

# Das Bedeutsamkeitsproblem in der Statistik

Reinhard Niederée und Rainer Mausfeld

Im letzten Kapitel (Niederée & Mausfeld, in diesem Band), das dem Thema *Skalenniveau, Invarianz und „Bedeutsamkeit“* gewidmet war, haben wir uns im wesentlichen mit der Rolle von skalenniveaubezogenen Invarianzkriterien für numerische Einzelfallaussagen und gesetzesartige Aussagen befaßt. Dabei haben wir zu zeigen versucht, daß solche Kriterien nicht als aprioristische Bedeutsamkeitspostulate mißverstanden werden dürfen, daß ihnen aber dennoch eine fruchtbare substanzwissenschaftliche Rolle zukommen kann. In diesem Kapitel wollen wir uns mit verwandten Fragen im Bereich der angewandten Statistik befassen. Dabei werden wir die einschlägigen, im letzten Kapitel besprochenen Konzepte als bekannt voraussetzen (vgl. insbesondere die Abschnitte zum Skalenniveau und zu „naiven“ Invarianzkonzepten).

## 1 Stevens' Bedeutsamkeitskonzeption und die Folgen

Tatsächlich hatte Stevens (z.B. 1946, 1951) primär Probleme der angewandten Statistik im Sinn, als er sein einflußreiches skalenniveaubezogenes Bedeutsamkeitskonzept vorschlug. Konzepte der sog. statistischen Bedeutsamkeit oder Zulässigkeit, welche in der Stevensschen Tradition stehen, finden sich bis heute in einer Vielzahl von einführenden Texten zur angewandten Statistik und zum experimentellen Design. Dabei werden für ein „gegebenes“ Skalenniveau der betrachteten Variablen nur solche statistischen Kennwerte (Statistiken) als „bedeutsam“ oder „zulässig“ eingestuft, welche gewisse Invarianzkriterien bezüglich der Menge der zulässigen Transformationen dieses Skalenniveaus erfüllen (sog. *meaningful*, *permissible* oder *appropriate statistics*).

Eine konzeptuell strengere Weiterentwicklung des Stevensschen Ansatzes wurde in Anlehnung an die in Suppes und Zinnes (1963) vertretene *meaningfulness*-Konzeption von Adams, Fagot und Robinson (1965) vorgeschlagen. Diese ersetzt zum einen die irreführende Sprechweise von „dem Skalenniveau“ einer Variablen durch entsprechende repräsentationstheoretische Konzepte (vgl. die entsprechende Diskussion im letzten Kapitel). Zum anderen werden verschiedene Invarianzkonzepte systematisch unterschieden und in diesem Zusammenhang der Blick auf bestimmte *Aussagen* gelenkt, in denen statistische Kennwerte verwendet werden sollen, wobei insbesondere solche Aussagen betrachtet werden, welche den Vergleich von Kennwerten betreffen. Auch wir wollen im folgenden diese weithin als notwendig anerkannte Umformulierung des statistischen Bedeutsamkeitsproblems zugrunde legen und uns in unserer Diskussion im wesentlichen auf statistische Aussagen/Hypothesen konzentrieren (vgl. auch die Anmerkungen zu numerischen Indizes am Ende des letzten Kapitels,

sowie Pfanzagl, 1971; Klein, 1984; Luce, Krantz, Suppes und Tversky, 1990, Kap. 22; Chiang, 1995).

Der hier angesprochenen Tradition zufolge gilt beispielsweise der ordinale Vergleich von Mittelwerten in bezug auf Ordinalskalen als unzulässig/unangemessen, da sich nach einer *ordinalen* Transformation – d.h. wenn man die betrachtete Variable  $X$  durch eine Variable  $\tau \circ X$  ersetzt, welche sich dadurch ergibt, daß man  $X$  mittels einer streng monoton wachsenden Abbildung  $\tau$  transformiert – die Ordnung der Mittelwerte umkehren kann. Der Wahrheitswert (wahr bzw. falsch) einer entsprechenden Aussage ist also unter Umständen nicht invariant bezüglich einer solchen Transformation.

Als „für Ordinalskalen zulässiges“ Maß der zentralen Tendenz gilt in einem solchen Vergleich dagegen der Median; ein Vergleich zweier arithmetischer Mittel würde dagegen als zulässig erachtet für Skalenfamilien, welche zumindest das Skalenniveau einer Intervallskala besitzen. In beiden Fällen bleibt bei einem Vergleich zweier Gruppen die Ordnung der jeweiligen Maße der zentralen Tendenz unter zulässigen Transformationen der Variablen erhalten. Auf diese Weise kann man zu einem Schema gelangen, welches beliebige statistische Kennwerte bzw. bestimmte Aussagen über solche Kennwerte als für bestimmte Skalenniveaus „erlaubt“ oder „unerlaubt“ klassifiziert (siehe z.B. Stevens, 1951; Luce & Krumhansl, 1988).

Stevens' restriktive Postulate wurden bereits von Lord (1953) und Burke (1953) als ungerechtfertigt und unangemessen zurückgewiesen, da die Statistik mit Zahlen beginne und mit Zahlen ende und deshalb unabhängig sei von den zugrunde gelegten Meßprozeduren und somit auch von Skalenniveaubetrachtungen. Dies leitete eine lang anhaltende – und durch einer Reihe von konzeptuellen Konfusionen beeinträchtigte – Kontroverse ein, deren gegensätzliche Positionen die psychologische Methodenlehre bis heute durchziehen.

