einer Messung (also einer Zahl) innewohnende Eigenschaft, sondern Ausdruck dessen, dass man eine zufriedenstellende Deutung (Interpretation) geben kann, also auch stets kontextabhängig.

Ähnlich problematisch war in der Frankfurter Schule der ständige Hinweis auf "Logik" und "Vergleichbarkeit". Man glaubte nur mit Logik zu einem allein richtigen Messkonzept gelangen zu können und mit einer "Logik des Vergleichs" (Herausarbeitung von Voraussetzungen und Grenzen der Vergleichbarkeit[27]) ein ähnlich achtbares Fundament für eine "logische" Statistik zu besitzen, wie es die mathematische Statistik mit der Wahrscheinlichkeitsrechnung besitzt. Man begab sich damit aber in die folgende Widersprüchlichkeit[28]

- man hält einerseits, ganz im Sinne von Descartes Traum[29] eine eindeutige Ableitbarkeit der einzig richtigen statistischen Methode (z.B. der einzigen richtigen Preisindexformel) allein mit den Mitteln der "Logik" für möglich (was allerdings klar illusionär ist),
- verwirft aber andererseits vehement, die auch mit Descartes Traum verbundene Idee der Einheit der Wissenschaft und plädiert stattdessen für einen Dualismus (Natur- und Geisteswissenschaft, Zahlen- und Sachlogik, mathematische und "logische" Statistik).

Aus solchen und anderen Gründen findet die Frankfurter Schule, trotz ihres nur allzu berechtigten Anliegens, heutzutage nicht mehr viel Anklang. Es ist auch klar, dass man - was die meisten Statistiknutzer nur zu gut kennen - endlos darüber streiten kann, ob etwas "messbar", "vergleichbar" oder "sinnvoll" ist. Das berechtigte Unbehagen an solchen Diskussionen sollte aber nicht dazu führen, bei Statistiken Zweifel an "Messbarkeit", "Vergleichbarkeit" oder "Sinnhaftigkeit" (wir haben hierfür, wie für das "Wesen" kein geistiges Wahrnehmungsorgan, keine untrügliche Intuition) zu unterdrücken und nicht mehr zu artikulieren. Aber genau das geschieht leider mehr und mehr.

4. Mit der "Punktsummenmethode" (PSM) ist buchstäblich alles messbar

Zur Beschreibung der PSM greifen wir von ihren unzähligen Anwendungen nur eine heraus. Es ist ein neuerlicher Index "Automatisierte Fahrzeuge" von einer Forschungsgesellschaft in Aachen, der "zeigt, welche Nationen auf dem Gebiete der selbstfahrenden Autos führend sind. Er umfasst zum einen Industrie-Indikatoren wie die Verfügbarkeit von (teil-) automatisierten Fahrfunktionen sowie Forschungsaktivitäten in diesem Bereich. Zum anderen erfasst der Index die rechtlichen Rahmenbedingungen für den Betrieb automatisierter Fahrzeuge sowie das aktuelle Marktvolumen für Fahrzeuge mit Fahrassistenzfunktionen wie z.B. Einparkhilfen."[30] Offenbar hat man hier Punktskalen entworfen

[26] Es hat sich gezeigt, dass Vertreter der Frankfurter Schule in vielen Fällen "sinnvoll" im Sinne von "anschaulich" oder "verständlich" gebraucht hatten. Wie die umfangreiche und meist nur schwer lesbare Philosophie des "Verstehens" zeigt ist bei solchen Begriffen nicht viel Klarheit und Exaktheit zu erwarten, obgleich viele Menschen meinen sie intuitiv richtig zu verstehen. Man glaubt zu wissen, wann etwas "sinnvoll" ist, hat aber Schwierigkeiten, das zu begründen.

[27] Bei "Vergleichbarkeit" ist die Frage, wie viele Merkmale einer Einheit betrachtet werden. Geht es nur um das Geschlecht männlich sind alle Männer, egal welchen Alters, welcher Nationalität usw. "vergleichbar". Fasst man dagegen sehr viele Merkmale ins Auge sind die Einheiten nur noch nicht vergleichbare Einzelfälle: "vergleichbar sein" ist also wie "sinnvoll sein" keine absolute, einer Sache innewohnende Eigenschaft.

[28] Sie wird besonders deutlich in der Position eines ihrer wichtigsten Protagonisten, Paul Flaskämper (vgl. v. d. Lippe 2013).

[29] Er beinhaltet die Idee (quasi das *Programm* des Rationalismus), dass eine (sichere) Methode "bei jeglicher wissenschaftlicher Forschung Anwendung finden sollte" (R&H, S. 23). Ähnlich dachte wohl auch Leibniz, der eine Methode für denkbar hielt, "mit deren Hilfe alle Probleme der Menschheit, seien sie nun wissenschaftlicher, juristischer oder politischer Natur, vernünftig und systematisch durch ein logisches Kalkül gelöst werden sollten" (R&H, S. 27).

[30] iw-dienst (Institut der deutschenWirtschaft Köln) Nr. 42 v. 8.10.2015, S. 8.

für die einzelnen Dimensionen, wie (A) automatisierte Fahrfunktionen, (F) Forschungsaktivitäten, (R) rechtliche Rahmenbedingungen und (M) Marktvolumen und die vom Land i erreichten Punkte A_i, F_i etc. gewogen (mit Gewichten g) gemittelt zu $y_i = g_A A_i + g_F F_i + g_R R_i + g_M M_i$ mit $g_A + g_F + g_R + g_M = 1$ oder ungewogen mit $g_A = \ldots = g_M = ¼$. Das beliebte aber wenig durchdachte Verfahren verlangt

1. eine Art Brainstorming, um festzulegen was alles als (Teil-)Dimensionen für das zu messende Konstrukt "relevant" erscheint (*Auswahl der Indikatoren*, wie hier A, F, R und M)
2. mit wie viel "Punkten" die verschiedene Ausprägungen in den (Teil-) Dimensionen zu bewerten sind (also das Problem der *Skalierung der Indikatoren*),
3. wie durch *Gewichtung* der unterschiedlich großen Relevanz der Indikatoren für das zu messende Konstrukt Rechnung zu tragen ist, und
4. ob überhaupt die m Indikatoren $x_1,\ldots,x_m$ (hier A, …, M) auf *eine* dahinterstehende, "*latente*" *Dimension* y abzubilden sind.

Es scheint für keines dieser Probleme der PSM eine statistisch fundierte "Theorie" und allgemein akzeptierte Empfehlung zu geben (v. d. Lippe u. Kladroba 2004). Punkt 4 ist ein Problem, das bei der PSM leider gar nicht erst aufkommt, weil die fälschlich für unproblematisch gehaltene *Summe* stets nur *eine* Zahl als Ergebnis liefert.

Was *Schritt 1* betrifft, so gibt es kein Kriterium dafür, ob und wann man "alle" Indikatoren berücksichtigt hat, ob einige redundant sind und andere fehlen. Das Problem ist nicht, Indikatoren und hierfür Zahlenangaben zu finden, das Problem ist, wie man garantieren kann, dass es die "richtigen" Indikatoren sind. Es ist heutzutage nicht mehr schwer, Daten über viele Indikatoren zu sammeln, so dass die Anwendung der PSM hieran nicht scheitern dürfte, ganz egal was mit ihr gemessen werden soll. Während die Angaben für A und F wohl auf einer Befragung von Unternehmen beruhen, wurden Daten für R und M offenbar mit einer Internetrecherche gewonnen. Die leichte Datenverfügbarkeit lässt erwarten, dass in Zeiten von "big data" mit noch viel mehr (auch fragwürdigen) Anwendungen der PSM zu rechnen ist.

