Populationsverteilungen von Merkmalen und Geltungsbereiche individuenbezogener Aussagen als Gegenstand der Inferenzstatistik in psychologischen Untersuchungen.

Albrecht Iseler

Mathematisch formalisierte methodologische Diskussionen geraten leicht in den Verdacht folgenloser Abstraktheit. Um diesem Eindruck entgegenzutreten, will ich einige konkrete Anregungen für die Methodik der Überprüfung psychologischer Hypothesen zunächst ohne ihren mathematischen Hintergrund darstellen. Im wesentlichen geht es dabei um den Verzicht auf repräsentative Stichproben, der sich ja in der Forschungspraxis in vielen Bereichen durchgesetzt hat. Auch wenn man von einem Methodiker vielleicht eher erwartet, daß er angesichts dieser Praxis mit erhobenem Zeigefinger die Anwendung wohlbekannter Stichprobentechniken einfordert, will ich in diesem Referat die Auffassung begründen, daß zur Überprüfung vieler Hypothesen die üblicherweise verwendeten *beliebig verzerrten* Stichproben meist nicht weniger informativ sind als repräsentative und daß *wohlüberlegt verzerrte* Stichproben sogar informativer als repräsentative Stichproben sein können. Die Grundidee kann auch ohne mathematische Formalisierung begründet werden, und das will ich in einem ersten Abschnitt versuchen; allerdings wird auch aufzuzeigen sein, warum es zusätzlich zu dem intuitiven Zugang einer mathematischen Formaliserung bedarf. Diese folgt dann im zweiten Abschnitt, und im dritten sollen methodologische Schlußfolgerungen gezogen werden.

1. Die Grundidee von Hypothesentests an wohlüberlegt verzerrten Stichproben

Für eine erste Demonstration des Aussagegehalts von Ergebnissen aus wohlüberlegt verzerrten Stichproben ist etwa die Kerzenaufgabe von Duncker (1935) geeignet. Deren Grundhypothese läßt sich - übertragen auf heutige Standards der Versuchsplanung - dahingehend explizieren, daß in randomisierten unabhängigen Stichproben die mittlere Lösungszeit in der Durchführungsbedingung mit funktionaler Gebundenheit der Lösungmittel größer ist als ohne funktionale Gebundenheit. Unmittelbar einsichtig ist dabei, daß eine Eingrenzung der psychologischen Hypothese auf einen Geltungsbereich von Personen erforderlich ist, denen der Umgang mit Heftzwecken ebenso vertraut ist wie das Tropfen von Kerzen, die nicht senkrecht stehen. Offenkundig ist auch, daß es wenig Sinn machen würde, dieselben Personen das Kernzenproblem unter beiden Durchführungsbedingungen bearbeiten zu lassen, obschon die psychologische Hypothese eigentlich auf einem für alle Personen des Geltungsbereichs theoretisch angestellten intraindividuellen Vergleich basiert: Unter der einen Bedingung ist nach dieser Hypothese die Lösungszeit tendenziell kürzer, als sie es bei derselben Person (!) unter der anderen Bedingung wäre.

Unter einer vor allem in den 70-er Jahren von einigen Autoren vertretenen Sichtweise (Überblick z.B. bei Westmeyer, 1979) könnte man kritisieren, daß die Methode der Wahl für

die Überprüfung universeller Hypothesen nicht die statistische Prüfung von Aggregathypothesen, sondern die Untersuchung von Einzelfällen ist. Diese Sicht will ich in ihrem wissenschaftstheoretischen Kern nicht infragestellen - im Gegenteil. Auch für mich zählt der Anspruch von Fisher (1955), daß die Inferenzstatistik eine Lösung des Induktionsproblems darstellt, zu den folgenschwersten Mißverständnissen ihrer Leistungsfähigkeit, gerade auch für die Psychologie. Daher möchte ich klarstellen: Wenn es - anders als bei der Kerzenaufghabe - sinnvollerweise möglich ist, eine universelle Hypothese unter Beachtung allgemeiner Prinzipien des strengen und fairen Hypothesentestens an einem einzelnen Element eines intendierten Geltungsbereichs zu prüfen (z.B. an einer einzelnen Person oder an einer einzelnen Mutter-Kind-Dyade), dann ist das auch für mich die Methode der Wahl. Ich will darüber aber noch hinausgehen: Ich möchte die zugrundeliegenden wissenschaftstheoretischen Prinzipien auch für Fälle anwendbar machen, in denen - wie bei der Kerzenaufgabe - Einzelfalluntersuchungen an *einem* Element des Geltungsbereichs nicht sinnvoll sind, so daß eine Überprüfung kaum anders als über Aggregathypothesen erfolgen kann. Für solche Situationen soll aufgezeigt werden, daß aus bestimmten auf alle Individuen eines Geltungsbereichs bezogenen Hypothesen ein Universum von Aggregathypothesen herleitbar ist. Man kann auch von einer universellen Aggregathypothese sprechen, die mit der universellen individuenbezogenen Hypothese logisch äquivalent ist. Die einzelne in einer Untersuchung geprüfte Aggregathypothese ist im Verhältnis zu dieser universellen Aggregathypothese ein Einzelfall, und das bedeutet zweierlei: Trifft eine Vorhersage nicht ein, dann ist (unter geeigneten Vorbehalten hinsichtlich der üblichen Fehlerrisiken) die universelle Aggregathypothese falsifiziert, und damit auch die äquivalente universelle individuenbezogene Hypothese. Umgekehrt ist eine Bewährung *einer* solchen singulären Aggregathypothese nicht mehr und nicht weniger als *ein* Bestätigungsfall für eine universelle Hypothese.

