Variable

Merkmal, das verschiedene Ausprägungen annehmen kann (z.B. unterschiedliche Körpergrößen, Einkommen, Alter, Geschlechter, Parteipräferenzen ...). Nahezu alle in den Sozialwissenschaften erhobenen Merkmale sind Variable.

Datenmatrix

Tabellarische Organisationsform von Daten für die computergestützte Datenanalyse. Datenmatrizen sind rechteckig. In den Zeilen der Matrix werden die einzelnen Merkmalsträger aufgelistet, in den Spalten die jeweils untersuchten Merkmale, in den Zellen der Matrix werden die Codezahlen der Merkmalsausprägungen eingetragen.

Skala (engl.: Scale)

Menge von Items zur Messung eines ganz bestimmten Sachverhaltes, deren jeweilige Codezahlen in der Datenanalyse zu einem Gesamtwert (Skalenwert) zusammengefaßt werden. Primär werden Skalen zur Messung von Einstellungen verwendet.

Die bekanntesten unter den in den Sozialwissenschaften eingesetzten Verfahren sind die Likert-Skala und (in der Praxis viel seltener eingesetzt) die Guttman-Skala.

Weitere, hier nicht behandelte (weil teilweise zu Recht, teilweise zu Unrecht selten eingesetzte) Verfahren sind die Thurstone-Skala, die Rasch-Skala und die Magnitude-Skala.

Messniveau (engl.: Level of Measurement)

Messen bedeutet (vereinfacht gesagt) die Zuordnung von Zahlen zu Beobachtungen. Das Messniveau – häufig auch als Skalenniveau oder Skalentyp bezeichnet – gibt an, wie man diese Zahlen interpretieren darf, und damit auch, welche Operationen mit den Zahlen sinnvoll sind. Es werden vier Messniveaus unterschieden: Nominal-, Ordinal-, Intervall- und Ratioskala. Bei Messung auf einem der beiden letzteren Niveaus spricht man auch von *metrischen* Merkmalen.

> Bei einer *Nominalskala* bedeuten unterschiedliche Zahlen nichts anderes als unterschiedliche Merkmalsausprägungen;sie stehen *nicht* für ein »Mehr« oder »Weniger«, »Größer« oder »Kleiner«.
> *Beispiele:* Parteipräferenz; Haarfarbe; ausgeübtes Hobby.
> Bei solchen Daten können z.B. weder das Arithmetische Mittel noch der Median berechnet werden (bzw. wären die Ergebnisse entsprechender Berechnungen sinnlos).
>
> Bei einer *Ordinalskala* drücken die Zahlen eine Rangfolge aus, aber sie sagen nichts über die Relationen der der Rangfolge zugrundeliegenden Eigenschaften.
> *Beispiele:* Angaben darüber, welches gesellschaftliche Problem (z.B.: Arbeitslosigkeit, Kriminalität, Umweltverschmutzung – selten wird gefragt: Korruption, Unfähigkeit von Politikern, Gewalt von Männern gegen Frauen) man für am wichtigsten hält, welches am zweitwichtigsten, usf. Oder: Welche Person aus einer vorgegebenen Liste am sympathischsten ist, welche am zweitsympathischsten, usf.
> Gleiche Abstände zwischen den Zahlenwerten bedeuten also nicht gleiche Abstände »in der Realität«. Bei einem Wettrennen wissen wir, dass die oder der Erstplazierte schneller als die oder der Zweite, und diese(r) wiederum schneller als die oder der Dritte war; aber die Zahlen sagen nichts darüber aus, *um wieviel* schneller oder langsamer die LäuferInnen im Vergleich

zu den anderen waren. Das gleiche gilt auch für die hier genannten sozialwissenschaftlichen Beispiele. – Bei solchen Daten kann z.B. der Median berechnet werden.

Bei einer *Intervallskala* geben die Zahlen Informationen über die Abstände zwischen den gemessenen Ausprägungen, aber es gibt keinen »echten« Nullpunkt.
Beispiele: Temperatur in Grad Celsius; Haushaltseinkommen, sofern auch negatives Einkommen in Form von Schulden einbezogen wird; Kontostand (dieser kann leider auch negativ sein ...).
Die Abstand zwischen 0 Grad Celsius und 10 Grad Celsius ist (physikalisch gesehen!) genauso groß wie der zwischen 10 und 20 Grad. Bei intervallskalierten Daten ist neben dem Median u.a. auch die Berechnung von arithmetischem Mittel und Varianz sinnvoll.