Eine befriedigende Analyse, welche der Vielschichtigkeit dieser Frage und dem damit verbundenen Zusammenspiel statistischer, wissenschaftstheoretischer und meßtheoretischer Aspekte gerecht werden könnte, bleibt unseres Erachtens noch zu leisten. An dieser Stelle kann daher nicht mehr als der Versuch einer Klärung unternommen werden. Hierzu werden wir eine mögliche Position jenseits der üblichen Lager skizzieren. Ebenso wie im letzten Kapitel wird sich dabei erweisen, daß zwischen verschiedenen Konzepten von „Bedeutsamkeit“ zu unterscheiden ist und daß die *normative* Verwendung skalenniveaubezogener Invarianzkriterien (etwa in Form von „Verboten“ der Art: „Mittelwertvergleiche sind auf Ordinalskalenniveau unzulässig“) in der üblichen Form nicht haltbar sind. Dennoch beruhen Bemühungen um ein Bedeutsamkeitskriterium auf einem durchaus berechtigten methodologischen Unbehagen (was ist zum Beispiel von einem Mittelwertvergleich bei Ratingskalen zu halten?). Diesem Problemkreis werden wir uns im zweiten Teil des Kapitels zuwenden.

Zwei allgemeine Hinweise wollen wir voranstellen: In unserer Erörterung statistischer Aussagen wird im folgenden gewöhnlich von *statistischen Hypothesen* im Sinne der Inferenzstatistik die Rede sein; soweit sich die folgenden Ausführungen auf die statistischen Hypothesen selbst beziehen, übertragen sie sich sinngemäß auf deskriptiv-statistische Aussagen. Dabei werden wir uns im folgenden exemplarisch auf Vergleiche zweier Populationen bezüglich „geeigneter“ Maße der zentralen Tendenz

(und hier wiederum beispielhaft auf den Median und den arithmetischen Mittelwert, d.h. den Erwartungswert) konzentrieren; auch hier verallgemeinern sich unsere Betrachtungen sinngemäß auf andere Klassen von Hypothesen.

## 2 Statistische Hypothesen vs. statistische Tests

Bei einer Diskussion von Bedeutsamkeitsbetrachtungen im Umfeld inferenzstatistischer Hypothesentestens muß zunächst streng zwischen zwei konzeptuellen Ebenen unterschieden werden, welche in diesem Zusammenhang häufig konfundiert werden: einerseits der Ebene der *statistischen Hypothesen* selbst, welche sich auf Populationen (bzw. Zufallsvariablen) beziehen, und andererseits der Ebene der für das Testen solcher Hypothesen anhand von Zufallsstichproben herangezogenen *Teststatistiken* und der mit ihrer Anwendung verbundenen Probleme (Verteilungsannahmen etc.). Wir werden uns in diesem Abschnitt zunächst den Hypothesen selbst unter dem Aspekt ihrer „semantischen Bedeutsamkeit“ zuwenden und sodann auf das Problem angemessener („zulässiger“) Tests zu sprechen kommen. Es herrscht unter den maßgeblichen Fachvertretern weitgehend Einigkeit darüber, daß für diese beiden im folgenden betrachteten Aspekte skalenniveaubezogene Invarianzkriterien in der Tat unangemessen sind.

### 2.1 Die semantische Bedeutsamkeit statistischer Hypothesen

Für den Fall von numerischen Einzelfall- und gesetzesartigen Aussagen wurde die Frage der semantischen Bedeutsamkeit bereits im Abschnitt „Naive numerische Invarianzpostulate“ des letzten Kapitels behandelt. Die dortigen Überlegungen behalten auch für statistische Aussagen ihre Gültigkeit. Betrachten wir beispielsweise für eine auf einer *fest gewählten Einzelskala* basierenden Zufallsvariable (ZV)  $X$  die entsprechenden Erwartungswerte  $\mu_1 = E(X_1)$  und  $\mu_2 = E(X_2)$ , welche sich wie üblich auf die Verteilung der Variablen  $X$  für zwei Populationen bzw. experimentelle Bedingungen beziehen. Sofern sich überhaupt sinnvoll von entsprechenden ZVn sprechen läßt, ist z.B. die Aussage  $\mu_1 > \mu_2$  wahr oder falsch und somit in diesem Sinne selbstverständlich semantisch sinnvoll (semantisch „bedeutsam“<sup>1</sup>). Der genannte triviale Befund hat offensichtlich nichts mit dem Skalenniveau einer Skalenfamilie zu tun, als deren Element man diese Einzelskala „auffaßt“, und nichts mit der Frage, ob eine Hypothese der Form  $\mu'_1 > \mu'_2$  stets den gleichen Wahrheitswert hat wie die Hypothese  $\mu_1 > \mu_2$  selbst, wenn Erwartungswerte der Gestalt  $\mu'_1 = E(\tau \circ X_1)$  und  $\mu'_2 = E(\tau \circ X_2)$  für ordinale Transformationen  $\tau$  betrachtet werden. Natürlich hängt der Wahrheitswert solcher Hypothesen von den zugrundeliegenden Verteilungen (und damit wesentlich von empirischen Gegebenheiten) ab, ist also strenggenommen sicher kein Problem, das „mit Zahlen beginnt und mit Zahlen endet“ (Burke, 1953, S. 74).

<sup>1</sup>Wie im vorigen Kapitel wollen wir uns hier den schwierigen philosophischen Fragen gegenüber, welche sich an Begriffe wie den der „Wahrheit“ von wissenschaftlichen und statistischen Aussagen knüpfen, dem *common sense* entsprechend „naiv“ verhalten, da dies für eine Klärung der hier zu behandelnden Fragen völlig ausreicht.

Fallen die Wahrheitswerte derartiger Hypothesen auseinander, so muß man selbstverständlich die betrachtete Skala so weit wie nötig spezifizieren. Insbesondere kann man in einem solchen Fall bei Bezugnahme auf eine Ordinalskala nicht pauschal, d.h. ohne Nennung der zugrundeliegenden Einzelskala, davon sprechen, daß eine Gruppe „im Mittel“ höhere Werte aufweise als die andere; eine solche Äußerung wäre dann in der Tat bereits in einem semantischen Sinne bedeutungslos. Die verschiedenen Hypothesen  $\mu_1 > \mu_2$ ,  $\mu'_1 > \mu'_2$  etc. sind in einer solchen Situation (anders als bei Bezug auf eine Intervallskala, wo sie mathematisch äquivalent sind) als *separate* Hypothesen zu betrachten. Die Frage, ob sich unter diesen Hypothesen auch solche befinden, die tatsächlich *inhaltlich interessant* sind, ist eine hiervon zu unterscheidende Frage, der in der Tat häufig nicht genügend Beachtung geschenkt wird; sie soll in Abschnitt 3 diskutiert werden.

