Stichwörter Empirische Forschungsmethoden aus dem Lexikon ILM

Punktschätzung

Unter P. versteht man die möglichst "gute" Schätzung eines Parameters der Grundgesamtheit aus einer gegebenen Stichprobe im Rahmen der Inferenzstatistik. Es geht hier also (*sehr* salopp gesprochen) um die Frage, welcher Parameter bei gegebener Stichprobe am "wahrscheinlichsten" ist. Da die P. aber stets mit Unsicherheit behaftet ist, sollte man sie i.a. durch eine Intervallschätzung ergänzen.

Intervallschätzung

Die I. ergänzt eine Punktschätzung durch die Angabe der Grenzen eines Intervalls - des Konfidenzintervalls -, das mit vorgegebener Wahrscheinlichkeit den wahren Parameter überdeckt.

Inferenzstatistik

Die I., auch induktive oder schließende Statistik genannt, beschäftigt sich mit der Frage, wie wir von einer Stichprobe, also einer Auswahl von Untersuchungseinheiten, auf die Grundgesamtheit zurückschließen können, aus der die Stichprobe stammt.

Im Rahmen des (noch) vorherrschenden "frequentistischen" Paradigmas geht man dabei von der Vorstellung aus, was geschehen würde, wenn man aus einer Grundgesamtheit viele Stichproben ziehen würde. Man kann daraus ableiten, wie groß bei einer gegebenen Annahme - der Null-Hypothese - die Wahrscheinlichkeit ist, einen gegebenen Stichprobenwert zu erhalten. Liegt der Stichprobenwert außerhalb des Bereichs, in dem er mit einer *vorher* festzusetzenden Wahrscheinlichkeit (der Sicherheitswahrscheinlichkeit) bei Gültigkeit der Nullhypothese liegen müßte - dieser Bereich wird Konfidenzintervall (oder Vertrauensintervall) genannt -, so wird die Nullhypothese verworfen. (Häufiger spricht man statt von der Sicherheitswahrscheinlichkeit von ihrer Gegenwahrscheinlichkeit, der sog. Irrtumswahrscheinlichkeit; dabei handelt es sich also um die Wahrscheinlichkeit, daß der Stichprobenwert auch bei Gültigkeit der Nullhypothese außerhalb des Konfidenzintervalls liegt).

Hypothese

Unter H. versteht man in der empirischen Sozialforschung eine anhand empirischer Daten zu prüfende Annahme. Im Rahmen der "quantitativen" (standardisierten) Sozialforschung meint man vor allem eine Annahme, die einem statistischen Test unterworfen werden kann. Diese Annahme richtet sich meistens darauf, dass zwischen zwei Merkmalen ein Zusammenhang, oder dass zwischen Gruppen ein Unterschied besteht. (Im Grunde sind das nur Varianten ein und desselben Sachverhalts.) Es sind aber auch zahlreiche weitere Hypothesen denkbar, so etwa solche über eine bestimmte Form des Zusammenhangs (etwa ein linearer, ein exponentieller, ein kurvilinearer), über Veränderungen usw.

Beim statistischen Test wird eine sog. *Nullhypothese* aufgestellt, die i.a. besagt, dass der postulierte Zusammenhang oder Unterschied *nicht* besteht. Es wird eine Teststatistik berechnet, die angibt, ob ein in den Daten beobachteter Zusammenhang oder Unterschied mit der Nullhypothese kompatibel ist. Überschreitet die Teststatistik einen bestimmten, vorab festzulegenden Wert, so wird die Nullhypothese verworfen und die eigentliche

Forschungshypothese, die *Alternativhypothese*, gilt als vorläufig nicht widerlegt (manche sagen auch: vorläufig akzeptiert). - Andere Nullhypothesen sind denkbar, z.B. dass ein Unterschied einen bestimmten Betrag nicht überschreitet, sie werden in der sozialwissenschaftlichen Forschungpraxis jedoch nur selten formuliert.

Im Rahmen der "qualitativen" Forschung geht man von einem weniger linearen Verhältnis zwischen H. und ihrer Prüfung aus. Insbesondere wird dort zumeist die Vorstellung abgelehnt, sich von vornherein mit einer bestimmten H. im Kopf dem Gegenstandsbereich anzunähern. Vielmehr sollen dort die H.en erst aus dem Datenmaterial entwickelt werden. Diese H.en müssen aber an neuem Material überprüft werden. "Neues Material" muss nicht unbedingt heißen "neue Fälle"; so versucht man bei der Prüfung von "Fallstrukturhypothesen" im Rahmen der Objektiven Hermeneutik, Annahmen über den Fall, die anhand eines Teils des Datenmaterials gewonnen wurden, anhand des übrigen Datenmaterials zu dem gleichen Fall zu bestätigen oder zu widerlegen.

Konfidenzintervall und Irrtumswahrscheinlichkeit

Die aus Stichproben geschätzten Parameter für eine Grundgesamtheit weichen notwendigerweise häufig von den wahren Parametern ab. Im Rahmen der Inferenzstatistik wird gezeigt, daß man aus der Stichprobe Intervalle schätzen kann, innerhalb derer der wahre Parameter mit einer vorgegebenen Wahrscheinlichkeit (der Überdeckungswahrscheinlichkeit) liegt. Diese Intervalle werden als K. (auch: Vertrauensbereich oder -intervall, in der älteren Literatur manchmal auch Mutungsbereich oder -intervall) bezeichnet. Andersherum wird häufig nach der Irrtumswahrscheinlichkeit gefragt, mit der der Parameter *nicht* in dem vorgegebenen Intervall liegt. Man erhält damit Aussagen der Art: Mit einer (Irrtums-)Wahrscheinlichkeit von p - häufig wählt man für p einen Wert von 0,05, manchmal aber auf 0,01 oder 0,001 - liegt der wahre Wert nicht innerhalb des Konfidenzintervalls mit der Untergrenze G_U und der Obergrenze G_O. Wie man das K. im einzelnen schätzt, hängt von Annahmen über die Verteilung der jeweiligen Variablen in der Grundgesamtheit ab.