Bei einer *Ratioskala* ist außerdem ein sinnvoll interpretierbarer Nullpunkt vorhanden.
Beispiele: Größe; Einkommen aus Erwerbstätigkeit; Temperatur in Grad Kelvin.
Daher kann man auch Verhältnisse zwischen verschiedenen Werten berechnen. Man kann legitimerweise sagen, dass eine Person, die 100 m in 10 Sekunden bewältigt, doppelt so schnell läuft wie eine, die dafür 20 Sekunden benötigt (während 20 Grad Celsius nicht doppelt so warm ist wie 10 Grad Celsius!).

Während das in der Theorie alles klar ist, ist in der Praxis manchmal nicht leicht zu bestimmen, welches Skalenniveau tatsächlich vorliegt. Vor allem stellt sich immer wieder die Frage, wie puristisch man sein darf oder soll. Ein wichtiges Beispiel sind die in den Sozialwissenschaften beliebten Likert-skalierten Items (Aussagen, hinsichtlich derer man Zustimmung oder Ablehnung auf einer Skala angeben kann, etwa von »stimme voll und ganz zu« bis »lehne voll und ganz ab«). Den Angaben werden (je nach Zahl der Ausprägungen in der Antwortvorgabe) Zahlen zugeordnet, z.B. von 1 bis 5. Realistischerweise wird man nun nicht behaupten können, dass es sich um eine Intervallskala handelt; wir wissen nicht, ob der Abstand zwischen einer Person, die beispielsweise im Fragebogen ein Kreuz ganz links macht (oder eine 1 einträgt) zu einer Person, die ihr Kreuz in der zweiten Kategorie von links einträgt (oder eine 2 angibt), wirklich der gleiche Abstand ist wie zwischen Personen, die die zweite bzw. die dritte Ausprägung angegeben haben. Auf der anderen Seite sind diese Daten sicherlich etwas anderes als z.B. die Daten, die man erhält, wenn man eine Schulklasse der Größe nach aufstellt und abzählen lässt. In diesem Fall haben wir tatsächlich keine Ahnung, wie groß die größte Person ist oder wie groß der Abstand zwischen dieser und der kleinsten Person ist; es kann sich also um eine Klasse handeln, in der alle Personen zwischen 1,75 und 1,78 m groß sind, sie können aber ebenso gut zwischen 1,45 und 2,10 m groß sein. Bei der Einstellungsskala ist zumindest die Idee eine ganz andere; intendiert ist im Grunde so etwas wie eine Ratioskala. Wir erwarten eigentlich, dass Personen, die angeben »lehne voll und ganz ab«, gleichsam den »Nullwert« haben, und Personen mit der Angabe »stimme voll und ganz zu« den denkbaren Maximalwert.

Wie gesagt – das sind die Absichten einer solchen Messung. Damit ist natürlich nichts darüber gesagt, ob diese Absichten tatsächlich realisiert worden sind; das dürfte höchst selten einmal der Fall sein und lässt sich obendrein, sollte es doch geschehen sein, wohl kaum nachweisen. Faktisch hat sich dennoch eingebürgert, mit solchen Items zu verfahren, als ob sie wenigstens Intervallskalenniveau aufweisen würden; daher werden z.B. Faktorenanalysen durchgeführt.

Die Problematik der Messniveaus hat sich insofern in den vergangenen Jahren »entschärft«, als nunmehr Verfahren der multivariaten Analyse für praktisch alle Skalenniveaus verfügbar sind (am wenigsten noch für »echte« Rangdaten im Sinne des Beispiels der Schulklasse).