Das methodologische Problem, um das es mir geht, läßt sich an der Frage festmachen, ob für eine solche Untersuchung nicht eigentlich eine "repräsentative" Stichprobe der zum Geltungsbereich der Hypothese gehörenden Personen erforderlich wäre. Es lassen sich leicht Beispiele anderer psychologischer Fragestellungen anführen, bei denen dieses Erfordernis zu bejahen ist. Wenn es etwa beim Vergleich zweier Varianten eines neu einzuführenden Verkehrsschilds darum geht, bei welcher von beiden der Bedeutungsgehalt besser erkannt wird, dann mag die Antwort auf diese Frage bei verschiedenen Personengruppen unterschiedlich sein. Aber da es in aller Regel nicht möglich ist, für jede dieser Gruppen das für sie günstigste Schild aufzustellen, kommt es darauf an, mit welcher Häufigkeit verschiedene Werte der (in geeigneter Weise operationalisierten) Erkennbarkeit in einer Population vorkommen, und damit bezieht sich die psychologische Fragestellung in der Tat auf die Populationsverteilung dieses Merkmals. Bei der Kerzenaufgabe ist das anders. Auf der Datenebene kann es zwar keine hypothesenwidrigen Einzelpersonen geben, weil es nicht

sinnvoll ist, von derselben Person Daten unter verschiedenen Bedingungen zu erheben. Die psychologische Hypothese läßt sich aber dahingehend explizieren, daß es für jede Person Verhaltensdispositionen unter beiden Bedingungen gibt, und auf dieser Ebene wäre eine einzelne Person, bei der diese Dispositionen sich nicht in der behaupteten Richtung unterscheiden, ein hypothesenwidriger Einzelfall, der allein genügen würde, den Anspruch zu Fall zu bringen, daß eine Hypothese für alle Elemente einer Menge intendierter Anwendungen gültig ist. Aus Gründen, die gleich erkennbar werden, will ich solche Personen, die eine zu untersuchende gesetzesartige Aussage durchbrechen, im folgenden als "Gesetzesbrecher" bezeichnen.

Dazu stellt sich dann allerdings die Frage, wie denn der Hypothese empirischer Gehalt verliehen werden kann, daß es solche auf der theoretischen Ebene von Dispositionen definierte Gesetzesbrecher nicht gibt. Offenbar könnte es weiterhelfen, wenn wir begründete Vermutungen haben, welche Personen noch am ehesten als Gesetzesbrecher in Frage kommen, wenn es überhaupt welche gibt: Dann könnten wir das Experiment mit lauter derartigen Personen durchführen. Relativ zu der Forderung nach repräsentativen Stichproben mit gleichen Selektionswahrscheinlichkeiten für alle Elemente des intendierten Geltungsbereichs einer gesetzesartigen Aussage würde es sich um eine verzerrte Stichprobe handeln; aber im Unterschied zu den üblicherweise erhobenen Personenmengen könnte man von einer wohlüberlegt verzerrten Stichprobe sprechen.

Als ich einmal in einer Vorlesung an diesem Punkt angekommen war, meinte ein Student, er könne sich z.B. gut vorstellen, daß zwanghaft ordentliche Menschen sich bei der Bedingung ohne funktionale Gebundenheit über die Unordnung ärgern und als erstes Zeit damit verlieren, Kerzen, Reißnägel und Streichhölzer in die entsprechenden Schachteln einzuordnen. Er wäre gespannt, was herauskäme, wenn man ein Experiment mit lauter solchen Menschen durchführen würde. Ob er mit seiner Vermutung recht hat, sei dahingestellt; jedenfalls hatte er offenbar verstanden, was ein ernsthafter Falsifikationsversuch ist.