Man unterscheidet zwischen zwei- und einseitigen K.n. Zweiseitige K. liegen im Regelfall symmetrisch um den geschätzen Parameter; die Wahrscheinlichkeit, daß der wahre Parameter über G_O liegt, soll gleich groß sein wie diejenige, daß er unter G_U liegt. Bei einem einseitigen K. wird eine der beiden Grenzen auf den Wert (+ oder -) "unendlich" festgelegt; man interessiert sich nur dafür, ob der Parameter über bzw. unter einem vorgegebenen Wert liegt. Ein Beispiel: Lautet die Forschungsfrage, ob Mädchen intelligenter *oder* weniger intelligent sind als Jungen, so muß man ein zweiseitiges Konfidenzintervall festlegen; will man aber die Hypothese prüfen, ob sie intelligenter sind, so wird man ein einseitiges Konfidenzintervall zugrundelegen.

Die Größe des Konfidenzintervalls hängt - bei vorgegebener Irrtumswahrscheinlichkeit - vor allem von zwei Faktoren ab: der Stichprobengröße und der Variabilität der Grundgesamtheit. Je größer - ceteris paribus - die Stichprobe, desto kleiner wird das Konfidenzintervall; ebenso wird das Konfidenzintervall kleiner bei kleinerer Variabilität der Grundgesamtheit.

Signifikanz

S. steht in der empirischen Sozialforschung im allgemeinen für *statistische* Signifikanz. Als signifikant in diesem Sinne gilt das Ergebnis eines Hypothesentests - des *Signifikanztests* -, wenn die Annahme berechtigt ist, daß ein theoretisch angenommener und in den Daten

vorgefundener Zusammenhang zwischen Merkmalen oder Unterschied zwischen Gruppen nicht alleine durch die Unschärfe erklärt werden kann, die mit der Stichprobenziehung verbunden ist. Die Berechtigung dieser Annahme kann nie mit Sicherheit erwiesen werden, sondern nur mit einer gewissen, *vorab* festzulegenden Wahrscheinlichkeit. Diese bezeichnet man als *Signifikanzniveau*. In den Sozialwissenschaften übliche Signifikanzniveaus sind 0,05, 0,01 und 0,001. Ein Signifikanzniveau von 0,05 festzulegen bedeutet, daß man ein Ergebnis als signifikant akzeptiert, welches rein zufällig nur in 5 Prozent aller Stichprobenziehungen auftreten würde.

Ob ein Test signifikant in diesem Sinne ausfällt oder nicht, hängt vor allem mit der Größe der Stichprobe zusammen. Mit zunehmender Größe lassen sich auch kleine und unbedeutende Zusammenhänge oder Unterschiede als signifikant absichern. Ein signifikantes Ergebnis kann daher nie mit einem wichtigen Einfluß gleichgesetzt werden.

Grundgesamtheit (Population)

Die Menge der Objekte, für die die Aussagen einer Untersuchung gelten sollen, beispielsweise "alle wahlberechtigten Bürger der Bundesrepublik Deutschland" oder "alle Personen im Alter von 14 bis 26 Jahren" (als eine von vielen möglichen Definitionen für "Jugendliche"). Wichtig ist die definitorische Abgrenzung der Grundgesamtheit, um eine nachvollziehbare Auswahl treffen und exakt angeben zu können, für wen die Untersuchungsergebnisse Gültigkeit beanspruchen.

Chi² (Chi-Quadrat)-Verteilung

Wahrscheinlichkeitsverteilung, die zur Signifikanzprüfung eingesetzt werden kann. Die wichtigsten Anwendungen sind:

1. Test auf Überzufälligkeit von Zusammenhängen in Kreuztabellen. Chi² errechnet sich als Summe der quadrierten Abweichungen der beobachteten von den erwarteten Zell-Häufigkeiten, jeweils dividiert durch die erwartete Zell-Häufigkeit. Die resultierende Teststatistik, oft auch als "Pearsons Chi²" bezeichnet, hat (bei Berechnung der erwarteten Häufigkeiten anhand der Randverteilung) (r-1)(c-1) Freiheitsgrade (r=Anzahl der Zeilen, c=Anzahl der Spalten der Tabelle).

In der Beispielstabelle beim Stichwort Kreuztabelle ergibt sich ein Chi² von 135,58 bei 4 Freiheitsgraden. Die Wahrscheinlichkeit, einen solchen Wert rein zufällig, also aufgrund von Stichprobenschwankungen zu erhalten, ist kleiner als 0,001 Prozent. Man wird daher annehmen, dass der beobachtete Zusammenhang auch in der Grundgesamtheit besteht.

Vor allem bei größeren Tabellen (mehr als 2x2 Zellen) wird oft übersehen, dass der Chi²-Test ein "globaler" Test ist; die Teststatistik gibt also nur an, ob sich *irgendwelche* überzufälligen Zusammenhänge bzw. Zellhäufigkeiten zeigen oder nicht, sie besagt aber nichts darüber, welche dies sind. In solchen Fällem ist es oft besser, die Zusammenhänge mit einem komplexeren Verfahren zu modellieren, z.B. einem ->log-linearen Modell.

Bei kleineren Stichproben (weniger als 60 Fälle) sollte nach Meinung vieler (aber nicht aller) Autoren ein Chi²-Wert mit sog. Yates-Korrektur verwendet werden. Wenn die *erwartete* Häufigkeit für wenigstens eine Zelle der Tabelle kleiner 5 ist (nach anderen Autoren: wenn mehr als 20 Prozent der erwarteten Häufigkeiten kleiner 5 sind), ist die Anwendbarkeit des

Chi²-Tests nicht mehr gegeben; dann sollten exakte Tests (z.B. nach Fisher) herangezogen werden.