## 2.2 Gibt es „unzulässige“ statistische Tests?

Gehen wir zuvor noch kurz auf Probleme ein, die mit der oben angesprochenen zweiten Ebene, nämlich der Anwendung *statistischer Test* für derartige Hypothesen verbunden sind, da häufig auch hier aufgrund von Skalenniveaubetrachtungen von zulässigen und unzulässigen *Tests* gesprochen wird, wie etwa in der Behauptung, daß der *t*-Test für Ordinalskalen unzulässig sei.

Häufig werden hierbei wieder die oben genannten Ebenen konfundiert: Ist die Hypothese selbst gemeint (hier ein Vergleich von Erwartungswerten) oder tatsächlich das Testverfahren? Nur letzteres wollen wir im Augenblick betrachten. Sei also eine bestimmte Hypothese (wie etwa  $\mu_1 > \mu_2$  oder  $\mu'_1 > \mu'_2$ ) vorgegeben, welche inferenzstatistisch untersucht werden soll. Es stellt sich dann die Frage: Gibt es skalenniveaubezogene Kriterien für die Zulässigkeit gewisser statistischer Tests (d.h. der darin verwendeten Teststatistiken) für diese vorgegebene Hypothese (bzw. ein entsprechendes Hypothesenpaar  $H_0, H_1$ )?

Ohne Frage gibt es für eine skalenniveaubezogene Unterscheidung zulässiger und unzulässiger Teststatistiken für eine gegebene Hypothese keinerlei Begründung; sie ist offensichtlich unangemessen (vgl. für das meßtheoretische Schrifttum Luce et al., 1990, Abschnitt 22.6).<sup>2</sup> Wesentlich ist lediglich die Frage, ob die *mathematisch-statistischen Voraussetzungen* für eine angemessene Anwendung im gegebenen Fall – also insbesondere entsprechende *Verteilungsannahmen* – erfüllt sind (oder zumindest nicht „zu grob“ verletzt). Ob dies der Fall ist, ist natürlich wiederum keine Frage, welche mit Zahlen beginnt und mit Zahlen endet; aber ebensowenig ist es eine Frage „des Skalenniveaus.“ Ist mit der Betrachtung einer Hypothese eine bestimmte Variable *X* festgelegt, so ist nur zu prüfen, ob für diese (und nur für diese) die entsprechenden Annahmen erfüllt sind, oder es sind entsprechende *Robustheitsüberlegungen* anzustellen.

Was ist somit von dem häufig formulierten Ratschlag zu halten, „für Ordinalskalen“ besser (wenn nicht ausschließlich) auf nichtparametrische Verfahren zurück-

<sup>2</sup>Auf einer Fehlinterpretation meßtheoretischer Konzepte, die hier nicht weiter diskutiert werden soll, beruht insbesondere der gelegentlich gehörte Einwand, daß der *t*-Test „für Ordinalskalen“ schon deshalb nicht zulässig sei, weil das Berechnen von Summen und Mittelwerten und damit auch das Berechnen der Teststatistik selbst in einem solchen Fall nicht „sinnvoll“ sei. Hier ist es in der Tat angebracht, von einer Frage zu sprechen, die mit Zahlen beginnt und mit Zahlen endet.

zugreifen? Im Sinne unserer obigen Unterscheidung sind hier zwei Aspekte zu unterscheiden. Das erste hier anklingende Problem ist rein innerstatistischer Art und bezieht sich lediglich auf das Testverfahren. Im Hinblick auf diesen Teilaspekt ist es, wie wir gesehen haben, unangebracht, eine entsprechende Empfehlung in besonderer Weise „für Ordinalskalen“ auszusprechen.

Betrachtet man den genannten Ratschlag genauer, so zeigt sich zudem, daß mit dem Wechsel des Verfahrens gewöhnlich auch ein Wechsel der getesteten Hypothesen selbst einhergeht, da nichtparametrische Tests gewöhnlich der Prüfung von Hypothesen dienen, deren Wahrheitswert unter ordinalen Transformationen invariant ist. In unserem Fall könnte dies also ein Vergleich von *Medianen* sein. Wären wir aber aus *inhaltlichen Gründen* etwa an einem Vergleich von Erwartungswerten, nicht jedoch an einem Vergleich von Medianen interessiert, so hilft uns der genannte Ratschlag nicht weiter.

In der genannten Empfehlung schwingt nun jedoch häufig noch eine zweite Erwägung mit, und diese betrifft die zu testenden Hypothesen selbst: Hypothesen der Art  $\mu_1 > \mu_2$ ,  $\mu'_1 > \mu'_2$  etc., so könnte man einwenden, mögen zwar semantisch bedeutsam sein; dennoch könnten sie in einem strengeren empirisch-inhaltlichen Sinn „für Ordinalskalen“ *nicht* „bedeutsam“ sein. Aus einer solchen Sicht wäre dann der Übergang zu einer invarianten Hypothese, etwa einem Vergleich von Medianen, im genannten Beispiel auch aus inhaltlichen Gründen geboten.

Dem Problem der *inhaltlichen* Angemessenheit statistischer Hypothesen wollen wir uns nun zuwenden. Tatsächlich ist die Frage, *welche* Hypothesen wir sinnvollerweise testen sollten, letztlich die primäre Frage; erst dann stellt sich das (eventuell nur schwer lösbare oder vielleicht im gegebenen Fall sogar unlösbare) Problem, *wie* dies in der jeweiligen Situation geschehen könnte. Mit letzterem wollen wir uns im folgenden nicht mehr befassen.