Schritt 2: Würde man die Indikatoren in ihren "natürlichen" Maßeinheiten messen (etwa A in (Stück-) Zahl der Fahrfunktionen, M in Mill. €) dann wäre sofort klar, dass man eine Summe gar nicht bilden kann. Die Punktvergabe ermöglicht aber immer eine Summe zu bilden, und zwar auch bei völlig beliebigen Indikatoren.[31] Ohne Punktvergabe würde das offensichtlich werden, was mit Punktvergabe nur verdeckt wird, nämlich dass eine Zusammenfassung (Summe) der fraglichen Variablen inhaltlich vielleicht gar nicht zulässig und sinnvoll ist: Wie unten gezeigt wird, entscheidet die mehr oder weniger willkürliche Vergabe von Gewichten in *Schritt 3* zusammen mit den empirischen Korrelationen r_{12}, r_{13}, … der Indikatoren x_1, x_2, … untereinander darüber, wie stark ein Indikator mit dem "Index" y korreliert (also über r_{y1}, r_{y2}, …. und damit darüber, *was* y eigentlich misst).

Schritt 4: Die Frage, ob man einen Mittelwert (gewogene Linearkombination) y bilden darf und auf welchem Skalenniveau man damit etwas gemessen hat[32] ist nur eine von den vielen auf der Hand liegenden "formalen" Schwierigkeiten der PSM.[33] Eine Mittelwertbildung *über* n Personen (allgemein *Einheiten*) beim gleichen Merkmal x ist eine klare Sache (der Mittelwert repräsentiert eine "typische", "repräsentative" Einheit aus der Grundgesamtheit), nicht aber ein Mittelwert über m Merkmale x_{1i},

[31] Aber auch bei gleicher Maßeinheit der Indikatoren ist die Summe nicht notwendig sinnvoll. Die 2 kg, die man erhält als Summe von 1 kg Mehl und 1 kg Zyankali haben keine Bedeutung.

[32] Der Index "Automatisierte Fahrzeuge" nimmt Werte an zwischen 0 und 5 und Deutschland nimmt mit 3,1 den Spitzenwert ein, noch vor den USA (3,0) und Schweden (2,5). Man mag sich fragen, ob der Abstand zwischen Deutschland und Schweden sechsmal so groß ist wie der zwischen Deutschland und den USA (was ja eine Messung auf dem Niveau der Intervallskala implizieren würde) und ob es letzterer mit nur 0,1 wirklich erlaubt, sich in seiner Führungsposition getrost zurückzulehnen. Man kann hierauf keine Antwort geben, weil uns die PSM keine brauchbare Maßeinheit für den Abstand von 0,1 liefert.

[33] Es ist nicht nur eine formale, sondern auch eine inhaltliche Frage, weil – wie gleich gezeigt wird – die y Variable von anderer Art ist, als es die x-Variablen sind.

x_{2i}, …, x_{mi} bei der gleichen Einheit i (was repräsentiert er? Was tritt hier an die Stelle der Grundgesamtheit?)[34]

Aber die PSM impliziert noch sehr viel mehr, nämlich die Vorstellung, man habe mit

(1) $y_i = g_1x_{1i} + g_2x_{2i} + \ldots + g_mx_{mi}$

eine Variable gewonnen*, die von anderer Art ist, als* es *die x-Variablen* sind; denn sonst hätte man ja zur Messung eine beobachtete Variable wie x_1, …, x_m, (etwa x_{m+1})heranziehen können statt extra eine Variable y konstruieren zu müssen. Eine x-Variable, wie die Anzahl der automatisierten Funktionen ist eine "manifeste" Variable, d.h. Ergebnis einer *Beobachtung*, aber wettbewerbsfähig, d.h. mehr oder weniger "führend" im Wettbewerb sein ist eine "latente" y-Variable, d.h. Ergebnis einer *Beurteilung*. Auf diese beiden unterschiedlichen Variablentypen kommen wir noch zurück. Zunächst nur so viel: Was folgt aus (1) für eine Variable y? Bei m = 2 standardisierten Zufallsvariablen x_1 und x_2, (standardisiert heißt, dass $E(x_1) = E(x_2) = 0$ und damit auch $E(y) = 0$ und $\sigma_1^2 = E\left(x_1^2\right) = \sigma_2^2 = E\left(x_2^2\right) = 1$ ist), erhält man für die Varianz von y[35]

(2) $\sigma_y^2 = E(y^2) = g_1^2 + g_2^2 + 2g_1g_2r_{12} = g_1(g_1+g_2r_{12}) + g_2(g_2 + g_1r_{12})$.

Exkurs: *ohne* die Annahme *standardisierter* Variablen x_1 und x_2 erhält man im ungewogenen Fall $g_1=g_2=1/2$

(2a) $\sigma_y^2 = \frac{1}{4}\sigma_1^2 + \frac{1}{4}\sigma_2^2 + \sigma_{12}$ und da der Index y umso mehr zwischen den Einheiten differenziert je größer σ_y^2 ist, erklärt (2a) auch die sog. "implizite Gewichtung", d.h. eine Variable fällt "implizit" bei der PSM umso mehr ins Gewicht (trägt umso mehr zur Differenzierung bei), je größer ihre Varianz ist. Zurück zu (2): Für die Kovarianzen der x-Variablen mit y erhält man jetzt

(3) $\sigma_{y1} = E(x_1y) = E[x_1(g_1x_1+g_2x_2)] = g_1+g_2r_{12} = r_{y1}\sigma_y$ und

(3a) $\sigma_{y2} = g_{2+} g_1r_{12} = r_{y2}\sigma_y$ so dass für die Korrelationen mit y

(4) $r_{y1} = (g_{1+} g_2r_{12})/\sigma_y$ und

(4a) $r_{y2} = (g_{2+} g_1r_{12})/\sigma_y$, so dass $r_{y1}r_{y2} = \left(g_1^2r_{12} + g_2^2r_{12} + 2g_1g_2\right)/\left(g_1^2 + g_2^2 + 2g_1g_2r_{12}\right)$ gilt.

Mit der *empirischen* (also mit der mit den Daten *gegebenen*) Korrelation r_{12} und mit der Wahl der Gewichte g_1 und $g_2 = 1- g_1$ steht also auch (quasi automatisch) fest

- die Varianz σ_y^2 des mit der PSM gewonnenen "Indexes" y, die auch als gewogene Summe von Kovarianzen $\sigma_y^2 = g_1\sigma_{y1} + g_2\sigma_{y2}$ darstellbar ist,
- die Größe der für die Interpretation der Aussagefähigkeit der Messung (also für die Frage der *Validität*, was y eigentlich misst) wichtigen Korrelationen r_{y1} und r_{y2} zwischen den Indikatoren x_1 bzw. x_2 und dem "Index" y gegeben, wobei jedoch
- diese Korrelationen r_{y1}, r_{y2},… sehr verschieden sein können von den Korrelationen r_{y*1}, r_{y*2},…, die man mit einem evtl. existierenden externen Kriterium der Validierung erhielte, also mit einer Variable $y^* \neq y$, die wirklich das misst, was gemessen werden soll.[36]

---

[34] Die PSM wird auch oft "Indexmethode" genannt, dabei ist der Unterschied zu einem Index (z.B. ein Preisindex) gravierend. Bei einem *Preisindex wird gemittelt über die gleiche Variable* (Veränderungsrate eines Preises) bei verschiedenen *Ein "Index" im Sinne der PSM ist aber ein Mittelwert über verschiedene Variablen* bei jeweils einer Einheit. Besonders beliebt ist die PSM beim Ranking von Städten. Es gibt unzählige Beispiele hierfür, aktuell etwa "Das Hoch imSüden" in Wirtschaftswoche 49/2015, S. 67-73, wo mit über 100 Einzelindikatoren für Arbeitsmarkt, Wirtschaftsstruktur ,Immobilienmarkt und Lebensqualität (z.B. Arztdichte, Kitaplätze, Lebenserwartung etc.) gearbeitet wurde

[35] Die dargestellten Zusammenhänge lassen sich unschwer für den Fall m > 2 verallgemeinern, bei dem entsprechend auch Korrelationen r_{13}, r_{23}, r_{y3} usw. auftreten: Auch jetzt sind die Korrelationen r_{y3}, r_{y4} usw. eindeutig mit den Gewichten g_1, g_2, g_3, g_4 usw. und den empirischen Korrelation r_{jk} zwischen den x-Variablen bestimmt.