Cronbachs Alpha

Alpha ist ein Maß zur Berechnung der internen Konsistenz einer aus mehreren Items zusammengesetzten Skala. Diese interne Konsistenz gilt als Schätzung der - >Reliabilität der Skala. Alpha kann - bei perfekter Konsistenz - ein Maximum von +1 erreichen; je kleiner der Wert (es sind auch negative Werte möglich!), desto geringer ist die Konsistenz. Die Formel für Alpha lautet:

Alpha = nr / (1 + r(n-1))

Dabei steht n für die Zahl der Items und r für den Mittelwert aus allen bivariaten Korrelationen zwischen den Items.

Deduktion

Verfahren des logischen Schließens von zwei Prämissen auf einen zu erklärenden Sachverhalt (das Explanandum). Die beiden Prämissen bestehen aus (a) einer allgemeinen Gesetzmäßigkeit (dem Explanans) und (b) einer Aussage, die einen Fall unter die Gesetzmäßigkeit subsumiert (häufig als Randbedingung bezeichnet).

Beispiel:
(Explanans) Alle Menschen sind sterblich.
(Randbedingung) Cäsar ist ein Mensch.
(Explanandum) Also ist Cäsar sterblich.

Deduktion steht im Zentrum der deduktiv-nomologischen Konzeption wissenschaftlicher Erklärungen nach Hempel/Oppenheim, sie ist aber auch wesentlich für das Prinzip der Falsifikation.

Deskriptive Statistik

Die deskriptive Statistik befasst sich mit Maßzahlen zur Charakterisierung von Daten: Wie kann ich die "zentrale Tendenz" eines Datenbündels kennzeichnen (siehe Lagemaße)? Wie die Streuung (Streuungsmaße)? Wie kann ich Zusammenhänge zwischen zwei oder mehreren Variablen charakterisieren (Korrelation, Regressionsanalyse)? Wie kann ich Daten oder "Fälle "bündeln (Clusteranalyse)? Auch viele Verfahren der grafischen Darstellung von Daten gehören hierher, hier finden sich aber auch Schnittstellen zur Explorativen Statistik.

Die D. S. heißt nicht etwa deswegen "deskriptiv", weil es hier "nur" um Beschreibung (statt um Erklärung) ginge, sondern deshalb, weil sie sich - im Gegensatz zur Inferenzstatistik - nicht damit beschäftigt, ob aus den vorliegenden Daten (bei denen es sich meist um Stichproben handelt) auf die Grundgesamtheit geschlossen werden darf, aus der die Daten stammen. Vielleicht hilft ein weiterer Begriff zum Verständnis: Manche Autoren bezeichnen die deskriptive Datenauswertung als "Datenreduktion" - es geht eben darum, ein- oder mehrdimensionale Datenbündel durch einfachere (im Sinne von: die Zahl der Zahlenwerte reduzierende) Kennwerte zu charakterisieren (vgl. Ehrenberg 1986).

Streuungsmaße

S., auch Dispersionsmaße genannt, geben an, wie stark die Merkmalswerte eines Datenbündels vom "Zentrum" (charakterisiert durch ein Lagemaß) abweichen. Die wichtigsten S. sind der Range, der Interquartilsabstand, der Interdezilbereich, die Standardabweichung, die Varianz und der Variationskoeffizient.

Range

Range, engl. für Spannbreite oder Variationsbreite.
Der R. ist die Differenz zwischen dem größten und dem kleinsten Wert eines Datenbündels; er ist mithin ein Streuungsmaß.

Da der R. sehr stark von einzelnen Werten abhängig ist, ist er im allgemeinen als alleinige Maßzahl zur Verdeutlichung der Streuung nicht gut geeignet.

Varianz

Die V. ist die Summe der quadrierten Abweichungen der einzelnen Werte eines Datenbündels vom Mittelwert, dividiert durch *n*, d.i. durch die Anzahl der Beobachtungen. Die V. ist also ein Maß dafür, wie weit die einzelnen Werte von Mittelwert entfernt liegen; es handelt sich mithin um ein Streuungsmaß.

Diese Definition gilt für die sogenannte "empirische Varianz", die ein gegebenes Datenbündel charakterisiert. Faßt man dieses Datenbündel als Stichprobe auf, wird die Summe der quadrierten Abweichungen nicht durch *n*, sondern durch *n - 1* dividiert.

Die V. kann nur bei metrischen Daten (siehe Meßniveau) berechnet werden. Da wegen der Quadrierung der Abweichungen Extremwerte ein größeres Gewicht haben, ist die Varianz u.U. nicht gut zur Verdeutlichung der Streuung geeignet.

Evaluation

Evaluation bezeichnet die systematische, datenbasierte Beschreibung und Bewertung von Programmen (z. B. Hilfe- oder Beratungskonzeptionen), zeitlich beschränkten Projekten (z. B. Modellvorhaben) oder Institutionen (z. B. Zulassung von Trägern) in Bildung, Sozialer Arbeit, Gesundheitswesen u. a.

Evaluation beschafft nützliche und abgesicherte Informationen für Auftraggeber, Beteiligte & Betroffene (engl. stakeholders). Sie unterstützt diese, entweder den bewerteten Gegenstand schrittweise zu stabilisieren / zu verbessern (*formative* oder Gestaltungs-Evaluation) oder zu bewerten (*summative* oder Bilanz-Evaluation).

Obwohl verwandt, folgt Evaluation einer anderen Logik als Forschung: Evaluation ist auf unmittelbar praktische Nützlichkeit ihrer Ergebnisse verpflichtet, weniger auf Mehrung theoretischer Erkenntnis; sie ist häufiger durch bei Beteiligten geankerte Fragestellungen gesteuert als durch theoretisch abgeleitete Hypothesen; sie hat als expliziten Auftrag, Werturteile zu fällen oder Beteiligte auszurüsten, dies informiert zu tun.