Eine derartige Darstellung leuchtet auch Laien ohne Vorkenntnisse in Statistik ein als Demonstration der Position, daß wohlüberlegt verzerrte Stichproben eher als repräsentative geeignet sind, Gesetzesbrechern auf die Spur zu kommen, falls es sie gibt. Ihnen, meine Damen und Herren, kann ich zusätzlich noch die Anwendung einer Methode vorschlagen, die in der Arbeitsgruppe um Cronbach vor allem für den Bereich der Pädagogischen Psychologie entwickelt worden ist: Die multivariate Analyse von Aptitude-treatment-Wechselwirkungen (Cronbach & Snow, 1977). Natürlich macht es keinen prinzipiellen methodischen Unterschied, ob die statistische Abhängigkeit des Vorzeichens einer Mittelwertsdifferenz von Kovariaten überprüft wird, um ggf. aufgrund dieser Kovariaten eine differentielle Treatment-Indikation vorzunehmen (so bei Cronbach), oder ob eine solche Abhängigkeit untersucht

wird, um Verstöße gegen eine Gesetzesaussage zu suchen, die dasselbe Vorzeichen einer Mittelwerts-Differenz für alle zu einem Geltungsbereich gehörenden Individuen behauptet.

Ein ernsthafter Falsifikationsversuch der Hypothese, daß die Bedingung "funktionale Gebundenheit" bei allen zu einem intendierten Geltungsbereich gehörenden Personen zumindest auf der Dispositionsebene zu einer Verlängerung der Lösungszeit führt, könnte demnach folgendermaßen aussehen: In einer Vorstudie wird das Experiment an unabhängigen Stichproben durchgeführt, und von den Vpn werden außerdem Kovariaten erhoben, die nach theoretischen Vorüberlegungen noch am ehesten geeignet sind, Gesetzesbrecher zu identifizieren, wenn es sie gibt. Dann wird mit den für die Erforschung multivariater Aptitude-treatment-Wechselwirkungen entwickelten Methoden ermittelt, in welcher Region (oder auch in welchen Regionen) des von diesen Kovariaten aufgespannten Variablenraums sich noch am ehesten Gesetzesbrecher befinden, sofern es überhaupt welche gibt. Für die Hauptuntersuchung werden dann an einer größeren, mit der Vorstudie nicht überlappenden Menge von Personen Werte für diese Kovariaten erhoben, und daraus werden für ein Experiment mit randomisierten unabhängigen Stichproben diejenigen Personen ausgewählt, die nach den Ergebnissen der Vorstudie aufgrund ihrer Werte in den Kovariaten noch am ehesten als Gesetzesbrecher infragekommen. Die Vorhersage lautet dann: Wenn es - wie die psychologische Hypothese behauptet - keine Gesetzesbrecher gibt, dann wird sich auch in einer solchen wohlüberlegt verzerrten, durch Randomisierung in zwei Teilgruppen aufgeteilten Stichprobe kein hypothesenwidriger Unterschied der mittleren Lösungszeiten ergeben. An dieser Stelle mag auch die Motivation der Bezeichnung "Gesetzesbrecher" deutlich werden: Das ganze Verfahren kann anschaulich als "Rasterfahndung nach Gesetzesbrechern" bezeichnet werden.

Betrachten Sie, bitte, die Wiederholung der Formulierung "noch am ehesten" nicht als Mangel an verbaler Flexibilität. Ich spiele damit auf eine Explikation des Konzepts der strengen Hypothesenprüfung an, die sich - vorbereitet durch Arbeiten von Meehl (1967/1970, 1978) - etwa bei Westermann und Hager (1986) oder bei Erdfelder und Bredenkamp (1994) findet: Die Wahrscheinlichkeit eines Resultats, das als Bewährung einer zu prüfendenden psychologischen Hypothese gilt, soll möglichst gering sein, falls die Hypothese falsch ist. Die Verwendung des Wahrscheinlichkeitskonzepts in dieser Explikation mag problematisch sein; trotzdem bleibt - mit entsprechend vorsichtigerer Formulierung - die Forderung sinnvoll, Hypothesen an Personen zu prüfen, bei denen noch am ehesten mit einer Entdeckung von Gesetzesbrechern zu rechnen ist, falls es sie gibt.

Nach dem üblichen Verfahren einer eher intuitiven Herleitung der in einer Untersuchung zu prüfenden Aggregathypothesen mag die Vorhersage eines hypothesenkonformen Unterschieds der mittleren Lösungszeiten als eine aus der psychologischen Hypothese

abgeleitete Aggregathypothese akzeptabel sein. Daß eine Verlängerung von Problemlösungszeiten in der Regel ein Indikator für eine Aufgabenerschwerung ist, gehört zu den in der Problemlösungsforschung weitgehend akzeptierten fach-methodischen Grundannahmen. Man kann aber fragen, ob diese Interpretation von Lösungszeiten auch auf Intervallskalenniveau gerechtfertigt ist, ob also gleiche Verlängerung auch Indikator für gleiche Erschwerung ist. Da andrerseits die ordinale Interpretation unproblematisch ist, könnte man in Erwägung ziehen, die an wohlüberlegt verzerrten Stichproben zu prüfende statistische Hypothese auf der Ebene von Medianen zu formulieren und vorherzusagen, daß der Median der Lösungszeiten unter der Bedingung funktionaler Gebundenheit höher ist als ohne funktionale Gebundenheit.