### 3 Zum Problem der „inhaltlichen Bedeutsamkeit“ von statistischen Hypothesen

Viele Autoren, welche skalenniveaubezogene Bedeutsamkeitskriterien diskutieren, haben in der Tat das Problem der „inhaltlichen Relevanz“ statistischer Hypothesen im Sinn, wie etwa Luce et al. (1990, Kap. 22.6), welche sich tendenziell der durch Adams et al. (1965, S. 124) ausgesprochenen Befürchtung anschließen, daß „the practice of ignoring scale type in statistical tests could lead to the formulation of *empirically meaningless hypotheses*“ (unsere Hervorhebung). An anderer Stelle sprechen Adams et al. (1965) auch von der möglichen *empirical nonsignificance* semantisch bedeutsamer Hypothesen. Wir teilen die Ansicht, daß die empirisch-inhaltliche Relevanz konventionell betrachteter statistischer Hypothesen in nicht wenigen Anwendungszusammenhängen tatsächlich fraglich oder zumindest unklar ist. Doch was ist von der Diagnose (und entsprechenden Therapieansätzen) zu halten, dies sei – wenn auch vielleicht nicht ausschließlich – eine Folge der Vernachlässigung von Skalenniveaubetrachtungen?

Der sachliche Kern entsprechender meßtheoretischer Invarianzbetrachtungen ist in etwa der folgende: Führt man die zugrundeliegenden Skalen(familien) in der im letzten Kapitel angesprochenen (und z.B. von Adams et al., 1965, und Luce et al.,

1990, zugrunde gelegten) repräsentationstheoretischen Weise ein, so läßt sich der qualitative Gehalt von statistischen Hypothesen, deren Wahrheitswert unter den betreffenden zulässigen Transformationen invariant bleibt – diese wollen wir in der Folge kurz als invariante Hypothesen bezeichnen –, häufig in einem gewissen Sinne in termini von Wahrscheinlichkeiten „auf“ der zugrunde gelegten qualitativen Struktur charakterisieren; unter gewissen Voraussetzungen ist Invarianz tatsächlich auch eine notwendige Bedingung hierfür. Ein Beispiel: Beruht  $X$  auf einer Ratingskala, so können wir die damit verknüpfte (schwache) qualitative Ordnung  $\succsim$  auf der Menge  $A$  der zu beurteilenden Objekte betrachten (wobei  $a \succsim b$  bedeutet, daß  $a$  mindestens so hoch eingestuft wird wie  $b$ ). Die Menge aller Homomorphismen (d.h. hier: der ordnungserhaltenden Abbildungen) der qualitativen Struktur  $\langle A, \succsim \rangle$  in die numerische Struktur  $\langle \mathbb{R}, \geq \rangle$  bildet dann eine Ordinalskala. Die invariante Hypothese, daß der Median von  $X_1$  größer sei als der von  $X_2$  (kurz:  $Med(X_1) > Med(X_2)$ ), läßt sich in der Tat unter ausschließlicher Bezugnahme auf Wahrscheinlichkeiten „auf“  $\langle A, \succsim \rangle$ . d.h. ohne Bezug auf Skalen, etwas umständlich wie folgt ausdrücken: †

Sind  $a_i$  ( $i = 1, 2$ ) Objekte mit der Eigenschaft, daß in Population  $i$  die Wahrscheinlichkeit, ein Objekt  $b$  mit der Eigenschaft  $a_i \succsim b$  zu „ziehen“, gerade 0.5 beträgt,<sup>3</sup> so gilt  $a_1 \succ a_2$ .

Wird das genannte Invarianzkriterium verletzt (wie z.B. unter Umständen im Falle eines Vergleichs von Erwartungswerten), so ist eine reichhaltigere qualitative Struktur zugrunde zu legen. Man könnte hier von *qualitativ-immanenter Bedeutsamkeit* statistischer Hypothesen *relativ* zu einer spezifischen qualitativen Struktur sprechen; da diese erweiterbar ist, handelt es sich hier jedoch nicht um ein absolutes Konzept „qualitativ-immanenter Bedeutsamkeit“. Im statistischen Kontext stehen einschlägige meßtheoretisch-technische *meaningfulness*-Konzepte am ehesten zu diesem Aspekt des Sammelbegriffs *Bedeutsamkeit* in Beziehung. Eine allgemeine Präzisierung dieses durchaus subtilen Konzepts – und seines Bezugs zu Invarianzkriterien – erfordert jedoch eine tiefergehendere Analyse, als sie zur Zeit vorliegt.

Eine solche „immanente“ Prüfung des empirischen Gehalts einer statistischen Hypothese relativ zu einer (ausgewählten) qualitativen Bezugsstruktur darf jedoch nicht, und schon gar nicht in einem normativen Sinne, als ein Kriterium für „Bedeutsamkeit“ im Sinne von (aktueller oder potentieller) „empirisch-inhaltlicher Relevanz“ *per se* mißverstanden werden. Wie bereits im letzten Kapitel ausgeführt wurde, kann letztere nur mit Blick auf den gesamten jeweiligen empirisch-theoretischen Forschungskontext oder praktischen Anwendungszusammenhang beurteilt werden.

Wenden wir uns einer Variante dieses Problems zu, welches – vor allem mit Blick auf das Problem der Auswahl geeigneter statistischer Hypothesen – eine präzisere Diskussion erlaubt, nämlich die Relevanz (oder „Bedeutsamkeit“) statistischer Hypothesen *im Hinblick auf* spezifische übergeordnete Fragestellungen. Wieder wird es sich also um ein *relatives* Konzept von „Bedeutsamkeit“ handeln, diesmal jedoch um eines, das einen unmittelbaren Bezug zu Fragen der „inhaltlichen Relevanz“ im Forschungsalltag aufweist. Dabei werden wir die Frage im Auge behalten, ob (bzw. in welcher Weise) skalenniveaubezogene Invarianzkriterien *in diesem Zusammenhang* eine Rolle spielen könnten.