Evaluation kann auf alle vier Hauptdimensionen von insbesondere pädgogischen Gegenständen gerichtet werden: Das Konzept (insbesondere Zielsetzungen), die Struktur (gesetzliche Bestimmungen, Ausstattung u.v.m.), den Prozess (z.B. Ablauf des pädagogischen Geschehens, Reaktionen der Teilnehmenden darauf) und das Ergebnis

(kognitive/affektive Lerneffekte bei den Zielgruppen bzw. Änderungen ihrer materiellen Situation).

Im Ablauf einer Evaluation werden in der Gegenstandsbestimmung zunächst solche Fragestellungen formuliert, deren Beantwortung die Beteiligten zu verbesserter Gestaltung / Entscheidung befähigt. Zur Informationsgewinnung bedient sich die Evaluation qualitativer und quantitativer Methoden der Sozialforschung (Inhaltsanalyse, Beobachtung, Befragung). Die Ergebnisvermittlung legt Beteiligten und ggf. einer weiteren Öffentlichkeit Datenquellen und Instrumente sowie Resultate und Schlussfolgerungen nachvollziehbar dar.

In Nordamerika hat Evaluation eine etwa 70jährige Tradition und ist fester Bestandteil der politischen Kultur. In Deutschland lösten Mitte der 90er Jahre veränderte öffentliche Steuerungskonzepte einen Evaluation-Boom aus: Schulen, Universitäten und andere öffentlich finanzierte Einrichtungen insbesondere der Sozialen Arbeit sind gehalten, die Qualität ihrer Prozesse und Leistungen datenbasiert auszuweisen.

Evaluation ist gefordert, bei konfligierenden Interessen fair und unparteiisch zu verfahren, indem sie ihre Unabhängigkeit gegenüber mächtigen Einflussgruppen (z.B. finanzierenden Auftraggebern) behauptet und auch den Perspektiven artikulationsschwacher Beteiligter Raum verschafft. Da sie meist in öffentlichem Auftrag und steuerfinanziert durchgeführt wird, bedarf es eines Minimalkonsenses über Gütekriterien von Evaluation selbst. Je stärker die Rechnungshöfe Evaluation zu ihrem originären Auftrag machen, und in je mehr Politikfeldern (wie Umwelt-, und Energiepolitik, der Verkehrs- und Forschungspolitik) systematische Evaluation stattfindet, desto wichtiger werden fachlich präzise und ethisch korrekte Evaluationsverfahren.

Gütekriterien

Kriterien, anhand derer die Qualität sozialwissenschaftlicher (wie sonstiger) Forschung beurteilt werden können soll.

Im Rahmen der quantitativen/standardisierten Forschung wurden G. vor allem im Hinblick auf die Messung entwickelt. Die wesentlichen G. sind hier Validität, die Reliabilität sowie die ->Objektivität. Andere Teile des Forschungsprozesses, namentlich die Wahl eines angemessenen Forschungsdesigns oder der adäquaten (und adäquat durchgeführten!) statistischen Analysemethode werden weitaus weniger systematisch reflektiert bzw. unterliegen nicht im gleichen Maße *allgemein* anerkannten Qualitätsstandards.

Im Bereich der qualitativen Sozialforschung sind weitere G. entwickelt worden (während die oben genannten teilweise - wenn auch möglicherweise modifiziert - akzeptiert, z. T. auch abgelehnt werden). Zu diesen weiteren (bzw. anderen) G. zählen die kommunikative Validierung, die theoretische Sättigung, die Authentizität und die intersubjektive Nachvollziehbarkeit (letzteres freilich ein Kriterium, welches auch als allgemeine Grundlage der G. der standardisierten, wenn nicht jeglicher Forschung formuliert werden kann; allerdings ist auch dieses Kriterium im Rahmen der qualitativen Forschung nicht unumstritten).

Validität

Unter V. versteht man die *Gültigkeit* von Messungen, d.h. die Eigenschaft, genau das zu messen, was gemessen werden soll. Sie ist zu unterscheiden von der Reliabilität oder

Zuverlässigkeit von Messungen. Messinstrumente können sehr exakt immer das Falsche messen; dann sind sie zwar reliabel, aber nicht valide.

Die Bestimmung der V. ist im allgemeinen nicht einfach, und sie kann fast nie als endgültig betrachtet werden. Dies schon deshalb, weil man streng genommen die V. nur mittels eines anderen Messinstruments prüfen kann, dessen V. bereits bekannt sein müsste - so dass man hier im Prinzip in einen infiniten Regress kommt. Die Bestimmung der V. ähnelt also - wie so vieles in der Sozialforschung - eher einer Detektivarbeit als einem eindeutigen und klar geregelten Vorgehen mit ebenso eindeutigen und klaren Ergebnissen.

Man unterscheidet heute im wesentlichen drei Arten von V.: Die *Inhaltsvalidität*, die *Kriteriumsvalidität* und die *Konstruktvalidität*.

Inhaltsvalidität

Die I. (englisch: content validity) bedeutet, dass die Gültigkeit der Messung mehr oder weniger für jedermann einsichtig aus den einzelnen Teilen des Messinstruments hervorgeht. Letztlich beruht sie auf der Kenntnis von 'Experten' über den betreffenden Gegenstand (wobei u.U. sehr viele oder alle Leute Experten sein können). So wird z.B. bei jeder Prüfung I. unterstellt: Eine Mathematikprobe sagt etwas über die 'Mathematik-Fähigkeit', weil ein Mathematiklehrer oder -professor in der Lage sein sollte zu beurteilen, was gute und was schlechte Mathematik-Fähigkeiten sind. Die Behauptung, ein Messinstrument habe I., bedeutet in der Forschungspraxis aber oft nichts anderes, als dass der Entwickler des Instruments selbst glaubt, das Instrument sei valide. Ehrliche Personen gebrauchen hier den Begriff der 'face validity', d.h. der 'augenscheinlichen Validität'.