Es gibt allerdings ein Problem bei dieser Art der Hypothesenableitung. In einem Artikel in der zu diesem Kongreß aus der Taufe gehobenen Zeitschrift "Methods of Psychological Research - Online" (Iseler 1996a) habe ich eine paradoxe Eigenschaft von Median-Hypothesen aufgezeigt, die sich auch auf das vorliegende Problem anwenden läßt. Dieses Median-Paradox besteht hier darin, daß die oben intuitiv entwickelte Aggregathypothese unzutreffend sein kann, obwohl für jede einzelne Person der Median der für diese Person charakteristischen Wahrscheinlichkeitsverteilung von Lösungszeiten unter der Bedingung "funktionale Gebundenheit" größer ist als ohne funktionale Gebundenheit. Gravierend wird dieses Problem vor allem dadurch, daß es ganz bestimmte Muster individuenbezogener Wahrscheinlichkeitsverteilungen von Lösungszeiten gibt, die dieses Median-Paradox produzieren können; denn dadurch kann es uns bei der "Rasterfahndung" passieren, daß unsere Kovariaten nicht wirkliche "Gesetzesbrecher" auswählen, sondern Personen, die aufgrund ihrer Verhaltensdispositionen zu einem Fall des Median-Paradoxes beitragen. Angesichts solcher Risiken ist es keineswegs nur eine Pflichtübung zwanghafter Formalisten, zu überprüfen, ob aus einer statistisch explizierten Hypothese, die sich auf alle Individuen eines Geltungsbereichs bezieht, auch eine entsprechende Eigenschaft des Aggregats folgt, die dann als Aggregathypothese empirisch prüfbar ist. Im folgenden Abschnitt 2 will ich deshalb zeigen, daß auch bei Hypothesentests anhand wohlüberlegt verzerrter Stichproben die zu prüfende Aggregathypothese aus einer entsprechenden individuenbezogenen Hypothese folgt, wenn wir Treatment-Effekte aufgrund von Erwartungswerten explizieren.

2. Herleitung von Erwartungswert-bezogenen Aggregathypothesen für Hypothesentests unter Verwendung von wohlüberlegt verzerrten Stichproben

Die Methode der Hypothesentests an wohlüberlegt verzerrten Stichproben ist keineswegs auf Mittelwertsdifferenzen beschränkt. Damit das Ganze nicht zu abstrakt wird, will ich mich aber für die folgende Demonstration auf diesen Hypothesentyp beschränken. Erst ganz zum Schluß wird zu klären sein, auf welcher Eigenschaft, die es auch bei vielen anderen

Hypothesenarten gibt, die methodologischen Schlußfolgerungen beruhen. Für die Behandlung von Erwartungswert-bezogenen Hypothesen stelle ich in einem ersten Unterabschnitt eine Formalisierung des zugrundeliegenden Stichprobenmodells dar, und in einem zweiten Unterabschnitt folgt dann die Hypothesen-Herleitung.

2.1 Das zugrundeliegende Stichprobenmodell

Zur Explikation des zugrundeliegenden Stichprobenmodell sollen zunächst einige Wahrscheinlichkeitsverteilungen der im folgenden als *Y* bezeichneten AV eingeführt werden, und wie es sich gehört, beginnt das mit einer Definition von Mengen. Die Menge *W* ist die Menge der möglichen Werte der AV; das sollen reellen Zahlen sein, und damit es keine meßtheoretischen Komplikationen bei der Bildung von Erwartungswerten gibt, wird Intervallskalenniveau dieser Meßwerte vorausgesetzt. Die Menge *D* (für 'domain') ist der Geltungsbereich der zu überprüfenden Hypothese. Das können - gleichsam prototypisch - Personen sein, aber beispielsweise auch Mutter-Kind-Dyaden oder "Personen in Situationen" - allgemein: Units. Schließlich besteht die Menge *C* (für 'conditions') aus den Bedingungen *a* und *b*, unter denen die units untersucht werden sollen. Zur Vereinfachung sei außerdem die Annahme eingeführt, daß auch die Mengen *W* und *D* endlich sind; das ermöglicht nämlich eine Summendarstellung an Stellen, wo sonst Lebesgue-Integrale erforderlich wären, mit denen ich vielleicht den einen oder anderen abhängen würde.

Nun zu den Wahrscheinlichkeitsverteilungen. Zunächst sei für jedes zum Geltungsbereich *D* gehörende unit *u*, jede zur Menge *C* gehörende Bedingung *c* und jedes zum Wertebereich *W* der AV gehörende *x* die Wahrscheinlichkeit $P_{uc}(Y=x)$ eingeführt: Es ist die für unit *u* unter Bedingung *c* anzusetzende Wahrscheinlichkeit, daß die AV *Y* den Wert *x* annimmt. Für alle Werte *x* der Menge *W* zusammengenommen, bilden diese für ein unit *u* und eine Bedingung *c* anzusetzenden Wahrscheinlichkeiten die Wahrscheinlichkeitsverteilung P_{uc}.