Zu wenig beachtet werden in der Lehrbuchliteratur andere Eigenschaften von Daten. So sind etwa Daten über Dauern (z.B.: wie lange ist jemand arbeitslos) – die eigentlich Ratioskalenniveau aufweisen – häufig nicht vollständig erhebbar (weil zum Erhebungszeitpunkt die Arbeitslosigkeit noch nicht beendet ist); man spricht von sog. »zensierten« Daten, die mit speziellen Verfahren der Verweildaueranalyse ausgewertet werden müssen. Auch für sog. Zähldaten (Count Data) sind eigene statistische Verfahren angemessen. Grundsätzlich versucht man in der Statistik immer mehr, nicht so sehr (oder nicht nur) auf das Messniveau zu achten, sondern auf den Prozess, durch den die Daten entstanden sind.

Deskriptive Statistik (engl.: Descriptive Statistics)

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 graphischen 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).

Arithmetisches Mittel (engl.: Arithmetic Mean)

Lagemaß zur Kennzeichnung von metrischen (also mindestens intervallskalierten) Daten. Oft auch einfach als "Mittelwert" bezeichnet, was streng genommen wegen der Existenz anderer Mittelwerte (etwa geometrisches oder harmonisches Mittel) nicht korrekt ist, aber sicher dann zulässig, wenn aus dem Kontext eindeutig hervorgeht, was gemeint ist. Es wird berechnet als:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Das bedeutet, dass die Summe der Einzelwerte des Datenbündels durch die Zahl der Elemente dividiert wird.

Werden die kleinsten und größten Werte des Datenbündels von der Berechnung ausgeschlossen, so spricht man vom *getrimmten* a. M. (z.B. 5-%-getrimmtes Mittel, wenn die 5 % kleinsten und 5 % größten Werte weggelassen werden). Dies kann als sinnvoll erachtet werden, um den Einfluss extremer Werte auf das a. M. zu reduzieren.

Varianz (engl.: Variance)

Die Varianz

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

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 im Durchschnitt von Mittelwert entfernt liegen; es handelt sich mithin um ein Streuungsmaß.

Die oben angegebene Formel charakterisiert ein gegebenes Datenbündel. Handelt es sich bei den Daten um eine Stichprobe und soll ein Schätzwert für die Varianz in der Grundgesamtheit berechnet werden, so wird statt dessen die Größe

$$S^2 \text{ (oder } \hat{\sigma}^2) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

herangezogen.

Wichtiger Hinweis: Hinsichtlich der sprachlichen Bezeichnung der beiden angegebenen Berechnungsmöglichkeiten für die Varianz hat sich leider eine offensichtlich unaufhebbare Konfusion eingebürgert. Wie man durch eine schnelle Internetrecherche ebenso wie durch einen vergleichenden Blick in eine Reihe von Lehrbüchern feststellen kann, werden die Begriffe "empirische Varianz", "Stichprobenvarianz" und manchmal auch nur "Varianz" alternativ für beide Formeln verwendet. Es kann also vorkommen, dass ein Buch oder eine sonstige Quelle den Begriff "empirische Varianz" für die erste und den Begriff "Stichprobenvarianz" für die zweite Formel verwendet und das nächste Buch genau umgekehrt verfährt! Wieder andere geben nur die eine oder die andere Formel an, ohne auf deren exakte Bedeutung hinzuweisen. Beim Konsultieren statistischer Texte ist also höchte Vorsicht geboten und es ist dringend erforderlich, sich jeweils ein genaues Bild darüber zu verschaffen, was mit "Varianz", "empirischer Varianz" oder "Stichprobenvarianz" gemeint ist.

Ein Konfidenzintervall für die Varianz in einer Grundgesamtheit lässt sich mit einer Irrtumswahrscheinlichkeit von α folgendermaßen bestimmen:

Untere Grenze: $\frac{(n-1)S^2}{\chi^2_{1-\alpha/2}(n-1)}$

Obere Grenze: $\frac{(n-1)S^2}{\chi^2_{\alpha/2}(n-1)}$

Dabei ist $\chi^2_{\alpha/2}(n-1)$ das alpha/2-Quantil der Chi-Quadrat-Verteilung mit n-1 Freiheitsgraden (n=Zahl der Untersuchungseinheiten); analoges gilt für den Nenner der ersten Formel.

Die V. kann sinnvollerweise nur bei metrischen Daten (siehe Messniveau) berechnet werden. Jedoch lassen sich Abwandlungen davon als allgemeine Ausdrücke der "Variabilität" von Daten auch auf andere Daten anwenden.