<sup>3</sup>Bzw. allgemeiner: daß diese Wahrscheinlichkeit  $\geq 0.5$  ist und daß es keine (im Sinne von  $\succsim$ ) „kleineren“ Objekte  $a_1$  bzw.  $a_2$  mit dieser Eigenschaft gibt ...

### 3.1 Bedeutsamkeit im Hinblick auf übergeordnete Fragestellungen

Betrachten wir ein für unsere Zwecke angemessen vereinfachtes Beispiel. Die Wirksamkeit zweier Substanzen soll verglichen werden, von denen angenommen wird, daß sie die Gesundheit von Pflanzen verbessern. Zur Beurteilung der Pflanzengesundheit soll ein Expertenrating herangezogen werden. Wir betrachten wieder entsprechende ZVn  $X_1$  und  $X_2$ , deren Werte durch das Rating definiert sind. Stellen wir uns einen typischen Vertreter skalenniveaubezogener Bedeutsamkeitskriterien vor und nehmen wir an, daß er mit der Begründung, daß es sich um eine „ordinalskalierte Variable“ handle,<sup>4</sup> insbesondere den Vergleich von Erwartungswerten als unzulässig ausschließt; eventuell gibt er auch die Empfehlung, invariante Hypothesen, wie etwa einen Medianvergleich, zu betrachten.

Es fällt auf, daß dieser Ratschlag nicht auf eine mögliche übergeordnete Fragestellung Bezug nimmt. Wir werden jedoch sehen, daß von dieser Fragestellung und dem jeweiligen *theoretischen Kontext* entscheidend abhängt, welche Hypothesen sich als angemessen erweisen. Betrachten wir aus der Vielzahl der Möglichkeiten einige instruktive Fälle, in welchen Vergleiche der zentralen Tendenz eine Rolle spielen könnten.

**Fall 1:** Nehmen wir an, es gehe lediglich darum zu belegen, daß die beiden Substanzen insofern einen unterschiedlichen Effekt haben, daß die Verteilungen von  $X_1$  und  $X_2$  *verschieden* sind. Sämtliche auf irgendeine Art der Verschiedenheit bezogene Hypothesen, insbesondere also auch die betrachteten Hypothesen in termini von Erwartungswerten und Medianen, formulieren hierfür *hinreichende* Bedingungen und können insofern als *mit Bezug auf unsere Zielsetzung* sinnvolle Hypothesen betrachtet werden.

**Fall 2:** Interessanter ist die Frage, ob die erste Substanz, welcher die ZV  $X_1$  entspricht, im Vergleich zu der zweiten, mit  $X_2$  verknüpften Substanz einen „günstigeren Effekt“ auf die Pflanzengesundheit hat. Diese noch recht unspezifische Fragestellung erlaubt unterschiedliche Präzisierungen, die getrennt zu behandeln wären. Betrachten wir hier das praktische Problem, ob der Einsatz der ersten Substanz zu einem höheren Verkaufserlös führen würde als der der zweiten; diese Frage soll, wenn möglich, auf der Grundlage der durch  $X$  gegebenen Daten entschieden werden. In diesem Fall (wie in anderen Fällen) bezieht sich die Formulierung „günstigerer Effekt“ bei genauem Hinsehen „eigentlich“ auf eine zweite Variable  $Y$  – hier: der Verkaufserlös –, welche in vielen psychologischen Beispielen eine (unter Umständen nur schwer spezifizierbare) latente Variable sein kann. Im hier vorliegenden Fall liegt es nahe, als „eigentlich“ interessante Hypothese im Sinne unserer Fragestellung die Hypothese

$$E(Y_1) > E(Y_2) \quad (1)$$

<sup>4</sup>Anhänger Stevensscher Bedeutsamkeitskriterien, welche hier einen abweichenden Standpunkt einnehmen, mögen die nun folgende Diskussion sinngemäß auf ein in ihrem Sinne „ordinalskaliertes“ Beispiel übertragen. (Man stelle sich z.B. vor, daß ein befragter Experte Paarvergleiche vornimmt, wobei wir annehmen wollen, daß dies zu einer Ordnung  $\succsim$  führe;  $X$  sei dann eine beliebig festgelegte Einzelskala aus der entsprechenden Ordinalskala, z.B. der Rang.)

zu betrachten (auf eine Begründung sei hier verzichtet). Damit dies überhaupt in Beziehung zu Hypothesen über  $X_1$  und  $X_2$  gesetzt werden kann, sind *theoretische Zusatzannahmen* erforderlich, wie etwa die Annahme, die wir hier zu Illustrationszwecken zugrunde legen wollen, daß

$$Y = \tau \circ X \quad (2)$$

für eine ordinale Transformation  $\tau$  (d.h. höhere Ratings entsprechen einem höherem zu erzielenden Verkaufserlös). Unterscheiden wir nun, je nach weiteren theoretischen Zusatzannahmen, drei Unterfälle:

**Fall 2a:** Nehmen wir an, es gibt Grund zu der Annahme, daß (i) die Verteilungen von  $Y_1$  und  $Y_2$  *symmetrisch* sind und damit  $E(Y_i) = Md(Y_i)$  ( $i = 1, 2$ ), oder daß (ii) eine der Verteilungen die andere in dem Sinne *stochastisch dominiert*, daß sich die beiden entsprechenden kumulativen Verteilungsfunktionen nirgendwo kreuzen oder berühren (vgl. z.B. Niederée, 1994, Abschnitt 5.5). Unter jeder dieser Voraussetzungen ist Hypothese (1) mathematisch äquivalent zu  $Md(Y_1) > Md(Y_2)$ , und damit wegen (2) auch äquivalent zur Hypothese

$$Md(X_1) > Md(X_2). \quad (3)$$

Unter diesen (oder eventuell anderen, ähnlich strengen) Voraussetzungen erweist sich ein Vergleich von Medianen in der Tat als inhaltlich angemessen für unsere Fragestellung. [Ein entsprechendes Resultat hätten wir unter Annahme von (2) natürlich insbesondere auch dann erhalten, wenn wir anstelle von (1) sofort eine entsprechende Hypothese in termini von Medianen zugrunde gelegt hätten, was in manchen Situationen angemessen sein könnte, in unserem Beispiel jedoch sicher nicht.]