Dennoch ist die Idee der I. sehr wichtig: Es geht letztlich eben darum, dass eine Messung das relevante Phänomen möglichst in allen Aspekten erfasst, und dies kann nur durch Forschen, Nachdenken und Kommunikation zwischen Wissenschaftlern herausgefunden werden und nicht durch bestimmte, immer funktionierende 'Techniken'.

Kriteriumsvalidität

Bei der K. (englisch: criterion-related validity) geht es um die Übereinstimmung eines Messinstruments mit anderen relevanten Merkmalen (sog. Außenkriterien). Genauer unterscheidet man hier zwischen der *Übereinstimmungsvalidität* (engl: concurrent validity) (das Außenkriterium wird gleichzeitig erhoben) und der *Vorhersagevalidität* (engl.: predictive validity), bei der das Außenkriterium erst später gemessen wird. Übereinstimmungsvalidität wird z.B. erhoben, wenn die Messung mit dem gleichen Merkmal in Beziehung gesetzt wird, wie es durch eine andere Messung ermittelt wurde (Beispiel: Ein kürzerer Intelligenztest wird mit einem längeren verglichen). Eine andere Form ist die Methode der 'bekannten Gruppen' (known groups). So sollte eine Skala, die Ausländerfeindlichkeit misst, bei Mitgliedern rechtsradikaler Parteien viel höhere Werte ergeben als bei solchen liberaler oder linker Parteien. Vorhersagevalidität ist z.B. eine wichtige (und keineswegs immer gegebene!) Eigenschaft von Studieneingangstests, sie wird also gemessen anhand des späteren Studienerfolgs.

Konstruktvalidität

K. (englisch: construct validity) ist ein komplexes Vorgehen, bei dem man eine Reihe von plausiblen oder sogar bestätigten Hypothesen prüft, die sich u.a. auf das Konstrukt beziehen, dessen V. geprüft werden soll. Wenn sich diese Hypothesen auch jetzt bestätigen, so ist

anzunehmen, dass das fragliche Messinstrument auch gültig ist. Eine Nicht-Bestätigung der Hypothesen kann allerdings auch bedeuten, dass die angeblich plausiblen oder bestätigten Hypothesen eben doch falsch waren, oder dass die anderen Variablen mit nicht validen Instrumenten gemessen wurden.

Eine besondere Form der Konstruktvalidität ist die Bestimmung mit Hilfe einer Multi-Trait-Multi-Method-Matrix. Hier werden mehrere Eigenschaften mit jeweils mehreren Instrumenten gemessen; die Messungen der gleichen Eigenschaften mit verschiedenen Instrumenten sollten dabei stärker untereinander zusammenhängen als die verschiedener Eigenschaften mit den gleichen Instrumenten.

Reliabilität (Zuverlässigkeit)

Neben der Validität (Gültigkeit) das zweite zentrale Qualitätskriterium bei Messungen. Meint, daß Messinstrumente bei wiederholter Messung unter gleichen Bedingungen auch das gleiche Ergebnis produzieren müssen.

Korrelation

K. ist ein statistischer Fachbegriff für "Zusammenhang". Alternative Begriffe sind Assoziation oder eben auch Zusammenhang. K.-smaße drücken die Stärke des Zusammenhangs zwischen zwei Variablen aus. Werden dabei Zusammenhänge zwischen diesen und weiteren Variablen berücksichtigt, so spricht man von Partialkorrelation oder partieller Korrelation.

Maße für die Stärke der K. werden meist als Korrelationskoeffizienten bezeichnet. Häufig können K.-skoeffizienten Werte zwischen minimal -1 und maximal +1 annehmen, wobei -1 einen perfekten negativen und +1 einen perfekten positiven Zusammenhang bezeichnet. Es gibt aber auch K.-skoeffizienten, die die Richtung des Zusammenhangs nicht durch unterschiedliche Vorzeichen ausdrücken. Manche K.-skoeffizienten können auch nicht das Maximum von +1 erreichen.

Dass eine (gegebenenfalls: partielle) K. zwischen zwei Merkmalen besteht, ist zwar eine notwendige, aber keine hinreichende Bedingung für die Annahme eines ->Kausalzusammenhanges.

Die Wahl des Korrelationskoeffizienten hängt vom Messniveau der Variablen ab. Die wichtigste K.-skoeffizienten sind:

> bei zwei nominalskalierten Merkmalen Phi, der Kontingenzkoeffizient C, Cramer's V, Tau (PRE-Maß von Goodman und Kruskal für nominalskalierte Daten) oder Lambda;
> bei zwei ordinalskalierten Merkmalen Tau-b, Tau-c, Somers' D, Gamma oder u. U. der ->polychorische Korrelationskoeffizient.
> bei zwei metrischen Merkmalen die Produkt-Moment-Korrelation (auch Bravais-Pearson'scher Korrelationskoeffizient genannt), die oft mit r abgekürzt wird. Manchmal wird dieses Korrelationsmaß einfach als "Korrelation" bezeichnet, wenn aus dem Zusammenhang klar ist, dass es sich um kein anderes Maß handelt.

Kovarianz

Die Kovarianz beschreibt den Zusammenhang zwischen zwei metrischen Merkmalen. Sie wird folgendermaßen berechnet: Für jeden Wert der beiden Variablen wird die Abweichung

vom jeweiligen Arithmetischen Mittel (durch Subtraktion desselben) berechnet. Für jeden Fall wird nun das Produkt dieser beiden Abweichungen gebildet, die Produkte werden aufsummiert und durch *n - 1* dividiert (*n* = Zahl der Fälle). Die Berechnung der K. ist nur bei metrischen Variablen sinnvoll.