Standardabweichung (engl.: Standard Deviation)

Die S. ist die Wurzel aus der Varianz eines Datenbündels. Durch das Wurzelziehen wird die Quadrierung der Abweichungen "rückgängig gemacht", so daß die S. die gleiche Maßeinheit hat wie die Datenwerte selbst.

Wie bei der Varianz ist zu unterscheiden zwischen der S., die die gegebenen Daten charakterisiert und der S., die aus Stichprobendaten als Schätzwert für die Grundgesamtheit berechnet wird. Es gilt also:

Standardabweichung, die gegebene Daten charakterisiert:

$$s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Standardabweichung als Schätzer für die Grundgesamtheit:

$$S \text{ (oder } \hat{\sigma}) = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Signifikanz (engl.: Significance), Signifikanztest (Test of Significance), Signifikanzniveau (Significance Level)

S. steht in der empirischen Sozialforschung im allgemeinen für *statistische* Signifikanz und bezieht sich auf das Problem des Schlusses von einer (Zufalls-)Stichprobe auf die Grundgesamtheit. 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 (Irrtums-)Wahrscheinlichkeit. Diese bezeichnet man in diesem Kontext als *Signifikanzniveau*. In den Sozialwissenschaften übliche Signifikanzniveaus sind 0,05, 0,01 und 0,001, also 5, 1 oder 0,1 Prozent. 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, wenn der vermutete Zusammenhang "in Wahrheit", d. h. in der Grundgesamtheit, gar nicht besteht.

Allgemein liegt dem S.-test eine Forschungshypothese zugrunde, die sich auf einen Zusammenhang, einen Unterschied oder einen Einfluss (etwa in Form eines Koeffizienten einer Regressionsgleichung) bezieht. Diese wird mit einer Nullhypothese konfrontiert, welche meist besagt, dass *kein* Zusammenhang/Unterschied/Einfluss besteht; es sind aber auch Nullhypothesen der Art denkbar, dass der Zusammenhang usw. einen bestimmten Betrag nicht überschreitet. (Die Nullhypothese wird oft als H_0 bezeichnet, die Forschungshypothese – auch Alternativhypothese genannt – als H_1). Es wird nun eine Teststatistik berechnet, deren Art von der Fragestellung und der Art der vorliegenden Daten (vor allem deren Messniveau) abhängt. Diese Teststatistik stellt eine Zufallsvariable dar und gibt die Wahrscheinlichkeit an, das Stichprobenergebnis zu erhalten, wenn in der Grundgesamtheit die Nullhypothese gilt. Je nach gewähltem Signifikanzniveau ergibt sich ein kritischer Wert für diese Teststatistik, der dem entsprechenden Quantil der Verteilung der betreffenden Zufallsvariablen entspricht (ist z. B. ein Signifikanzniveau von 5 Prozent oder 0,05 gewählt worden, so trennt der kritische Wert die 95 Prozent unter der Nullhypothese wahrscheinlichsten von den 5 Prozent am wenigsten wahrscheinlichen Werten ab). Ist die aus den Daten errechnete Teststatistik (gegebenenfalls: im Absolutbetrag) größer als dieser kritische Wert (man sagt dann auch: sie liegt im Ablehnungsbereich), so wird die Nullhypothese verworfen, andernfalls wird sie (bis auf weiteres) beibehalten. Genauer ist hier zwischen einseitigen und zweiseitige Hypothesen zu unterscheiden. Im ersteren Fall trennt der kritische Wert typischerweise (bei einem gewählten Signifikanznviveau von α Prozent) die untersten $100-\alpha$ von den obersten α Prozent der Verteilung ab bzw. im Falle negativer Unterschiede/Zusammenhänge die untersten α von den obersten $100-\alpha$ Prozent; im zweiteren Fall gibt es zwei kritische Werte, nämlich einen beim $\alpha/2$-Quantil und einem beim $100-(\alpha/2)$-Quantil.

Häufig verwendete S.-tests sind z.B. der t-Test, der F-Test der Varianzanalyse oder der Chi-Quadrat-Test für Kreuztabellen.