**Fall 2b:** In der Regel werden wir jedoch nicht ohne weiteres davon ausgehen können, daß Voraussetzungen der in Fall 2a genannten Art erfüllt sind.<sup>5</sup> Nehmen wir an, wir hätten statt dessen aufgrund von Voruntersuchungen hinlänglich gesicherte Hypothesen darüber, für *welche* Transformation  $\tau$  Gleichung (2) gilt. In einem solchen theoretischen Kontext wäre dann natürlich die betreffende statistische Hypothese

$$\mu'_1 := E(\tau \circ X_1) > \mu'_2 := E(\tau \circ X_2). \quad (4)$$

zu (1) äquivalent und somit eine mit Blick auf unsere Fragestellung angemessene auf  $X$  bezogene Hypothese, während sich die medianbezogene Hypothese (3) nicht mehr in stringenter Weise in Beziehung zum genannten Problem setzen ließe (und somit zu einer schmerzlichen Fehlentscheidung führen könnte).

**Fall 2c:** Dies ist der Fall, in dem wir weder die in Fall 2a noch die in Fall 2b vorausgesetzten (oder vergleichbar starke) Zusatzinformationen zur Verfügung haben.

<sup>5</sup>Nehmen wir an, sie seien auch nicht in „hinlänglicher Annäherung“ erfüllt. Wäre letzteres der Fall, so könnte man immerhin noch „einigermaßen“ – aber nicht mehr vollständig – sicher sein, daß die Wahrheitswerte der genannten auf den Median bzw. auf Erwartungswerte bezogenen Hypothesen übereinstimmen. Derartige pragmatische „Robustheitsüberlegungen“ auf der Ebene der Hypothesen selbst seien hier ausgeklammert, da sie für das grundsätzliche Verständnis des hier betrachteten konzeptuellen Problems nicht wesentlich sind.

Dann wissen wir zwar, daß – unter Voraussetzung der Monotonieannahme (2) – für gewisse uns unbekannt Transformationen  $\tau$  Hypothese (4) inhaltlich angemessen für unser Problem wäre, doch nicht für welche. Falls noch nicht einmal (2) oder eine ausreichend starke Abschwächung hiervon vorausgesetzt werden kann, läßt sich noch nicht einmal sagen, ob überhaupt eine auf  $X$  bezogene Hypothese der hier betrachteten Form für unsere Fragestellung angemessen ist. Im ungünstigsten Fall könnte sich sogar herausstellen, daß die betrachtete Variable  $X$  (in der durch die zugrundeliegende Operationalisierung festgelegten Form) für unser Problem keine brauchbaren Informationen liefert und durch eine andere Variable ersetzt werden muß.

Wir sehen, daß sich die *positive Empfehlung*, die medianbezogene Hypothese (3) zu betrachten, in der im Fall 2 betrachteten Situation *nur* unter ganz bestimmten Zusatzannahmen angemessen ist; diese Angemessenheit, soll sie nicht im Ungefähren verbleiben, ist *ausschließlich* eine Frage des hinreichend präzise zu modellierenden Zusammenhangs zum übergeordneten Problem. Gleiches würde entsprechend für andere unter ordinalen Transformationen invariante Hypothesen gelten.

Als ebenso unangemessen im Hinblick auf den hier diskutierten Aspekt des Bedeutsamkeitsproblems erweist sich in unseren Beispielen aber auch der rein *restriktive* Gebrauch entsprechender skalenniveaubezogener Kriterien, welcher auf positive Empfehlungen verzichtet, aber gewisse Hypothesen – in unserem Fall also insbesondere Vergleiche von Erwartungswerten – ohne Berücksichtigung möglicher übergeordneter Fragestellungen für unzulässig erklärt (siehe Fall 1 und Fall 2b; selbst im Fall 2c könnte man immerhin noch von einer *potentiellen* Bedeutsamkeit solcher Hypothesen sprechen).

Andererseits wird aber auch die Position, daß Statistik mit Zahlen beginne und mit Zahlen ende, dem hier behandelten Aspekt des Bedeutsamkeitsproblems nicht gerecht. Dies zeigt deutlich Fall 2c: In der Tat ist in solchen Situationen die *Warnung* am Platze, daß ohne weitere Begründung eine willkürlich herausgegriffene Hypothese der Form (4) – insbesondere also die traditionell bevorzugte Hypothese  $\mu_1 > \mu_2$  – durchaus für das betrachtete Problem unangemessen sein kann. Allerdings erstreckt sich diese Warnung in diesem Beispiel insbesondere eben auch auf die medianbezogene Hypothese (3). In *diesem* Sinne liegt in derartigen Fällen also tatsächlich ein „Bedeutsamkeitsproblem“ (*im Hinblick auf* die jeweils betrachtete übergeordnete Fragestellung) vor. Es hat aber offensichtlich nichts gemein mit den üblichen in der Literatur anzutreffenden Skalenniveaubetrachtungen.<sup>6</sup> In solchen Fällen der relativen Unwissenheit, welche in der Psychologie die Regel sind (häufig ist nicht einmal die übergeordnete Frage hinlänglich präzise faßbar) kann strenggenommen nur eine – leider in vielen Situationen kaum realisierbare – Empfehlung ausgesprochen werden: Soll die Angemessenheit von statistischen Hypothesen im Hinblick auf eine übergeordnete substantielle Fragestellung stringent beurteilt werden, so ist zunächst diese Fragestellung selbst zu klären und sodann die Beziehung zwischen den betrachteten

<sup>6</sup>Insbesondere sehen Diskussionen darüber, ob Ratingskalen nicht unter Umständen doch „intervallskaliert“ seien, üblicherweise von einer übergeordneten Fragestellung ab und nehmen damit (selbst dort, wo das irreführende Konzept des „wahren Skalenniveaus“ vermieden wird) eine „immanente“ Perspektive ein, welche dem hier betrachteten Problem nicht gerecht werden kann.