[36] Es wäre auch nicht zu erwarten, dass diese Korrelationen r_{y*1}, r_{y*2} bei gegebenem r_{12}, … auch so eindeutig durch die Wahl der Gewichte g_1, g_2 bestimmt wären, wie es die Korrelationen r_{y1}, r_{y2} sind.

Was in der PSM völlig ausgeblendet wird, ist die Größe der mit den Daten *gegeben* Korrelationen r_{12} etc. zwischen den beobachteten Variablen x_1 und x_2 etc. Sie ist – wie gesagt – entscheidend dafür, ob die beobachteten Indikatoren x_1, x_2,…. überhaupt auf *eine* und nur eine "dahinterstehenden" gemeinsamen Dimension y abzubilden sind.[37] Die PSM kombiniert die x-Variablen, egal, ob und wie stark sie untereinander korrelieren. Dabei sind wie (4) und (4a) zeigen, die Konsequenzen für die mit r_{y1} und r_{y2} gemessenen Aussagefähigkeit von x_1 bzw. x_2 für den Index y sehr unterschiedlich, je nachdem wie groß r_{12} ist. Wenn $r_{12} = 0$ ist, dann ist $r_{yj} = g_j \Big/ \sqrt{g_1^2 + g_2^2}$ (j = 1, 2) und falls ist , gilt $r_{y1} = r_{y2} = 1$; aber dann, bei $r_{12} = 1$ stellt sich die Frage, wozu man neben x_1 auch noch die hierzu faktisch identische Variable x_2 zur Messung braucht. Betrachtet man die Literatur zur PSM (vgl. v.d.Lippe/Kladroba 2004) so zeigt sich, dass es - was diese Korrelation betrifft - keine statistisch-methodisch begründete Empfehlung gibt. Es gibt also keine klare Antwort auf die so naheliegende Frage, ob Indikatoren, wie A, F usw. oder allgemein x_1, x_2,.. bei der PSM untereinander hoch oder gering (bzw. gar nicht $r_{12} = r_{13} = \ldots = 0$) korreliert sein sollten. In der Literatur werden beide Standpunkte vertreten,[38] wonach sowohl betragsmäßig hohe, als auch möglichst niedrige Korrelationen wünschenswert sind.

Ein weiteres formales Problem betrifft das Konzept, mit dem die Anordnung der Einheiten auf eine Skala zu begründen ist. Das folgende einfache Zahlenbeispiel weist auf eine Alternative zur Summenbildung hin:

Daten		
	x_1	x_2
A	2	18
B	3	17
C	17	3
D	18	2

Distanzen dividiert durch $\sqrt{2}$				
	A	B	C	D
A	0	1	15	16
B	1	0	14	15
C	15	14	0	1
D	16	15	1	0

Gemessen am Konzept der PSM stehen alle vier Einheiten gleich gut da (die Punktsumme ist jeweils 20, sie erlaubt also keine Differenzierung zwischen den Einheiten). Bei einer graphischen Darstellung (die Punkte A bis D liegen auf der fallenden Gerade $x_2 = 20 - x_1$) zeigt sich aber, dass gemessen an der jeweils nur $\sqrt{2}$ betragenden euklidischen Distanz zwischen den Einheiten A und B einerseits und C und D andererseits zwei Cluster (untereinander ähnliche Einheiten) vorliegen. Wir haben es mit zwei und nicht mit nur einem Cluster zu tun, weil es die großen Distanzen zwischen A und C mit $15\sqrt{2}$ und A und D mit $16\sqrt{2}$ (entsprechend zwischen B und C, bzw. B und D mit $14\sqrt{2}$ und $15\sqrt{2}$) nicht rechtfertigen, alle vier Einheiten in einen Topf zu werfen. Aber genau das geschieht wegen gleicher Punktzahl bei der PSM. Wir haben es also mit unterschiedlichen Konzepten der "Ähnlichkeit" zu tun: Der lineare Ansatz der PSM erlaubt es, dass ein Minus bei einem Indikator ($x_{2C} = 3 < x_{2B} = 17$) durch ein Plus bei einem anderen Indikator ($x_{1C} = 17 > x_{1B} = 3$) *kompensiert* wird. Die Clusteranalyse sieht keine solche Kompensation vor. A und B (bzw. C und D) wird als ähnlich einstuft weil *beide* Einheiten *jeweils in beiden Dimension* (x_1 *und* x_2) *gleichermaßen* hohe, bzw. niedrige Werte haben.

---

[37] Aber genau das kann – wie gesagt – bei der PSM kein Thema sein, solange die Summenbildung nicht in Frage gestellt wird; denn für eine Summe erhält man immer *eine* und nur eine Zahl und damit eine Messung in nur *einer* Dimension. Würde man im Fall der Index "Automatisierte Fahrzeuge" zu A, F, R und M auch noch P, die durchschnittliche PS Zahl der Fahrzeug eines Landes addieren, so könnte man argumentieren, dass die Summe der fünf statt vier Indikatoren unsinnig wäre und P auf einer anderen Dimension liegt. Dabei wird i.d.R. inhaltlich argumentiert, obgleich es hierfür auch "formale" Hinweise gibt. So dürfte P mit A (und ähnlich auch mit F, R und M) kaum korreliert sein, was vermuten lässt, dass P auf einer anderen Dimension liegt. Ein Hinweis auf latente Mehrdimensionalität wäre eine intransitive Anordnung von Einheiten, bei der für drei Einheiten, i, j und k gilt: wenn i < j und j < k ist, kann gleichwohl i > k sein.

[38] Mit m = 2 und $g_1 + g_2 = 1$ erhält man – wie leicht zu sehen ist – bei $r_{12} = 0$ die Werte $(\sigma_y)^2 = 1 - 2g_1$ und $r_{y1} = g_1/(1 - 2g_1)$. Noch bedenklicher ist die Annahme $r_{12} = 1$ (oder $r_{12} \approx 1$); denn dann ist quasi automatisch auch $\sigma = r_{y1} = r_{y2} = 1$.
Auch aus inhaltlichen Erwägungen dürfte die Aussage, die Indikatoren sollten unkorreliert sein ($r_{12} = 0$) eindeutig falsch sein.

Es gibt also wenig, was "formal" für die PSM spricht. Bevor wir zu formal fundierteren Alternativen zur PSM kommen, sollte noch ein Wort gesagt werden zu "inhaltlichen" Fragen (deren Diskussion allerdings oft unergiebig ist) und zur grenzenlosen Anwendbarkeit der PSM. Beim Index "Automatisierte Fahrzeuge" könnte man z.B. fragen, ob und in welchem Maße das aktuelle Marktvolumen M auch ein Indikator dafür ist, ob ein Land auf dem Gebiet auch "führend" ist (bzw. dies auch künftig sein wird, was der Index ja messen will), oder ob man nicht einen wichtigen Indikator vergessen hat. Das Problem ist, dass man sich bei solchen Diskussionen über den Wert einzelner Indikatoren und die Aussagefähigkeit des Indexes in die gleiche unbequeme Position begibt, nämlich verbal in Kategorien wie "Sinn" und "Relevanz" etc. argumentieren zu müssen, die heutzutage jemand hat, der begründen möchte, warum etwas "nicht aussagefähig" oder gar "nicht messbar" ist, weshalb so etwas auch meist unterbleibt.