Was bedeutet das ganze? Man kann sich vor Augen halten, was die einzelnen Produkte besagen: Wenn ein positiver Zusammenhang zwischen den beiden Variablen besteht, so heißt das, daß relativ häufig *entweder* ein überdurchschnittlich großer Wert in der einen Variablen (nennen wir sie X) mit einem überdurchschnittlich großen Wert in der anderen (nennen wir sie Y) *oder* ein unterdurchschnittlicher Wert in X mit einem unterdurchschnittlichen Wert in Y gemeinsam auftritt. In beiden Fällen ist das Produkt der beiden Abweichungen positiv, da entweder zwei Werte mit positiven oder zwei Werte mit negativen Vorzeichen miteinander multipliziert werden.
Besteht dagegen ein negativer Zusammenhang zwischen den beiden Merkmalen, so wird relativ häufig *entweder* ein überdurchschnittlich großer Wert in X mit einem unterdurchschnittlich großen Wert in Y *oder* ein unterdurchschnittlicher Wert in X mit einem überdurchschnittlich großen Wert in Y gemeinsam auftreten. Die Produkte der beiden Abweichungen werden in diesen Fällen negativ, da jeweils ein Wert mit positivem und ein Wert mit negativem Vorzeichen miteinander multipliziert werden.
Bei einem positiven Zusammenhang überwiegen also Werte mit positivem Vorzeichen, bei einem negativen Zusammenhang dagegen Werte mit negativem Vorzeichen.

Das heißt also: Ist der Wert der K. positiv, so besteht auch ein positiver (gleichsinniger) Zusammenhang zwischen den beiden Variablen; ist die K. negativ, so besteht ein negativer (gegensinniger) Zusammenhang. Allerdings besagt die Größe der Kovarianz noch relativ wenig, da sie ja vom Maßstab der beiden Variablen abhängt. Daher wird zur Beschreibung von Zusammenhängen zweier (metrischer) Variablen meist die Produkt-Moment-Korrelation berechnet, welche gleichsam eine standardisierte K. darstellt.

Phi

Maßzahl für die Stärke des Zusammenhangs zwischen zwei nominalskalierten Variablen bei einer 2x2-Tabelle. (Wenn mindestens eines der beiden Merkmale mehr als zwei Ausprägungen hat, sollte stattdessen der Kontingenzkoeffizient herangezogen werden).

Die Formel für Phi lautet: Phi = Sqrt(Chi² / N).
Dabei steht Chi² für die gleichnamige Teststatistik für Kontingenztabellen nach Karl Pearson, N für die Gesamtzahl der Fälle in der Tabelle. Es handelt sich also um ein sog. Chi-Quadrat-basiertes Zusammenhangsmaß.

Phi nimmt Werte zwischen 0 (kein Zusammenhang) und 1 (perfekter Zusammenhang) an. Die Richtung des Zusammenhangs wird aus der zugrundeliegenden Kreuztabelle ersichtlich.

Pragmatik

Eine sehr einfache Defition könnte lauten: Die Untersuchung (bzw. Lehre) vom Gebrauch der Sprache in der Kommunikation. Insbesondere bezieht sich Pragmatik auf die Beziehung zwischen Sprache und Sprecher/Hörer/Interpret.

Faktisch ist Pragmatik (der Begriff geht auf Charles W. Morris zurück) der am schwersten zu fassende Teil der Sprachwissenschaft. Andere Definition beziehen sich etwa auf die

Kontextabhängigkeit der Sprache. Immer geht es aber darum, Sprache nicht rein grammatikalisch oder nur hinsichtlich ihres abstrakten Bedeutungsgehaltes zu untersuchen, sondern sie im Kontext ihrer Verwendung zu erforschen.

Obwohl auch Morris selbst sich mit der philosophischen Strömung des Pragmatismus (Charles S. Peirce, William James, John Dewey) auseinandergesetzt hat, bestehen allenfalls sehr oberflächliche Beziehungen zwischen dieser Richtung und der Pragmatik im sprachwissenschaftlichen Sinne.

Produkt-Moment-Korrelation

Die P.-M.-K., auch als Bravais-Pearson'scher Korrelationskoeffizient bezeichnet, ist ein Zusammenhangsmaß für metrische (siehe Meßniveau) Variablen. Sie wird berechnet als Kovarianz der beiden interessierenden Variablen, dividiert durch das Produkt der Standardabweichungen der beiden Variablen. Die P.-M.-K. kann Werte zwischen +1 (perfekter positiver Zusammenhang) und -1 (perfekter negativer Zusammenhang) annehmen. Ein Wert von 0 indiziert die Abwesenheit eines (linearen) Zusammenhanges.

Standardisieren

Metrische Variablen werden standardisiert durch die Transformation

$Z = (X - M) / S$

mit X als ursprünglichem Messwert, M als arithmetischem Mittel der Variablen x und S als Standardabweichung von x. Die standardisierten Werte werden häufig auch als Z-Werte bezeichnet. Sie haben einen Mittelwert von 0 und eine Standardabweichung von 1. Wie der Name schon sagt, dient die S. hauptsächlich dazu, verschiedene Messwerte vergleichbar zu machen.

Standardfehler

Der S. ist die Streuung von Stichprobenkennwerten um dem wahren Wert des gesuchten Parameters in der Grundgesamtheit. Je größer der S., desto geringer die Wahrscheinlichkeit, dass der Stichprobenkennwert den Parameter richtig schätzt. Die Größe des S. hängt ab von der Varianz der Messwerte in der Grundgesamtheit (je geringer diese Varianz - also die Unterschiedlichkeit der Werte-, desto geringer auch der Standardfehler) und vom Umfang der Stichprobe.