Variablen und dieser Fragestellung herauszuarbeiten. Solange dies nicht in befriedigender Weise möglich ist, sind wir auf ein „vorsichtig interpretierendes“ Vorgehen angewiesen, wobei skalenniveaubezogene Kriterien bestenfalls die Rolle behutsam zu verwendender Heuristiken spielen können, welche – wie wir in den obigen Beispielen gesehen haben – wie jede Heuristik auch in die Irre führen können.

Der Bezug auf eine übergeordnete Fragestellung kann natürlich verschiedene Formen annehmen. Idealerweise wird er aber – wie in unseren obigen Beispielen – ein deduktiver sein (gegebenenfalls relativ zu gewissen theoretischen Vorannahmen, wie z.B. der Monotonieannahme (2)). Von besonderem Interesse sind dabei Situationen, in denen in einem geeigneten theoretischen Rahmen ein inhaltlich motiviertes *stochastisches Modell* vorgeschlagen worden ist.<sup>7</sup> Zum Zwecke einer empirischen Überprüfung eines solchen Modells – dieses Ziel definiert in diesem Fall die übergeordnete Fragestellung – können wir nun statistische Hypothesen untersuchen, welche Folgerungen aus diesem Modell sind und deren Gültigkeit daher eine *notwendige* Bedingung für die Gültigkeit des Modells ist. Gelegentlich finden sich auch in Teilbereichen der Psychologie, wie z.B. der Reaktionszeitforschung, Beispiele für einen solchen klaren theoretischen Zusammenhang. So implizieren z.B. im Rahmen von Sternbergs (1969) *Methode additiver Faktoren* (vgl. Funke, in diesem Band) entsprechende Modelle (bei Zugrundelegung geeigneter struktureller Annahmen wie etwa der Unabhängigkeit der Dauer gewisser Teilprozesse) häufig spezifische erwartungswertbezogene statistische Hypothesen. Hierbei spielen separate Skalenniveaubetrachtungen keine Rolle; allein entscheidend ist jeweils der logische Bezug zwischen Hypothese und Modell. In anderen derartigen Situationen könnten es natürlich gerade „ordinale“ Hypothesen sein, welche die gewünschten Beziehungen zum betrachteten Problem aufweisen.

### 3.2 Meßtheoretischer Epilog

Abschließend sei kurz der Frage nachgegangen, ob nicht – unbeschadet der vorhergehenden Ausführungen – skalenniveaubezogene Intuitionen im Bereich der angewandten Statistik dennoch einen präzisierbaren „wahren Kern“ besitzen könnten. Im Falle der oben angedeuteten „immanenten“ Perspektive hatten wir diese Möglichkeit bereits vorsichtig bejaht. Wie dort gilt aber auch für die nachfolgend grob skizzierten Gedanken, daß die gegenwärtige Meßtheorie zwar bereits wichtige Bausteine für entsprechende Analysen bereithält, daß deren Erstellung aber im großen und ganzen ein Desiderat zukünftiger meßtheoretischer Forschung bleibt.

Wie steht es mit den Beispielen des letzten Abschnitts? Scheiterte in Fall 2 das Kriterium der Invarianz unter ordinalen Transformationen als notwendige Bedingung für Bedeutsamkeit im Hinblick auf eine spezifizierbare Fragestellung nicht einfach daran, daß hier die Fragestellung (und damit die Variable *Y*) unberücksichtigt blieb?

<sup>7</sup>Die auf dieser Ebene der *inhaltlichen* stochastischen Modellbildung eventuell eingehenden (unter Umständen recht schwachen) Verteilungsannahmen sind – im Sinne unserer obigen Trennung der Ebene der statistischen Hypothesen von der Ebene des Testens derselben – im Prinzip von dem statistischen Modell zu unterscheiden, welches einem entsprechenden Test zugrunde liegt und unter Umständen stärkere und „inhaltlich ungerechtfertigte“ Verteilungsannahmen voraussetzen kann. Glücklicherweise lassen sich letztere in der Regel im Sinne entsprechender Robustheitserwägungen pragmatisch „abschwächen“, so daß der Test auch in etlichen Fällen anwendbar ist, welche nur die eventuell schwächeren Voraussetzungen des er genannten inhaltlich motivierten Modells erfüllen.

Und ist es wirklich nur „Zufall“, daß für viele auf Reaktionszeit-Modelle bezogene Fragestellungen die betrachteten inhaltlichen Hypothesen (wie der Vergleich von Reaktionszeit-Erwartungswerten) invariant sind unter zulässigen Transformationen der mit der extensiven Messung der Zeitdauer einhergehenden Verhältnisskala?

Eine Klärung entsprechender Intuitionen setzt offensichtlich die explizite Einbeziehung *wesentlicher Aspekte der Fragestellung* voraus und ist nur in einem angemessen reichhaltigen (meta-)theoretischen Rahmen möglich (etwa auf der Grundlage des geeignet zu erweiternden Repräsentationsansatzes der axiomatischen Meßtheorie; vgl. das letzte Kapitel, sowie Niederée & Narens, in diesem Band). Es ist durchaus denkbar, daß ein entsprechendes *metatheoretisches* Theorem in mathematisch strenger Weise zeigen könnte, daß für Situationen und Hypothesen, welche gewisse Voraussetzungen erfüllen, die betreffenden Hypothesen sich als invariant unter gewissen Skalentransformationen erweisen (nicht: sein *sollen!*). Die Tatsache, daß ein solches Metatheorem (noch) nicht vorliegt, schafft aber keine grundsätzlichen Schwierigkeiten für die Diskussion der inhaltlichen Bedeutsamkeit gegebener statistischer Hypothesen im Hinblick auf spezifische Fragestellungen: Letztlich entscheidend für die Vermeidung des Fehlers, eine für eine Fragestellung unangemessene Hypothese auszuwählen, (Hand, 1994, spricht von einem *error of the third kind*) ist, wie wir gesehen haben, „lediglich“ ein ausreichendes Verständnis der entsprechenden theoretischen Bezüge.