Man könnte eine beeindruckende Aufzählung von Beispielen für die PSM vornehmen. Neben dem Index "Automatisierte Fahrzeuge" sei hier nur noch ein Beispiel erwähnt. Im Sport, der längst auch schon ein Gegenstand für die Statistik geworden ist, gibt es inzwischen ernst zu nehmende Versuche, den "Wert" y eines Sportlers, z.B. eines Baseballspielers durch Aggregation über sogar unzweifelhaft "messbare" Indikatoren x_1, x_2,… zu quantifizieren, wie etwa

x_1 =*Offensivrate* eines Spielers (Anzahl der Läufe, die er pro Auszeit hatte)
x_2 =*Verantwortlichkeit* eines Spielers (Anzahl der Auszeiten , die er verursacht hat) usw.[39]

"Es gibt kaum etwas, dem wir keine Zahlen zuordnen könnten"(D&H, S.36). Daraus wird oft gefolgert, dass man *allem* Zahlen zuordnen (also alles messen) kann und soll, zumal es mit der PSM so einfach ist. Aber warum hat diese Methode auf so naheliegende Fragen, wie

- wie viele und welche Indikatoren (Gilt je mehr desto besser? Wie wählt man sie aus?)
- sollen die Indikatoren x_j (j = 1, 2, ..., m Indikatoren) untereinander betragsmäßig nur gering oder aber hoch korrelieren und
- was rechtfertigt es, sie durch Summation $y_i = \Sigma_j g_i x_{ij}$ (i = 1,2,…,n Einheiten, j = 1,2,…,m Merkmale) zu *einer* Dimension y zu "verdichten"?

keine überzeugende Antwort hat? Das liegt daran, dass ihr kein Messmodell zugrundeliegt (die PSM ist eben ein typisches Beispiel für "measurement without theory"). Dabei ist die Messung mit einem *Messmodell* in der Statistik gar nicht so unüblich:[40]

- man kann die Menschen nicht danach fragen, wie lange sie noch leben werden und aus den Befragungsergebnissen die mittlere fernere Lebenserwartung e_x einer x-jährigen Person berechnen wollen,[41] gleichwohl ist aber eine Berechnung der Lebenserwartung möglich, und zwar durch das *Modell der stationären Bevölkerung*;[42]
- entsprechend kann man die Intelligenz einer Person nicht dadurch messen, indem man sie fragt, für wie intelligent sie sich hält (eine Messung ist aber trotzdem durchaus möglich dank des *Modells der Faktorenanalyse* [FA]).

Wie die FA ist die Latent Structure Analysis [LSA] ein *stochastisches* Messmodell (vgl. v. d. Lippe 1972). In beiden Modellen wird eine stochastische Beziehung (etwa $x_{ji} = \alpha_j + \beta_j y_i + \varepsilon_{ji}$ mit ε als Zu-

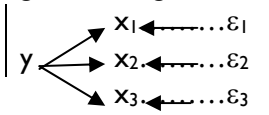
[39] vgl. Davis u. Hersh (D&H), S. 40f.

[40] Das wird z.B. beim "Adäquationsproblem" der oben erwähnten Frankfurter Schule übersehen. Man hat dort stets an Messung *nach vorangegangener Definition* (einer Eigenschaft oder einer Einheit, wie Betrieb, Unternehmen usw.)gedacht. Aber das betrifft nur einen Teil der Messprobleme. Man kann z.B. auch durch noch so viel Nachdenken über das "Wesen" der Intelligenz allein wohl kaum feststellen, wie intelligent im konkreten Fall jemand ist.

[41] Wenn man von *der* Lebenserwartung spricht, ist e_0 die Lebenserwartung eines (bzw. einer) Nulljährigen gemeint. Es ist klar, dass die Lebenserwartung vom Alter abhängt und etwa $e_{50} < e_{20}$ ist, weil Fünfzigjährige ja schon dreißig Jahre länger gelebt haben als Zwanzigjährige.

[42] oder "Sterbetafelbevölkerung": es ist ein *deterministisches* Modell (d.h. ohne Störgröße), dem die Vorstellung zugrunde liegt, dass ein heute 20 Jähriger (allgemein x-Jähriger) in 10 Jahren (allgemein in Δx Jahren) die gleiche einjährige Sterbewahrscheinlichkeit q_{30} (allgemein $q_{x+\Delta x}$) haben wird, die ein heute 30 Jähriger hat. q_x ist die *bedingte* Wahrscheinlichkeit, das Alter x+1 nicht mehr zu erreichen, *wenn* (bedingt dadurch dass) man das Alter x erreicht hat.

fallsvariable) zwischen m "manifesten" Indikatoren x_j, (j = 1,…, m), wie z.B. die Ergebnisse bei Intelligenztestaufgaben und der "latenten" Variable y (wie "Intelligenz")[43] postuliert.



Man könnte auch von einem "*datengenerierenden Prozess*" (DGP) sprechen, weil im Modell "Daten" wie x_j durch die zu messende "komplexe", dahinter stehende (latente) Variable y "generiert" werden.[44] Anders als mit der PSM kann man mit einem Messmodell auch mit den geschätzten Parameter $\hat{\alpha}_j$, $\hat{\beta}_j$ die Daten (d.h. die Korrelationen r_{jk} zwischen den manifesten Variablen x_j, x_k) "reproduzieren" (also $\hat{r}_{jk}$ schätzen) und so durch Vergleich von $\hat{r}_{jk}$ mit r_{jk} die Angepasstheit des Modells an die Daten überprüfen. So etwas kann die PSM nicht leisten, weil sie kein Modell eines datengenerierenden Prozesses liefert, keine Störgröße hat und auch durch Summenbildung *allen Daten*, d.h. allen Indikatoren x_1, x_2,… x_m, egal welchen Inhalts und wie sie korreliert sind) eine latente Eindimensionalität[45] aufzwingt.

Es ist hier nicht der Platz, um auf die FA oder die LSA näher einzugehen. Unser Ziel war nur, zu zeigen, dass die Statistik bei der Messung komplexer Variablen mehr zu bieten hat als die PSM. Der Preis dafür ist jedoch eine deutlich kompliziertere, d.h. schwerer zu erklärende und auch rechenaufwändigere Methode, die auch als deutlich überzogen erscheinen mag bei so manchen "kleineren" und weniger ernst genommenen Messübungen in der empirischen Sozialforschung. Trotz ihrer Mängel wird die PSM also in solchen Fällen meist als ausreichend akzeptiert wird, zumal ihre statistisch-methodischen Mängel wenig bekannt sind und wenig diskutiert werden."Messung"; denn leider gibt es so gut wie keine Auseinandersetzung von Statistikern mit der PSM. Wir geben uns also mit einer erkennbar unbefriedigenden Methode der Messung zufrieden und wir empfinden so etwas als eher als akzeptabel als ein Verzicht auf eine Messung. Dass "messen" per se besser ist als "nicht-messen" setzt stillschweigend voraus, dass die Manie, alles messen zu wollen[46] frei von Risiken und Nebenwirkungen ist.

5. Schaden, den die Obsession für alles Zahlen zu verlangen, anrichtet

Es gibt aber durchaus auch gravierende, Konsequenzen der Obsession für alles Zahlen haben zu müssen, und zwar solche,

1. von denen jeder Mensch, quasi "im Alltag" betroffen sein kann, und solche,
2. die speziell für die Statistik besonders relevant sein dürften.

zu 1: Was dies betrifft, so findet man im Buch von Davis & Hersh (D&H) einen Abschnitt mit der Überschrift "Die soziale Tyrannei der Zahlen", der einzelne Kapitel enthält über

- das Überreden (Rhetorik) mit (angeblich) unbezweifelbarer Mathematik/Statistik,
- die Auswahl von Bewerbern oder potentiellen Ehepartnern aufgrund digitalisierter, per Fragebogen ermittelter Merkmale und Selbsteinschätzungen, und die
- Bewilligung von Mitteln, Gewährung von Rechten etc. auf Basis von Schwellenwerten,
- das "Sieben" von Studenten im Studium mit "Scheinen" in Mathematik und Statistik

[43] Ein anderes Beispiel ist die Messung einer *latenten "attitude"* y (wie Rassismus) *durch manifeste* Antworten auf "*opinions*" x_1, x_2, …also mit ja oder nein zu beantwortende Meinungsäußerungen (z.B. Ich würde mich im Bus neben einen Schwarzen setzen, oder Ich würde einen Schwarzen als Schwiegersohn akzeptieren usw.).