Für eine weitere, im folgenden als π_u bezeichnete Wahrscheinlichkeit ist wohl eine begriffliche Vorklärung angebracht. Es geht um den Begriff der "Zufallsauswahl" eines units, der ja häufig im Sinne eines klassischen Urnenmodells verstanden wird, bei dem für jede Kugel in der Urne die Auswahl-Wahrscheinlichkeit gleich sein muß. In einem allgemeineren Sinn kann man von Zufallsauswahl aber auch dann sprechen, wenn für jedes unit eine - möglicherweise unbekannte - Auswahlwahrscheinlichkeit besteht, die von unit zu unit verschieden sein kann, und das ist die ‘Selektionswahrscheinlichkeit’ π_u. Einzige Voraussetzung ist, daß diese Selektionswahrscheinlichkeiten die Axiome der Wahrscheinlichkeitstheorie erfüllen, und das bedeutet bei einer endlichen Menge *D* lediglich, daß es sich um nicht-negative Zahlen handeln muß, deren Summe *1* ist. Und genauso, wie die Komponenten x_1, x_2 usw. eines Vektors zusammen den Vektor *x* bilden, genauso bilden die

Selektionswahrscheinlichkeiten π_u der verschiedenen units die "Selektionsverteilung" π.

Wofür wir diesen Begriff brauchen: Wir betrachten die Daten einer Stichprobe von units, die unter einer Bedingung *c* untersucht wurden, als Ergebnis der mehrfachen Realisierung des folgenden Zufallsprozesses: Unter Zugrundelegung einer Selektionsverteilung π wird ein unit ausgewählt, und für dieses ausgewählte unit wird unter Bedingung *c* einen Wert der AV *Y* erhoben. Auf diesen Prozeß bezieht sich die nächste zu definierende Wahrscheinlichkeit: $P_{\pi c}(Y=x)$ (mit $c \in C$ und $x \in W$) sei die Wahrscheinlichkeit, daß bei diesem Prozeß die AV *Y* den Wert *x* annimmt. Faßt man die für eine Selektionsverteilung π und eine Bedingung *c* gültigen Wahrscheinlichkeiten zu einer Wahrscheinlichkeitsverteilung zusammen, dann ergibt sich die Verteilung $P_{\pi c}$. Sie kann als Mischverteilung betrachtet werden, da die einzelnen Wahrscheinlichkeiten sich nach der folgenden, üblicherweise als "Mischungsaxiom" bezeichneten Gleichung ergeben:

$$\forall \pi, c, x: P_{\pi c}(Y=x) = \sum_{u \in D} \pi_u \cdot P_{uc}(Y=x) \qquad (2.1)$$

Es wird also angenommen, daß die Wahrscheinlichkeiten $P_{\pi c}(Y=x)$ aus den entsprechenden, für die einzelnen units anzusetzenden Wahrscheinlichkeiten durch eine gewichtete Mittelung entstehen, bei der die Selektionswahrscheinlichkeiten die Gewichte sind.

Das Stichprobenmodell für die folgende Hypothesenherleitung besteht nun darin, die Daten der Gruppe, die unter Bedingung *c* untersucht wird, als mehrfache Realisierung der Mischverteilung $P_{\pi c}$ anzusehen. Damit ist *eine* Voraussetzung des mathematischen Stichprobenbegriffs erfüllt: Es muß sich um mehrere identisch verteilte Zufallsgrößen handeln. Eine weitere Voraussetzung - die Unabhängigkeit dieser Zufallsgrößen - ist dagegen verletzt, da wir kaum einer Person die Gelegenheit geben werden, zweimal am Experiment teilzunehmen. M.a.W., wir ziehen “ohne Zurücklegen”. Das Argument, mit dem diese Abweichung vom mathematischen Stichprobenmodell üblicherweise gerechtfertigt wird, läßt sich aber auf unseren verallgemeinerten Begriff der Zufallsauswahl übertragen: Beim Ziehen ohne Zurücklegen sind die Zufallsfehler geringer als beim Ziehen mit Zurücklegen. Wenn wir also unsere Zufallsfehler nach dem Modell ‘mit Zurücklegen’ berechnen, ist unser Erhebungsverfahren in Wirklichkeit besser, als in unserem Fehlermodell angenommen wird, und das wird üblicherweise in Kauf genommen.