Der S. ist normalerweise unbekannt und wird aus der Stichprobe geschätzt. Verschiedene Verfahren, z.B. geschichtete Stichproben, sind geeignet, den S. zu verringern.

t-Test

Der t-Test (manchmal nach seinem Erfinder auch Student's t-test genannt - Student war übrigens ein Pseudonym) ist ein Verfahren der Inferenzstatistik zur Prüfung, ob sich die Mittelwerte (genauer: arithmetischen Mittel) zweier Stichproben überzufällig unterscheiden (oder - seltener - ob der Unterschied der Mittelwerte einen bestimmten Betrag überschreitet oder nicht). Dabei kann es sich um zwei unabhängige oder zwei abhängige (verbundene) Stichproben handeln. Der erste Fall liegt vor, wenn Stichproben aus zwei verschiedenen

Grundgesamtheiten gezogen wurden, der zweite z.B. dann, wenn das gleiche Merkmal an ein und derselben Stichprobe zweimal gemessen wurde.

Die Logik des t-Tests ist folgende: Es wird eine Prüfgröße berechnet, die einer t-Verteilung mit k Freiheitsgraden folgt (k wird häufig berechnet als n+m-2, mit n und m als Umfang der beiden Stichproben, es finden sich in der Literatur aber auch andere Formeln). Dabei wird vorausgesetzt, daß die Varianz der Grundgesamtheit(en) aus den Stichproben geschätzt werden muß. Ist die Varianz der Grundgesamtheit(en) bekannt (wird sie also nicht aus der Stichprobe geschätzt), so folgt die Prüfgröße der Standardnormalverteilung, ebenso bei größeren Stichprobenumfängen (darunter versteht man meist, daß n + m > 30 sein sollen, andere Autoren geben n + m > 50 als Grenze an).

Anwendungsvoraussetzungen: Der t-Test kann eingesetzt werden, wenn die zu untersuchende abhängige Variable mindestens (mehr oder weniger) intervallskaliert ist (siehe Meßniveau). (Das geht schon daraus hervor, daß Mittelwerte untersucht werden). Voraussetzung für die Gültigkeit des t-Tests sind ferner Varianzhomogenität (also gleiche Varianz in den Gruppen) sowie Normalverteilung der abhängigen Variablen. Die Varianzhomogenität wird z.B. durch den Levene-Test geprüft. (Das Ergebnis dieses Tests sollte *nicht* signifikantsein , d.h., die Varianzen sollten sich gerade nicht signifikant voneinander unterscheiden). Sind Normalverteilung und/oder Varianzhomogenität nicht gegeben, kann ein ->verteilungsfreies (nonparametrisches) Prüfverfahren gewählt werden; es stehen aber auch Modifikationen des t-Tests für den Fall von Varianzinhomogenität zur Verfügung. Der t-Test ist insgesamt - vor allem bei großen Stichproben - relativ robust gegen Verletzungen der zugrundeliegenden Annahmen; eine genauere Diskussion findet sich bei Bortz.

Varianzanalyse

Eine Klasse von statistischen Analyseverfahren zur Durchführung von Mittelwertvergleichen zwischen mehreren Gruppen (bei zwei Gruppen siehe auch t-Test).

Werden Mittelwertunterschiede *einer* (abhängigen) Variablen geprüft, so spricht man von *univariater* V. (ANOVA), bei simultanen Tests mehrerer abhängiger Variablen von *multivariater* V. (MANOVA). Werden die Untersuchungspersonen (oder allgemeiner: Untersuchungsobjekte) hinsichtlich *eines* Merkmals in Gruppen eingeteilt, spricht man von *einfaktorieller* V. (die Gruppenzugehörigkeit wird auch als Faktor bezeichnet), bei mehreren Gruppierungsmerkmalen von *mehrfaktorieller* (bzw. konkret zwei-, drei- usw. -faktorieller) V. Bei der mehrfaktoriellen V. können auch Interaktionseffekte geprüft werden, d. h. unterschiedliche Wirkungen eines Faktors in Abhängigkeit von den Ausprägungen eines oder mehrerer anderen Faktors/Faktoren.

Wichtige Spezialfälle sind die V. bei *Messwiederholungen* sowie die *Kovarianzanalyse* (ANCOVA) (die Durchführung von Gruppenvergleichen bei simultaner Berücksichtigung des potenziellen Einflusses weiterer, i. a. metrischer Variablen).

Die V. ist wohl das wichtigste statistische Auswertungsverfahren in der Psychologie, da sie sich nicht zuletzt besonders gut für die Auswertung experimenteller Daten eignet. Dieser kurze Einführungsartikel kann der Komplexität der verschiedenen Varianten der V. keinesfalls gerecht werden, hierzu muss auf die ein- und weiterführende Literatur verwiesen werden.

Grundlagen

Die *Grundidee der V.* besteht darin, die gesamte Varianz des zu erklärenden Merkmals, der abhängigen Variablen (oder mehrerer solcher Variablen) aufzuteilen (zu "zerlegen") in die Varianz *zwischen* den Gruppen - die Abweichung der Gruppenmittelwerte vom Gesamtmittelwert über alle Gruppen bzw. Untersuchungseinheiten - und die Varianz *innerhalb* der Gruppen (die Abweichung der einzelnen Messwerte innerhalb der Gruppen vom Gruppenmittelwert, auch Residualvarianz oder Fehlervarianz genannt). Sind die Unterschiede *zwischen* den Gruppen relativ groß bei gleichzeitig nicht allzu großer Varianz innerhalb der Gruppen, so kann man davon ausgehen, dass die Gruppenzugehörigkeit einen Einfluss auf die abhängige Variable hat. Formal wird dies geprüft über den F-Test

$F = s_{zw} / s_{in}$

wobei s_{zw} (üblicherweise geschrieben als kleines griechisches sgima mit einem Dach) die (geschätzte) Varianz zwischen den Gruppen und s_{in} die (geschätzte) Varianz innerhalb der Gruppen darstellt. Einzelheiten dazu im Folgenden bei der Darstellung der wichtigsten Verfahren.