[44] Im Unterschied zur LSA sieht die FA auch mehrere latente Variablen y_1, y_2,… vor. Bei nur einer latenten Variablen spricht man vom "Generalfaktormodell". Jedenfalls kann man mit der FA auch feststellen. ob überhaupt eine (latente) Eindimensionalität zugrunde liegt, die bei der PSM durch Summenbildung den Daten einfach aufgezwungen wird.

[45] Dass es nur *eine* latente Variable y gibt, das zu Messende also nur in einer Dimension abzubilden (mit nur einer Zahl zu kennzeichnen ist) ist keine Selbstverständlichkeit. Man hat ein Indiz dafür, dass mindestens noch eine zweite Dimension im Spiel ist, wenn es – wie gesagt – eine intransitive Anordnung von Einheiten (i, j, k) gibt, wenn i < j und j < k aber i > k ist.

[46] Eine bemerkenswert oberflächliche Betrachtung, nach der alles, wirklich auch alles, messbar sein soll, und zwar einfacher als man denkt, ist das Buch D. W. Hubbard, How to measure anything. Finding the Value in "Intangibles" in Business, 2nd. ed. (Wiley), 2010, ein Buch, das es immerhin zum Bestseller in den USA geschafft hat.

- und über Stellenbesetzungen nach dem erzielten Ergebnis in psychologischen Tests.[47]

"Testen ist immer ein Verfahren, um Menschen … rechnerisch handhabbar zu machen… Eine solche Entscheidung ist *automatisierbar* (das heißt sie kann von Maschinen ausgeführt werden) und *objektiv* (es tritt dabei kein menschliches Wesen offen in Erscheinung, von dessen Vorurteilen die Entscheidung abhängt) … In Wirklichkeit liegt das genaue Gegenteil vor: *Die Entscheidung ist zeitabhängig, fragwürdig und könnte auch ganz anders ausfallen*".[48]

Man kann zusammenfassend folgern, dass die "Tyrannei der Zahlen" darin besteht, dass Zahlen "Instrument sozialer Kontrolle zur Erhaltung des Status quo sind" (D&H, S. 139). Alles messbar machen erleichtert es, mit sanftem Druck konformes Verhalten zu erreichen.[49]

zu 2: Mit " kann von Maschinen ausgeführt werden", und es tritt "kein menschliches Wesen … in Erscheinung" erweckt Statistik den Eindruck des Objektiven und Unbezweifelbaren. Das mag es auch erklären, warum Statistik in einem modernen Verständnis von Politik und Wissenschaft so beliebt ist und dort zugleich aber auch so missbraucht wird, nämlich in Gestalt

a) von überzogenen Erwartungen an Daten der amtlichen Statistik in der Politik und von

b) statistischen Berechnungen als einen unentbehrlichen, aber oft ganz unverstandenen Bestandteil von so gut wie allen modernen wissenschaftlichen Arbeiten.

2a) amtliche Statistik: Es ist im Nachhinein oft grotesk, wenn nicht gar ziemlich peinlich, was alles einmal zur angeblich unverzichtbaren Datengrundlage für eine erfolgreiche Politik deklariert wurde und wie die amtliche Statistik zum Sündenbock gemacht wurde, weil sie diese speziellen Daten nicht lieferte.[50] So wollte man mit dem Arbeitsförderungsgesetz von 1969 präventiv die Entstehung *individueller* Arbeitslosigkeit im Ansatz verhindern. Der damalige Bundesarbeitsminister H. Ehrenberg forderte in einem Buch eine computergestützte tagesaktuelle Übersicht aller offenen Stellen im Rahmen eines regional und sektoral tief gegliedertem Frühwarnsystem, als ob das Ministerium ohne solche Daten, die es auch bis heute noch nicht gibt, seiner Aufgabe nicht gerecht werden könnte. Als die Öffentlichkeit für Probleme der Gesundheitspolitik so weit sensibilisiert war, dass man 1985 einen "Sachverständigenrat für die Konzertierte Aktion im Gesundheitswesen einrichtete, war es seine erste Tat, eine umfassende Gesundheitsberichterstattung zu fordern, die u.a. auch Modellrechnungen[51] und Daten über "psychische Risiken" von Personen (-gruppen) enthalten sollte (als ob nicht jeder zu jeder Zeit mit solchen Risiken konfrontiert wäre). Will man hier von "nicht notwendig" sprechen ist das Problem ähnlich wie bei "nicht messbar": man kann die "Nützlichkeit" auch ohne Beweis behaupten, verlangt aber von der "Nutzlosigkeit" einen Beweis, den zu erbringen kaum möglich ist.

In diesem Zusammenhang ist ein sehr fundamentales Prinzip der amtlichen Statistik, wonach jede Erhebung der Rechtsgrundlage durch ein Gesetz oder eine Rechtsverordnung bedarf, sehr hilfreich. Die amtliche Statistik bestimmt danach ihr Arbeitsgebiet nicht selbst, sie arbeitet im Auftrag des Gesetz- bzw. Verordnungsgebers und ist damit auch vor Angriffen enttäuschter Statistiknutzer, die eini-

---

[47] "*Der Intelligenzquotient ist ein Maß dafür, wie gut man mit einem Intelligenztest klarkommt*"(D&H, S. 137). Die Gefahr der Selektion mit Zahlen ist die "Bildung einer wissenschaftlichen Priesterkaste…, die vermutlich aus zweitrangigen Personen bestehen wird" (D&H, S. 121; Hervorhebungen jeweils im Original).

[48] D&H, S. 135. Hervorhebung im Original.

[49] Auch die erwähnte akademische Fließbandproduktion von Signifikanztests ist so ein konformes Verhalten.

[50] Die folgenden zwei Beispiele stammen aus dem Abschnitt "Computopia" in meinem Buch "Wirtschaftsstatisitk", erstmals in der vierten Auflage (1990, S. 216ff.) und dann in der fünften Auflage (1996, S. 261ff.). Mit "Computopia" (einem Ausdruck aus der Theorie der Wirtschaftsplanung in "sozialistischen" Ländern) ist die Geisteshaltung gemeint, dass "die Politik … umso besser ist, je umfassender die Lage- und Erfolgsbeurteilung ist. Der Optimierung aller gesellschaftlicher Prozesse steht damit nur die unzureichende Informationsbasis im Wege." Diese Geisteshaltung hat sogar die Wissenschaft ergriffen. Einige Volkswirte haben auf den Vorwurf, die Volkswirtschaftslehre habe die Finanzkrise im Jahre 2007 nicht vorhergesehen (und sei deshalb unbrauchbar!) geantwortet, man habe nicht rechtzeitig die richtigen statistischen Daten gehabt. Als ob alle Volkswirte Zahlen, wenn es denn die "richtigen" zur "richtigen Zeit" sind, gleich - und zwar "richtig" - interpretieren würden. Lutz Arnold, Die VWL steckt nicht in der Krise, ifo Schnelldienst 14/2009.

[51] z.B. auch über risikobezogene Kosten und Nutzen der GKV.

ge ihnen wichtig erscheinende amtliche Daten nicht finden, oder andere für überflüssig halten, geschützt.

Aus Sicht der Statistiknutzer in der Politik ist der Erfolg der Statistik primär ein Beweis für die Nützlichkeit von Sammeln und Aufbewahren (Speichern) um es später zu nutzen, und es gibt eine starke Tendenz hiermit fortzufahren z.B. in Richtung "big data". Das ist vor allem eine Herausforderung für die amtliche Statistik. So werden schon jetzt Scannerdaten im Einzelhandel und große Datensammlungen aufgrund eines Durchforstens des Internets zu privaten Preisindizes verarbeitet und diese dann als ebenbürtige Konkurrenzprodukte zum amtlichen Verbraucherpreisindex also der offiziellen Inflationsrate angesehen.