Für unsere weiteren Überlegungen ist vor allem ein Ergebnis wichtig: Der von uns zugrundegelegte verallgemeinerte Begriff einer Zufallsauswahl läßt es zu, die der mathematischen Stichprobentheorie zugrundeliegende Annahme unabhhängiger, identisch verteilter Zufallsvariablen auch anders zu erfüllen als durch die lehrbuchmäßigen Zufallsstichproben mit gleichen Selektionswahrscheinlichkeiten für alle Elemente einer

Population. Wenn wir mit den aus derart definierten Stichproben gewonnenen Daten einen Signifikanztest durchführen, dann übernehmen die Mischverteilungen $P_{\pi c}$ die Rolle der "Populationsverteilung", auf die sich die von dem Signifikanztest geprüfte Hypothese bezieht. Ein *t*-Test für unabhängige Stichproben, die unter den Bedingungen *a* bzw. *b* untersucht wurden, würde also z.B. die Nullhypothese prüfen, daß die Erwartungswerte der Mischverteilungen $P_{\pi a}$ und $P_{\pi b}$ identisch sind. Daher wird im folgenden Abschnitt geklärt, wie diese Erwartungswerte mit den Erwartungswerten der individuellen Verteilungen P_{uc} zusammenhängen.

2.2 Formulierung und Herleitung von Hypothesen für die Erwartungswerte der AV

Die Hypothesen beziehen sich auf die Erwartungswerte der Verteilungen P_{uc} und $P_{\pi c}$ der AV, die als μ_{uc} bzw. $\mu_{\pi c}$ bezeichnet werden. Es ist also für jedes unit *u* und jede Bedingung *c*

$$\mu_{uc} := \sum_{x \in W} x \cdot P_{uc}(Y=x),$$

und diesen Erwartungswert kann man im Sinne der probabilistisch reformulierten klassischen Testtheorie (Novick, 1966, Zimmerman, 1976) als true score von unit *u* unter Bedingung *c* bezeichnen. Und bei den Mischverteilungen $P_{\pi c}$ sind die durch

$$\mu_{\pi c} := \sum_{x \in W} x \cdot P_{\pi c}(Y=x)$$

definierten Erwartungswerte $\mu_{\pi a}$ und $\mu_{\pi b}$ diejenigen Erwartungswerte, die nach den Ergebnissen des vorangehenden Unterabschnitts 2.1 vom *t*-Test verglichen werden.

Für die Hypothesenformulierung ist es hilfreich, die Differenzen dieser Erwartungswerte unter den Bedingungen *b* und *a* mit Namen zu versehen. Ich lehne mich dazu an eine bei Steyer, Gabler und Rukai (1995, 1996) vorgeschlagene Verallgemeinerung einer auf Neyman (1923) zurückgehenden Terminologie an. Für jedes unit *u* soll die Differenz $\mu_{ub} - \mu_{ua}$ als der "individuelle kausale Effekt für unit *u* von Bedingung *b* (relativ zu Bedingung *a*) auf den Erwartungswert der AV" bezeichnet werden. Die psychologische Hypothese sei dahingehend expliziert, daß dieser individuelle kausale Effekt für jedes zum Geltungsbereich *D* gehörende unit *u* größer als null ist. Explikationen sind nicht richtig oder falsch; es gibt aber Adäquatheitskriterien, die mehr oder weniger gut erfüllt sein können, und es sei angenommen, daß der Psychologe, um dessen Hypothese es geht, diese Explikation als adäquate Reformulierung dessen akzeptiert, was er meint, wenn er auf der Ebene von theoretischen Dispositionen für jedes zum Geltungsbereich *D* gehörende unit einen positiven Erffekt von Bedingung *b* behauptet.

Ähnlich sei definitionsgemäß die Differenz $\mu_{\pi b} - \mu_{\pi a}$ der Erwartungswerte der Mischverteilungen der "mittlere kausale Effekt von Bedingung *b* (relativ zu Bedingung *a*) auf den Erwartungswert der AV unter Selektionsverteilung π". Die Behauptung, daß dieser Effekt für eine bestimmte Selektionsverteilung π größer als null ist, kann man als "singuläre Aggregathypothese" bezeichnen und damit diese Hypothese von der "universellen Aggregathypothese" unterscheiden, die für alle Selektionsverteilungen in *D* einen positiven mittleren kausalen Effekt behauptet.

Von zentraler Bedeutung ist nun die folgende, für jedes Selektionsverteilung π gültige Gleichung:

$$\mu_{\pi b} - \mu_{\pi a} = \sum_{u \in D} \pi_u \cdot (\mu_{ub} - \mu_{ua}). \qquad (2.2)$$

Sie besagt, daß der mittlere kausale Effekt $\mu_{\pi b} - \mu_{\pi a}$ das gewichtete Mittel der individuellen kausalen Effekte $\mu_{ub} - \mu_{ua}$ ist, wobei die Selektionswahrscheinlichkeiten π_u die Gewichte bei dieser Mittelung sind. Diese Gleichung ist eine fast triviale Konsequenz aus dem Mischungsaxiom (2.1); aber so manches Problem wird ja einfach, wenn wir nur die für seine Lösung brauchbaren Trivialitäten ausschöpfen.