Als Maß für die Erklärungskraft der untersuchten Faktoren (der Determination des abhängigen Merkmals durch die Gruppenzugehörigkeit) steht das PRE-Maß Eta² zur Verfügung.

Voraussetzung für die Gültigkeit der inferenzstatistischen Absicherung der Mittelwertunterschiede ist die Normalverteilung der abhängigen Variablen in der Grundgesamtheit sowie die Gleichheit der Varianzen in den einzelnen Gruppen (Varianzhomogenität). Die Varianzanalyse ist aber einigermaßen robust gegen moderate Verletzungen dieser Annahmen, vor allem wenn die Gruppen nicht zu klein und (in etwa) gleich groß sind (vgl. dazu die einschlägigen Angaben bei Bortz).

Die wichtigsten (einfacheren) Verfahren

Einfaktorielle univariate Varianzanalyse

Die Prüfgröße F ist hier F-verteilt mit p-1 und n-p Freiheitsgraden (n = Anzahl aller Untersuchungseinheiten, p = Zahl der Gruppen). Sie gibt Auskunft darüber, ob *irgendwelche* Unterschiede zwischen den Gruppen bestehen (anders formuliert: ob die Gruppen aus der gleichen Grundgesamtheit stammen oder nicht), es handelt sich also um eine globale Teststatistik. Um zu prüfen, ob spezifische Gruppenmittelwerte höher oder niedriger (als der Gesamtmittelwert oder als andere Gruppenmittelwerte) sind, sollte man A-priori-Hypothesen durch Angaben von Kontrasten testen. Es stehen aber auch A-posteriori-Tests zur Verfügung, von denen z. B. der Scheffé-Test empfehlenswert (weil eher konservativ) ist. Die Annahme der Varianzhomogenität wird beispielsweise durch den Bartlett-Test geprüft. (Das Ergebnis dieses Tests sollte *nicht* signifikant sein , d. h., die Varianzen sollten sich gerade nicht signifikant voneinander unterscheiden).

Mehrfaktorielle univariate Varianzanalyse

Hier können nicht nur die Effekte der einzelnen Faktoren (Haupteffekte), sondern auch gemeinsame Effekte (Interaktionseffekte) geprüft werden. Neben dem globalen F-Test wird für jeden dieser Effekte ein separater F-Test durchgeführt. Dabei ist (am Beispiel der zweifaktoriellen V. erläutert) die Zahl der Freiheitsgrade des Haupteffekts für Faktor A = p-1 (p = Zahl der Gruppen in Faktor A), für Faktor B = q-1 (q = Zahl der Gruppen in Faktor B), für den Interaktionseffekt AxB = (p-1)(q-1) und für die Fehlervarianz n - (p*q).

Prüfungen von Hypothesen über spezifische Effekte einzelner Faktoren sind hier nicht immer ohne weiteres möglich. Sind signifikante Interaktionseffekte vorhanden, ist die Interpretation der Haupteffekte problematisch. Auch entstehen Probleme, wenn (wie es bei nicht-experimentellen Designs häufig der Fall ist) die einzelnen Faktoren nicht voneinander unabhängig sind. In diesem Fall kann die Varianz nicht mehr eindeutig in ihre verschiedenen Anteile zerlegt werden; die Gruppen weisen gemeinsame Varianzanteile auf. Zur Lösung dieses Problems werden verschiedene Strategien angeboten: Bei *hierarchischem* Faktoreneinschluss wird zunächst der Beitrag des ersten Faktors zur Varianzerklärung geprüft und anschließend derjenige Beitrag des zweiten (und gegebenenfalls weiterer) Faktors/Faktoren, der durch den ersten Faktor noch nicht erfasst wurde. Die gemeinsame Varianzerklärung beider Faktoren wird damit dem ersten Faktor zugeschlagen, und die Ergebnisse können sehr stark von der Reihenfolge des Einschlusses der verschiedenen Faktoren abhängen. Bei *simultanem* Einschluss der Faktoren wird ähnlich wie beim linearen Regressionsmodell versucht, für jeden Faktor den nur durch diesen Faktor erklärten Varianzanteil zu bestimmen.

Auch Tests spezifischer Kontraste sind hier nicht ohne weiteres durchzuführen, die verfügbaren Statistik-Programme erlauben aber im allgemeinen Tests der Abweichung einzelner Gruppen vom Gesamtmittelwert oder von einer Referenzgruppe und häufig weitere Tests.

Auch bei der mehrfaktoriellen V. ist die Gültigkeit der Annahme der Varianzhomogenität durch eine geeignete Statistik zu testen.

Multivariate Varianzanalyse

Hier verkomplizieren sich die bislang angestellten Überlegungen noch beträchtlich; da es mehrere abhängige Variablen gibt, kann nicht *eine* Quadratsumme berechnet und für Tests herangezogen werden. Mehrere alternative globale Teststatistiken stehen zur Verfügung, etwa Wilk's Lambda, Hotelling's Spur-Kriterium, Roy's größter Eigenwert oder Pillai's Spur-Kriterium. Auch die Anwendungsvoraussetzungen der Signifikanztests sind komplexer. Es sollte zunächt geprüft werden (mit Bartletts Sphärizitätstest), ob die abhängigen Variablen untereinander zusammenhängen; ist dies nicht der Fall, können statt der multivariaten V. mehrere univariate V.n durchgeführt werden, die leichter zu berechnen und zu interpretieren sind. Ferner ist nicht nur die Homogenität der Varianzen in den einzelnen Gruppen zu prüfen; auch die Kovarianzen zwischen den abhängigen Variablen innerhalb der Gruppen sollen sich (vor allem bei kleineren Stichproben) nicht erheblich voneinander unterscheiden (Prüfung durch Box' M).