Die Statistik begann mit Datensammeln und zahlenmäßigen Beschreiben (zunächst nur im Sinne von Häufigkeiten auszählen). Erst später kam die Wahrscheinlichkeitsrechnung dazu (die übrigens auch als Teil der Mathematik relativ spät entstanden ist). Von Anfang an machte es den Reiz der Statistik aus, dass Sammeln, Speichern und Rechnen etwas sehr Mechanisches ist, was ohne ein "menschliches Wesen" "von Maschinen ausgeführt" werden kann und damit unangreifbar erscheint. Gerade weil die Statistik hiermit so erfolgreich ist, dürfte man es in dieser Richtung noch weiter treiben. Aber mit welchem Ergebnis?

2b Rolle der Statistik in der Wissenschaft: Gemeint ist hier vor allem die erwähnte akademische Fließbandproduktion von Asterisk-Ökonometrie und von Signifikanztests, mit denen selbst dort gerechnet wird, wo gar keine Stichprobe vorliegt und wo sich die Frage stellt

- welche "Hypothese" über die Grundgesamtheit man denn testen will, wenn die "Stichprobe" praktisch (vom Umfang her) die Grundgesamtheit(GG)[52], oder ein sog. convenience sample von gerade anwesenden Personen darstellt[53] und (schlimmer noch)
- was überhaupt die zu den Daten gehörende GG ist; denn man hat eine Stichprobe und muss (mit viel Phantasie) die dazu passende GG "nachliefern", und die Frage ist
- ob man überhaupt mit Tests und Konfidenzintervallen rechnen darf[54], wenn man irgendwelche Daten, z.B. von gerade einmal in einer Vorlesung anwesenden Studenten (ein sog. "convenience sample")[55] zu Stichprobendaten erklärt, obgleich *gar keine "Ziehung" nach dem Zufallsprinzip* aus einer eindeutig definierten GG erfolgt ist.

Ohne Auswahl (von n < N Einheiten) nach dem Zufallsprinzip gibt es keine Stichprobenverteilung einer Kenngröße (etwa des arithmetischen Mittels $\bar{x}$), d.h. keine Wahrscheinlichkeitsverteilung aller Werte, die man für $\bar{x}$ erhielte, wenn man alle $\binom{N}{n}$ Stichproben des Umfangs n aus einer Grundgesamtheit des Umfangs N nach dem Zufallsprinzip zöge. Man wendet bei den heutzutage geradezu obligatorischen empirischen Arbeiten Formeln für ein Konfidenzintervall $\hat{\mu}_{1,2} = \bar{x} \pm z_{\alpha}\left(\sigma/\sqrt{n}\right)$, bzw.

---

[52] Man stellt doch auch keine Vermutungen (Hypothesen!) an über das, was man bereits kennt.

[53] Ein Beispiel für ziemlich kühne Entscheidungen über sehr komplexe Theorien findet man in Strebinger et al. Dort wurden Beurteilungen von Besuchern einer Einkaufpassage (also keine Zufallsauswahl aus einer wohldefinierten Grundgesamtheit), die aufgrund von Fragen in Links- und Rechtshemisphäriker eingeteilt wurden (wie sicher ist das und gibt es überhaupt eine trennscharfe Unterscheidung dieser beiden Menschentypen?) über verbale vs. graphische Produktpräsentationen herangezogen. Mit Signifikanztests (hier wurden also alle Register der modernen Statistik gezogen) wurde über Parameter-Hypothesen (bzw. ganze Marketingtheorien reduzieren sich also auf Zahlen wie μ_0 für die Grundgesamtheit) befunden. Eine solche Hypothese war, dass "Linkshemisphäriker … versuchen einen analytischen und sequentiellen Denkstil anzuwenden, und sich darin von bildlichen Stimulusmaterial gestört fühlen …". Je nachdem, welchen Wert eine Prüfgröße annimmt, wenden also Linkshemisphäriker (d.h. Menschen die bevorzugt die linke Hirnhälfte nutzen) z.B. einen "sequentiellen Denkstil" (wie exakt das auch immer definiert sein mag) an, oder sie tun es nicht. Und das alles wird entschieden mit der *Stichprobenverteilung* einer Teststatistik, die es hier in Ermangelung einer Zufallsstichprobe gar nicht gibt (*ohne Zufall keine Wahrscheinlichkeit*).

[54] Man *kann* (rein rechnerisch) schon, *darf* es aber nicht, weil die Voraussetzungen der Rechenformel nicht gegeben sind.

[55] Ein Beispiel für eine solche Befragung, bei der dann die "strukturgleiche" Grundgesamtheit quasi nachgeliefert (man hat ja auch gar nicht aus einer definierten Grundgesamtheit gezogen) werden muss ist Müller, Voigt und Erichson.

für einen Test der Hypothese H_0: $\mu=\mu_0$ mit der Prüfgröße (Teststatistik) $z = \frac{|\bar{x} - \mu_0|}{\sigma/\sqrt{n}}$ an, die aber eine Stichprobenverteilung einer Kennzahl (Statistik) wie $\bar{x}$ mit der Varianz $\sigma_{\bar{x}}^2 = \sigma^2/n$ voraussetzt. Folgendes sollte eigentlich einleuchten (was aber offensichtlich nicht der Fall ist[56]): ohne Zufallsstichprobe keine Stichprobenverteilung (eine Wahrscheinlichkeitsverteilung). Man rechnet also mechanisch mit einer Formel, die nur als black box fungiert und gar nicht anwendbar ist. Und man entscheidet "objektiv" aufgrund des Rechenergebnisses ("von Maschinen ausgeführt") für z darüber, ob eine zu einer Zahl μ_0 verdichteten "Theorie"[57] richtig (nicht signifikant) oder falsch (signifikant) ist:

Wie sehr dies alles nur ein unverstandenes Ritual ist, wird auch daran deutlich, dass vielen noch nicht einmal klar ist, dass Schätzen und Testen zwei Seiten der gleichen Medaille sind. Dabei liegt der Zusammenhang auf der Hand; denn ein und die gleiche Gleichung wird nur nach der einen oder der anderen Größe aufgelöst: Ist die Zahl μ_0 innerhalb des Konfidenzintervalls $\mu_0 < \bar{x} + z_\alpha\left(\sigma/\sqrt{n}\right)$ oder $\mu > \bar{x} - z_\alpha\left(\sigma/\sqrt{n}\right)$, dann ist H_0 nicht abzulehnen (nicht signifikant)[58] liegt μ_0 außerhalb des Intervalls, dann ist H_0 abzulehnen (signifikant).[59]

Es gibt auch zahlreiche Anwendungen[60] der multiplen Regressionsrechnung mit Tests der geschätzten Koeffizienten $\hat{\beta}_j$ "gegen Null" (H_0: $\beta_j = 0$, Prüfgröße $\left|\hat{\beta} - \beta_0\right| / \hat{\sigma}_{\hat{\beta}}$) bei denen es sehr fraglich ist, welchen Wert die so gewonnene Erkenntnis haben soll.

So wurde – um hier nur einmal *ein* Beispiel zu nennen – mit Daten der WHO der Einfluss des "Mobbings" (engl. bullying) unter Kindern auf die Gesundheit (das subjektive Wohlbefinden) der Opfer untersucht und man stellte fest "being a victim of bullying reduces child subjective well-being substantively" (Klocke et al.).[61] Was ist daran so sensationell? Überraschend wäre es doch nur gewesen, wenn einem Kind das Gequält werden von anderen Kindern überhaupt nichts ausmachen würde. Dass es darunter leidet ist doch nur zu verständlich. Alles andere als unerwartet ist auch, dass der negative Einfluss (Betrag von $\hat{\beta}_j$)des bullying umso größer ist (sie sind auch alle hochsignifikant also ***), je häufiger ein Kind Opfer solcher Angriffe ist:

once or twice	-0,366	2-3 times per month	-0,623	once a week	-0,711	several times a week	-0,962

Über den Stichprobencharakter der Daten wird nichts gesagt. Dafür werden aber viele Größenvergleiche zwischen den Koeffizienten vorgenommen, nach dem Muster x_1 hat mehr Einfluss als x_2 auf y, weil $\left|\hat{\beta}_{y1.2...}\right| > \left|\hat{\beta}_{y2.1...}\right|$ ist.[62] So wurde z.B. gefolgert, dass "bullying" (x_1) schädlicher für das Wohlbefin-

[56] vgl. den Text "Das statistische Paralleluniversum der BWL" auf meine Website.