Zwei Konsequenzen der Gleichung (2.2) sind für uns vor allem wichtig. Zunächst ist das die Feststellung, daß ein positiver kausaler Effekt "aggregationsstabil" ist. Der Begriff der Aggregationsstabilität einer Eigenschaft, der an anderer Stelle (Iseler, 1996b) allgemeiner behandelt wird, bedeutet für die hiesige Situation: Sind alle individuellen kausalen Effekte positiv, dann gilt das auch für den mittleren kausalen Effekte bei jeder beliebigen Selektionsverteilung π. Damit folgt insbesondere auch für die bei einer wohlüberlegt verzerrten Stichproben zugrundegelegte Selektionsverteilung π, daß der entsprechende mittlere kausale Effekt $\mu_{\pi b} - \mu_{\pi a}$ positiv ist. Und damit können wir uns für die hier betrachtete Art psychologischer Fragestellungen endgültig von der Forderung nach Stichproben mit gleicher Selektionswahrscheinlichkeit für alle Elemente eines Geltungsbereichs verabschieden.

Aus Gleichung (2.2) läßt sich noch eine weitere Konsequenz herleiten: Gibt es in der Menge *D* Gesetzesbrecher (also Personen mit nicht-positivem individuellem kausalen Effekt), dann ist der mittlere kausale Effekt $\mu_{\pi b} - \mu_{\pi a}$ zumindest bei den Selektionsverteilungen nicht-positiv, bei denen die Selektionswahrscheinlichkeit π_u nur bei Gesetzesbrechern von *0* verschieden ist. Damit besteht aber eine Äquivalenz zwischen der universellen individuenbezogene Hypothese ("für jedes zur Menge *D* gehörende unit *u* ist der individuelle Effekt $\mu_{ub} - \mu_{ua}$ größer als *0*") und der folgenden universellen Aggregathypothese: "Für jede Selektionsverteilung π im Geltungsbereich *D* ist der entsprechende mittlere kausale Effekt

$\mu_{\pi b} - \mu_{\pi a}$ größer als *0*".

3. Methodologische Schlußfolgerungen

Die Überprüfung einer universellen Aggregathypothese kann nach den Prinzipien der Überprüfung von universellen Hypothesen an den daraus ableitbaren singulären Hypothesen erfolgen. Für die daraus abzuleitenden methodologischen Schlußfolgerungen muß ich mich im wesentlichen auf Thesen beschränken. Ich hoffe aber, daß diese Thesen aufgrund der vorangehenden Formalisierung meines Ansatzes verständlich genug sind, um sie zum Gegenstand einer Diskussion zu machen.

1. *Eine* bewährte singuläre Aggregathypothese ist nicht mehr und nicht weniger als *ein* singulärer Bestätigungsfall für eine universelle Hypothese (die universelle Aggregathypothese). Bakan (1955): An aggregate is not 'the general', but rather 'a particular'.

2. Gibt es im Geltungsbereich *D* nur wenige Gesetzesbrecher, kann deren Entdeckbarkeit an Grenzen der Teststärke scheitern. (Bei den meisten universellen Hypothesen kann es hypothesenwidrige Einzelfälle geben, die wegen Grenzen der Meßgenauigkeit nicht entdeckbar sind!)

3. Bewährt sich eine psychologische Hypothese auch bei Verwendung von Selektionsverteilungen, bei denen noch am ehesten eine Chance zur Entdeckung evtl. existierender "Gesetzesbrecher" besteht, dann ist eine induktive Gehaltserweiterung auf andere Selektionsverteilungen besser fundiert, als nach einer Hypothesenprüfung an beliebig verzerrten oder "repräsentativen" Stichproben.

4. Gerade weil die für die Praxis-Anwendung von Theorien erforderlichen induktiven Gehaltserweiterungen problematisch sind, sollten Wissenschaftler sich der damit verbundenen Verantwortung stellen und es nicht den Praktikern überlassen, sich damit die Finger schmutzig zu machen. Zur Entwicklung einer Theorie sollte daher auch die Benennung von Kovariaten gehören, die noch am ehesten zur Identifikation von "Gesetzesbrechern" (bzw. "nicht erfolgreichen Anwendungen") beitragen können, sofern solche existieren. Damit wird aus einer unstrukturierten "Menge intendierter Anwendungen" ein Raum intendierter Anwendungen, über den durch wohlüberlegt verzerrte Stichproben ein Netz von Überprüfungen gelegt werden kann.

5. Ein verallgemeinerter Populationsbegriff beruht auf Gewichten, mit denen eine psychologische Aussage verschiedene Untermengen ihres Geltungsbereichs gewichtet.

Gleichgewichtung aller Elemente des Geltungsbereichs ist dabei lediglich ein Spezialfall.