Bedeutsamer als derartige metatheoretische Erwägungen wäre daher der Versuch, die qualitativ-strukturelle Denkweise der axiomatischen Meßtheorie und die im letzten Kapitel beschriebene substanzwissenschaftlich-theoretische Interpretation von Invarianzkonzepten auch in stochastischen Kontexten fruchtbar zu machen (so ließe sich z.B. vermutlich ein tieferes Verständnis des im Fall 2 ohne weitere Diskussion zugrunde gelegten Kriteriums (1) gewinnen). In diesem Fall entsprechen Invarianzbeobachtungen spezifischen empirischen Hypothesen und sind damit *Teil* der jeweiligen inhaltlichen Theoriebildung.

Selbst im Hinblick auf Probleme der angewandten Statistik sollte man daher das Konzept des Skalenniveaus keineswegs generell als obsolet verwerfen. Jedoch ist – bzw. wäre – ein theoretisch wohlbegründeter und sinnvoller Bezug hierauf nur im Rahmen eines entsprechenden präzisen theoretischen Ansatzes möglich, wie er etwa in einem außerstatistischen Zusammenhang im letzten Kapitel angedeutet wurde. Herkömmliche, der Stevensschen Tradition verhaftete Skalenniveaudiskussionen und daran geknüpfte Invarianzkriterien – welche sich insbesondere in der Psychologie großer Popularität erfreuen – sind daher in aller Regel mit größter Zurückhaltung zu betrachten. Obwohl ihren Protagonisten der inhaltliche Bezug quantitativer Aussagen zu Recht am Herzen liegt, verdunkeln derartige Diskussionen häufig das eigentliche Problem, nämlich das Fehlen einer inhaltlichen Theoriebildung, welche die Untersuchung spezifischer statistischer Hypothesen „schlüssig“ rechtfertigen könnte.

#### 4 Weiterführende Literatur

Zur Einstimmung in die geschilderte Kontroverse sei die Lektüre klassischer Arbeiten wie etwa Stevens (1946, 1951, 1959) und Lord (1953) empfohlen (vgl. auch den Sammelband von Haber, Runyon & Badia, 1970). Meßtheoretische Präzisierung

gen und Weiterführungen von Stevens' Position finden sich u.a. bei Adams et al. (1965), Pfanzagl (1971), Klein (1984), und Luce et al. (1990, Kap. 22). In Luce et al. (1990, S. 294) finden sich auch Literaturhinweise auf das überaus umfangreiche Schrifttum zur nach wie vor nicht abgeschlossenen statistischen Bedeutsamkeitskontroverse. Trotz gewisser erzielter konzeptueller Klärungen scheint es uns jedoch an einer hinlänglich differenzierten und klaren Darstellung der verschiedenen Aspekte dieses Problems bisher zu fehlen. Die in diesem Kapitel vertretene Position ist im Einklang mit der in Niederée (1994, Abschnitt 5.5) vertretenen Auffassung (das dortige Konzept der „*type-3 meaningfulness*“ entspricht dem hier verwendeten Begriff der Bedeutsamkeit im Hinblick auf eine Fragestellung).

## Literaturverzeichnis

- Adams, E. W., Fagot, R. F. & Robinson, R. (1965). A theory of appropriate statistics. *Psychometrika*, 30, 99–127.
- Burke, C. J. (1953). Additive scales and statistics. *Psychological Review*, 60, 73–75.
- Chiang, C.-Y. (1995). Invariant functions on measurement structures. *Journal of Mathematical Psychology*, 39, 112–116.
- Haber, A., Runyon, R. P. & Badia, P. (Eds.) (1970). *Readings in statistics*. Reading: Addison-Wesley.
- Hand, D. J. (1994). Deconstructing statistical questions. *Journal of the Royal Statistical Society, A* 157, 317–356.
- Klein, I. (1984). *Das Problem der Auswahl geeigneter Maßzahlen in der deskriptiven Statistik*. Würzburg: Physica.
- Lord, F. M. (1953). On the statistical treatment of football numbers. *American Psychologist*, 8, 750–751.
- Luce, R. D., Krantz, D. H., Suppes, P. & Tversky, A. (1990). *Foundations of measurement. Vol. 3.: Representation, axiomatization, and invariance*. San Diego: Academic Press.
- Luce, R. D. & Krumhansl, C. L. (1988). Measurement, scaling, and psychophysics. In R. C. Atkinson, R. J. Herrnstein, G. Lindzey & R. D. Luce (Eds.), *Stevens' handbook of experimental psychology. Vol. I: Perception and motivation* (pp. 3–73). New York: Wiley.
- Niederée, R. (1994). There is more to measurement than just measurement: Measurement theory, symmetry, and substantive theorizing. *Journal of Mathematical Psychology*, 38, 527–593.
- Pfanzagl, J. (1971). *Theory of measurement* (2. rev. Aufl.). Würzburg: Physica.
- Sternberg, S. (1969). The discovery of processing stages: Extensions of Donder's method. *Acta Psychologica*, 30, 276–315.
- Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, 103, 677–680.
- Stevens, S. S. (1951). Mathematics, measurement, and psychophysics. In S. S. Stevens (Ed.), *Handbook of experimental psychology* (S. 1–49). New York: Wiley.
- Stevens, S. S. (1959). Mathematics, measurement, and utility. In C. W. Churchman & P. Ratoosh (Eds.), *Measurement: Definitions and theories* (S. 18–63). New York: Wiley.
- Suppes, P. & Zinnes, J. (1963). Basic measurement theory. In R. D. Luce, R. R. Bush & E. Galanter (Eds.), *Handbook of mathematical psychology* (Vol. 1, S. 1–76). New York: Wiley.