[57] Die Essenz der "Theorie" ist $\mu \neq \mu_0$ und die abzulehnende Aussage ist die Nullhypothese H_0: $\mu = \mu_0$. Es wird gefolgert: wenn H_0 abgelehnt wird, also "signifikant" ist die fragliche "Theorie" richtig. Man freut sich, wenn H_0 abgelehnt werden kann.

[58] Das sollte nicht zum Fehlschluss verleiten, dass Ablehnung (statt Annahme) von H_0 das grundsätzlich erwünschtere Ergebnis ist. Testet man z.B., ob die bei einer Regressionsanalyse implizit gemachten Voraussetzungen (etwa Streuungsgleichheit oder fehlende Autokorrelation der Störgrößen) erfüllt sind, ist man an der Annahme von H_0 interessiert.

[59] Das ist eine 1:1 Beziehung (also zwei Seiten einer Medaille). Es war mir nicht möglich, dies einem Herausgeber einer betriebswirtschaftlichen wissenschaftlichen Zeitschrift klar zu machen. Er glaubte es mir einfach nicht.

[60] Wir sprachen von einer "Fließbandproduktion" von methodisch unoriginellen und, wie das folgende Beispiel zeigt, auch inhaltlich uninteressanten empirischen Arbeiten.

[61] Das beste Modell von fünf durchgerechneten Modellen erreichte mit immerhin 17 Regressoren nur ein R^2 von gerade einmal 23,5% für die Selbsteinschätzung des Wohlbefindens als die zu erklärende Variable.

[62] Die Koeffizienten $\beta_{y1.2...}$ heißen partielle Regressionskoeffizienten (Einfluss von x_1 auf y bei Konstanz anderer Einflussfaktoren x_2, x_3 usw.). Man könnte sie auch multiple Regressionskoeffizienten nennen weil sie in einer multiplen

den (y) ist als andere Variablen x_2, …, wie Alkohol oder Nikotin sowie ungünstige Daten des betreffenden Landes, bei Inlandsprodukt, Jugendarbeitslosigkeit oder Staatsausgaben für Jugend und Familie. Das betrifft ein grundsätzliches Problem, das leider oft (und so auch in dem hier zur Diskussion stehenden Aufsatz) völlig ignoriert wird, nämlich:

Kann man aus einem Größenvergleich der Regressionskoeffizienten in einer multiplen Regression auf die relative Wichtigkeit einzelner Einflüsse (oder gar Ursachen) schließen? Das ist aus vier Gründen in der Regel nicht so einfach, nämlich weil

1. Regressions- im Unterschied zu Korrelationskoeffizienten maßstabsabhängig sind (mit dem lineartransformierten Regressor $x_1^* = a+bx_1$ statt x_1 erhält man einen anderen Regressionskoeffizient $\hat{\beta}_{y1^*.2...} \neq \hat{\beta}_{y1.2...}$), und man kann - wie unten gezeigt wird - deshalb auch solche Vergleiche allenfalls nur mit *standardisierten* Regressionskoeffizienten $\hat{\beta}^*_{y1.2...}$ nicht aber mit nichtstandardisierten Koeffizienten $\hat{\beta}_{y1.2...}$ durchführen;[63]
2. da die Regressoren x_1, x_2, … meist untereinander korreliert sind, messen die geschätzten partiellen Regressionskoeffizienten in der Regressionsgleichung (mit der Störgröße u_i, wobei wir der Einfachheit halber hier nur mit zwei Regressoren, also mit $y_i = \alpha + \beta_{y1.2}x_{1i} + \beta_{y2.1}x_{2i} + u_i$ arbeiten), auch immer den indirekten (über die Korrelation r_{12} vermittelten) Einfluss der jeweils anderen Variablen und nur wenn die Regressoren untereinander unkorreliert sind, ist auch die multiple Bestimmtheit R^2 einfach die Summe der partiellen (einfachen) Bestimmtheiten $R^2_{y.12} = R^2_{y1} + R^2_{y2}$;[64] und hinzu kommt, dass
3. solche Aussagen über die relative Wichtigkeit einzelner Einflüsse (oder gar Ursachen) eigentlich nur auf Basis der "wahren" Koeffizienten $\beta_{y1.2}$, $\beta_{y1.2}$ der Grundgesamtheit durchzuführen wären, und nicht auf Basis der Stichprobenkoeffizienten $\hat{\beta}_{y1.2}$ und $\hat{\beta}_{y2.1}$, die ja nur Schätzwerte für $\beta_{y1.2}$ und $\beta_{y2.1}$ sind; und schließlich wäre
4. mit entsprechenden Hypothesentests zu prüfen, ob die bei einer Regressionsgleichung implizit gemachten Modellannahmen (z.B. über die Störgröße u_i, aber auch über die Funktionsform, also ob ein *linearer* Ansatz vertretbar ist, und die Regressoren wirklich exogen sind usw.) erfüllt sind.

Wir gehen im Folgenden nur auf die ersten beiden Punkte ein. Dass Punkt 3 relevant ist, dürfte einleuchten, denn wenn man so einfach auf Basis von $\hat{\beta}_{y1.2}$ und $\hat{\beta}_{y2.1}$ auf die "Wichtigkeit". von x_1 relativ zu x_2 schließen könnte, bräuchte man eigentlich gar keine Schätz- undTesttheorie in der Statistik. Auch auf den weitgehend unbekannten Punkt 4 einzugehen würde den Rahmen unserer Darstellung sprengen; denn dieser Gegenstand wird üblicherweise in der Ökonometrie behandelt. Auch auf inhaltliche Fragen, ob hier nicht z.B. eine Scheinkorrelation vorliegt, kann hier aus Platzgründen nicht eingegangen werden.

Regressionsgleichung auftreten. Partielle und multiple Regressionskoeffizienten sind das Gleiche, aber partiellen und multiplen Korrelationskoeffizienten bezeichnen unterschiedliche Dinge. Man spricht von multipler und nicht multivariater Regression und Korrelation weil zwar mehrere *erklärende (unabhängige)* X-Variablen, aber immer nur eine *zu erklärende (abhängige)* Y-Variable im Spiel ist. In der multivariaten Analyse hat man es auch mit mehreren Y Variablen zu tun.

[63] Bei den nichtstandardisierten Regressionskoeffizienten kann allein die unterschiedlich große Standardabweichung der Regressoren x_j, x_k dafür verantwortlich sein, dass einer von zwei Regressionskoeffizienten der größere ist.

[64] Bei der einfachen Bestimmtheit schreibt man meist r^2, nicht R^2, wie bei der multiplen Bestimmtheit, aber die einfache Bestimmtheit (nur ein Regressor) ist ein Spezialfall der multiplen Bestimmtheit ($m \geq 2$ Regressoren).