Ich möchte diesen Thesen zwei Erläuterungen hinzufügen. Zunächst ist wohl zur fünften These eine Klarstellung angebracht, und dazu möchte ich noch einmal auf das Beispiel des Vergleichs zweier Varianten eines Verkehrsschilds zurückkommen. Zur Präzisierung der zu untersuchenden Fragestellung müßte geklärt werden, ob die Angehörigen verschiedener Personengruppen möglicherweise mit unterschiedlichem Gewicht zu verrechnen sind, je nach dem, mit welcher Häufigkeit oder Wahrscheinlichkeit sie mit entsprechenden Verkehrssituationen konfrontiert sind. Auf dieser Gewichtung basiert dann gemäß These 5 die Definition einer Population, und dementsprechend ist auch diese Gewichtung in eine entsprechende Selektionsverteilung umzusetzen. Das ergäbe dann aber im Verhältnis zu der im Sinne von These 5 verstandenen Population keine wohlüberlegt verzerrte, sondern eine unverzerrte Stichprobe. Insbesondere wären gleiche Selektionswahrscheinlichkeiten für alle Elemente der Population hier nur in dem Spezialfall angebracht, in dem sich aus der Fragestellung eine Gleichgewichtung aller Elemente der Population ergibt. Natürlich bleibt es dabei - ebenso wie bei jeder wohlüberlegt verzerrten Stichprobe - unbenommen, Abweichungen einer realisierten Selektionsverteilung von einer intendierten durch geeignete Hochrechnungsverfahren zu kompensieren.

Die zweite angekündigte Erläuterung bezieht sich dann wieder auf die Überprüfung universeller Hypothesen. Die am Beispiel einer Mittelwertsdifferenz-Hypothese demonstrierten methodologischen Konsequenzen beruhen allein auf der Äquivalenz der universellen individuenbezogenen Hypothese und der universellen Aggregathypothese und sind damit auf andere Hypothesenarten übertragbar, in denen eine derartige Äquivalenz gilt. Welche Hypothesen das sind - das ist eine andere Geschichte, und die soll ein andermal erzählt werden.

Bakan, D. (1955). The general and the aggregate: A methodological distinction. Perceptual and Motor Skills, 5, 211-212.

Cronbach, L.J. & Snow, R.W. (1977). Aptitudes and instructional methods: A Handbook for research on interactions. New York: Irvington.

Duncker, K. (1935). Zur Psychologie des produktiven Denkens. Berlin: Springer.

Erdfelder, E. & Bredenkamp, J. (1994). Hypothesenprüfung. In Th. Herrman & W.H. Tack (Hrsg.), Methodologische Grundlagen der Psychologie (S. 604-648). (Enzyklopädie der Psychologie, Themenbereich B Methodologie und Methoden, Serie 1 Forschungsmethoden der Psychologie). Göttingen: Hogrefe.

Fisher, R.A. (1955). Statistical methods and scientific induction. Journal of the Royal Statistical Society, Series B, 17, 69-78

Iseler, A. (1996a). A paradoxical property of aggregate hypotheses referring to the order of medians. Methods of Psychological Research - Online, http://www.hsp.de/MPR/issue0/art2/article.html.

Iseler, A. (1996b). Aggregation Stability of Properties and the Derivation of Aggregate Hypotheses from Hypotheses Referring to Individuals. Unveröff. Manuskript, FU Berlin. Vgl. http://userpage.fu-berlin.de/~iseler/papers/index.html.

Meehl, P.E. (1967/1970). Theory-testing in psychology and physics: A methodological paradox. In Morrison & Henkel (1970, p. 252-266). (Nachdruck aus Philosophy of Science, 34, 1967, 103-115.)

Meehl, P.E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. Journal of Consulting and Clinical Psychology, 46, 806-834.

Morrison, D.E. & Henkel, R.E. (Eds.) (1970). The Significance Test Controversy. Chicago: Aldine Publishing Company.

Neyman, J. (1923/1990). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. Statistical Science, 5, 465-472.

Novick, M.R. (1966). The axioms and principal results of classical test theory. Jounal of Mathematical Psychology, 3, 1-18.

Steyer, R., Gabler, S. & Rucai, A.A. (1995). Confounding in Regression Models: Individual Causal Effects, Average Causal Effects, and Unconfoundedness. Unveröff. Manuskript, Universität Magdeburg.

Steyer, R., Gabler, S. & Rucai, A.A. (1996). Individual causal effects, average causal effects, and unconfoundedness in regression models. In F. Faulbaum & W. Bandilla (Eds.), SoftStat '95: Advances in Statistical Software 5, p. 203-210. Stuttgart: Lucius & Lucius.

Westermann, R. & Hager, W. (1986). Error probabilities in educational and psychological research. Journal of Educational Statistics, 11, 117-146.

Westmeyer, H. (1979). Wissenschaftstheoretische Grundlagen der Einzelfallanalyse. In F. Petermann & F.J. Hehl (Hrsg.), Einzelfallanalyse (S. 17-34). München: Urban und Schwarzenberg.

Zimmerman, D. W. (1976). Test theory with minimal assumptions. Educational and Psychological Measurement, 36, 85-96.

Autor:

Prof. Dr. Albrecht Iseler
Freie Universität Berlin,
Institut für Entwicklungspsychologie, Sozialpsychologie und Methoden der Psychologie,
Habelschwerdter Allee 45
14195 Berlin
E-Mail: iseler@zedat.fu-berlin.de