Die geschätzten partiellen Regressionskoeffizienten $\hat{\beta}_{y1.2}$ und $\hat{\beta}_{y2.1}$ hängen mit den einfachen Regressionskoeffizienten $\hat{\beta}_{12}$ aus der Regression von x_1 (abhängige Variable, Regressand) auf x_2 (Regressor) sowie $\hat{\beta}_{21}$ (Regression von x_2 auf x_1) wie folgt zusammen:

(5) $$\begin{bmatrix} 1 & \hat{\beta}_{21} \\ \hat{\beta}_{12} & 1 \end{bmatrix}\begin{bmatrix} \hat{\beta}_{y1.2} \\ \hat{\beta}_{y2.1} \end{bmatrix} = \begin{bmatrix} \hat{\beta}_{y1} \\ \hat{\beta}_{y2} \end{bmatrix}$$, so dass in der Regel $\hat{\beta}_{y1.2} \neq \hat{\beta}_{y1}$ und $\hat{\beta}_{y2.1} \neq \hat{\beta}_{y2}$ sein wird, es sei denn x_1 und x_2 korrelieren nicht miteinander, wobei dann $\hat{\beta}_{12} = \hat{\beta}_{21} = 0$ und $r_{12}^2 = \hat{\beta}_{12}\hat{\beta}_{21} = 0$ ist. Den geschätzten nichtstandardisierten partiellen Regressionskoeffizienten $\hat{\beta}_{y1.2}$ kann man auch wie folgt schreiben $\hat{\beta}_{y1.2} = \frac{s_y}{s_1} \cdot \frac{r_{y1} - r_{y2}r_{12}}{1 - r_{12}^2}$ ($\hat{\beta}_{y2.1}$ entsprechend). Wie man sieht, hängt er von der Standardabweichung s_1 des Regressors x_1 ab. Multipliziert man dies mit s_1/s_y erhält man $\hat{\beta}^*_{y1.2} = \frac{r_{y1} - r_{y2}r_{12}}{1 - r_{12}^2}$, den geschätzten *standardisierten* partiellen Regressionskoeffizienten, der jedoch nicht zu verwechseln ist mit dem partiellen Korrelationskoeffizienten $r_{y1.2} = \frac{r_{y1} - r_{y2}r_{12}}{\sqrt{\left(1 - r_{y2}^2\right)\left(1 - r_{12}^2\right)}}$, der die Korrelation zwischen x_1 und y misst, bei Konstanz von x_2. Man kann also nicht einfach aus $\hat{\beta}_{y1.2} > \hat{\beta}_{y2.1}$ folgern, dass x_1 einen größeren Einfluss auf y hat als x_2. Genau das ist aber in der hier beispielhaft herausgegriffenen Arbeit über "bullying" geschehen, wenn es dort heißt. Die Regressionsrechnung "indicates that it is not the economy (GDP) or the level of spending on family policies which can foster child well-being. Rather it is the … school climate … child well-being looks to be more a result of the micro (family) and meso (school) level rather than the macro (society) level" (Klocke et al., S. 10). Auch das dürfte wieder eine Trivialität sein: warum sollte ein Kind mehr darunter leiden, wenn es in einem Land mit einem geringen Inlandsprodukt aufwächst als wenn es von Klassenkameraden geärgert wird? Wir behaupten nicht, dass dies ein falscher empirischer Befund ist. Wir sagen nur, dass dies nicht als eine durch die Regressionsrechnung bewiesene Tatsache hingestellt werden darf, d.h. der Vergleich von Regressionskoeffizienten liefert nicht den statistischen *Beleg*, den man hier glaubt, erbracht zu haben. Grenzen der Statistik sind hier aus Gründen mangelnder Kenntnisse der Methoden der Statistik klar nicht erkannt worden.

Das war nur ein Beispiel für eine Arbeit, in der eine Trivialität "entdeckt" wurde (Kinder leiden unter bullying), *und* statistische Methoden nicht beherrscht wurden. Man kann ohne Schwierigkeiten viele Arbeiten von ähnlicher Art finden und auch ziemlich sicher sein, dass

- wenn schon jetzt kein Mangel an solchen Arbeiten besteht. in Zukunft derartige methodische Unzulänglichkeiten empirischer Arbeiten eher mehr als weniger zu erwarten sind, einfach schon deshalb weil mit der gestiegenen Bedeutung solcher statistischer Berechnung keineswegs auch die Neigung gestiegen ist, sich ernsthaft mit den Feinheiten der statistischen Methoden zu beschäftigen (denn für viele ist heutzutage die Anwendung der Statistik ein Muss, ein unverstandenes Ritual, aber das Fach "Statistik" weiterhin ein Graus, dem man aber glaubt Herr zu werde, weil man mit Statistiksoftware wie SPSS einigermaßen erfolgreich umzugehen gelernt und geübt hat (Fehlanwendungen der Statistik dürften in Zukunft auch immer leichter werden an hierfür erforderliche Daten und Software heranzukommen).[65] Es könnte sogar sein,

[65] Aber die Kritik solcher massenhaft produzierter, aber ziemlich wertloser Anwendungen der Statistik ist nicht sonderlich gefragt. Sie ist zumindest viel weniger beliebt (wenn nicht gänzlich unüblich) als die in letzter Zeit in Mode gekommene Jagd nach fehlenden oder falschen Fußnoten in Dissertationen. Sie verlangt auch Kenntnisse der wissenschaftlichen Literatur und der Statistik, die – im Unterschied dazu – bei der so beliebten sog. "Plagiatsjagd" wohl eher entbehrlich sind.

- dass es in Zukunft, gemessen an den Fortschritten, die in letzter Zeit auf dem Gebiet der künstlichen Intelligenz erzielt wurden, sogar möglich sein wird, solche "empirische" Fließbandarbeiten auch – mitsamt der einleitenden verbalen Übersicht über die bisher publizierten Arbeiten auf dem fraglichen Gebiet, sowie der ebenfalls verbalen zusammenfassenden Würdigung der Ergebnisse – voll und ganz vom Computer selbst, ohne jede Mitwirkung menschlichen Autoren – erzeugt werden können (das könnte sogar auch für manche Auswertung von Erhebungen in der amtlichen Statistik gelten).

Die zu erwartende Entwicklung wird also wohl dahin gehen, dass wir noch mehr mit weitgehend automatisch generierten Zahlen über alles und jedes versorgt werden und vielleicht sogar auch noch – mit ebenfalls automatisch vom Computer generierten – Worten, die uns sagen, was von den Zahlen zu halten ist. Für Statistik im Alltagsleben bedeutet das, dass noch mehr von dem oben erwähnten (und nicht selten auch falschen) "Überreden … mit … unbezweifelbarer Mathematik/Statistik" (D&H) stattfinden wird und für Statistik in der Wissenschaft bedeutet das, dass man sich dann wohl, was die Art der Publikationen betrifft, mit denen sich der Nachwuchs qualifizieren kann, etwas anderes einfallen lassen muss.

Literatur

Davis, Philip, J. und Reuben Hersh, Descartes' Traum. Über die Mathematisierung von Zeit und Raum. Von denkenden Computern, Politik und Liebe (engl. Titel: Descartes' Dream. The World According to Mathematics), Frankfurt/Main 1988 (zitiert als D&H)

Klocke, A., Clair, A. u. J. Bradshaw, Being a victim of bullying reduces child subjective well-being substantively, An International Comparison, in: Informationsdienst soziale Indikatoren (ISI), Ausg. 53, (4/2015)

Müller, Holger, Voigt Steffen und Bernd Erichson, Ermittlung von Zahlungsbereitschaften mittels monadischer Preis- und Kaufabfragen, in: Marketing, ZFP, 2/2010, S. 117 – 127.

Popper Karl R. u. John C. Eccles, Das Ich und sein Gehirn, 2. Aufl., München u. Zürich, 1982.

Strebinger, Andreas, Hoffmann, Sabine, Schweiger, Günter und Thomas Otto, Realitätsnähe der Conjointanalyse in: Marketing, ZFP, 1/2000, S. 55 - 74.

von der Lippe, Peter (1972), Statistische Methoden zur Messung der sozialen Schichtung, Göppingen.

von der Lippe, Peter (1995), Die Messung des Lebensstandards, in W. Fischer (Hrsg.) Lebensstandard und Wirtschaftssysteme, Studien im Auftrag des Wissenschaftsfonds der DG Bank, Frankfurt.

von der Lippe, Peter (2013), Die "Frankfurter Schule" in der Statistik und ihre Folgen, Darstellung einer deutschen Fehlentwicklung am Beispiel der Indextheorie von Paul Flaskämper, AStA Wirtschafts- und Sozialstatistisches Archiv Bd. 7 S. 71- 89.

von der Lippe, Peter u. Andreas Kladroba (2004), Messung komplexer Variablen als Summe von Punktzahlen: Eine beliebte Methode des measurement without theory, in: Jahrbücher für Nationalökonomie und Statistik, Bd. 224, S. 115 – 134.

IBES

ISSN-Nr. 2192-5208 (Print)
ISSN-Nr. 2192-5216 (Online)