Ob ein statistischer Test signifikant ausfällt oder nicht, hängt neben dem Signifikanzniveau (der Irrtumswahrscheinlichkeit) vor allem von der Größe der Stichprobe ab. Mit zunehmender Größe lassen sich auch kleine und unbedeutende Zusammenhänge oder Unterschiede als signifikant absichern. Ein signifikantes (Test-)Ergebnis kann daher nicht ohne nähere Prüfung mit einem wichtigen (Forschungs-)Ergebnis gleichgesetzt werden.

Man beachte, dass der Begriff "Signifikanztest" in der statistischen Literatur nicht ganz einheitlich gebraucht wird. Die hier verwendete Erläuterung entspricht etwa dem Sprachgebrauch bei Fahrmeir et al. oder Hartung et al. Kühnel & Krebs hingegen reservieren den Begriff nur für Tests, bei denen die Nullhypothese lautet, dass in der Grundgesamtheit *kein* Zusammenhang (oder Unterschied) besteht, die Alternativhypothese hingegen, dass der Zusammenhang (oder Unterschied) von Null verschieden ist. (Tatsächlich könnten Nullhypothesen etwa auch lauten, dass der Zusammenhang/Unterschied einen bestimmten Betrag nicht überschreitet; im Unterschied zu Kühnel & Krebs würden im hier verwendeten Sprachgebrauch auch solche Tests als S.-test bezeichnet). Daly et al. schließlich sprechen von Signifikanztest, wenn anhand der errechneten Teststatistik das "empirische Signifikanzniveau", d. h. die Wahrscheinlichkeit, bei Gültigkeit der Nullhypothese einen Testwert der errechneten Größe zu erhalten, bestimmt wird. (Das hier in Übereinstimmung mit der übrigen Literatur skizzierte Vorgehen, eine Nullhypothese zu verwerfen, wenn die vorher festgelegte Irrtumswahrscheinlichkeit bzw. das Signifikanzniveau unterschritten wird, wird dort als "fixed-level testing" bezeichnet).

Siehe auch: Inferenzstatistik, Konfidenzintervall.

Chi² (Chi-Quadrat)-Verteilung (engl.: Chi-Square Distribution)

Verteilung einer Zufallsvariablen, die für Signifikanztests und zur Berechnung von Konfidenzintervallen eingesetzt werden kann. Sie entsteht durch die Summierung von n quadrierten standardnormalverteilten Zufallsvariablen (sog. z-Variablen). Es gibt also viele verschiedene Chi-Quadratverteilungen, die sich durch die Zahl der aufsummierten z-Variablen unterscheiden. Man spricht in diesem Zusammenhang von den Freiheitsgraden der Chi²-Verteilung; durch Summierung von drei z-Variablen entsteht beispielsweise eine Chi²-Verteilung mit drei Freiheitsgraden. Die Werte von Chi²-Verteilungen mit unterschiedlichen Freiheitsgraden sind in allen Statistiklehrbüchern tabelliert.

Die wichtigsten Anwendungen sind:

1. Test auf Überzufälligkeit von Zusammenhängen in Kreuztabellen.

Gegeben sei eine Kreuztabelle mit I Zeilen und J Spalten. Es wird folgende Teststatistik berechnet:

$$\chi^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(n_{ji} - e_{ij})^2}{e_{ij}}$$

Dabei ist n_{ij} die beobachtete (absolute) Häufigkeit der Zelle in der i-ten Zeile und j-ten Spalte der Tabelle und e_{ij} die unter der Nullhypothese eines nicht vorhandenen Zusammenhangs erwartete (absolute) Häufigkeit in der entsprechenden Zelle. Es wird also in jeder Zelle die Differenz der beobachteten und erwarteten Häufigkeiten gebildet, das Ergebnis wird quadriert und wiederum durch die erwarteten Häufigkeiten dividiert. Die Werte der einzelnen Zellen werden über alle Spalten und alle Zeilen aufsummiert. Die resultierende Teststatistik, oft auch als Pearsons Chi² bezeichnet, hat (I-1)(J-1) Freiheitsgrade.

Die erwarteten Häufigkeiten werden nach folgender Formel ermittelt:

$$e_{ij} = \frac{n_{i\bullet} \cdot n_{\bullet j}}{n}$$

Hier ist $n_{i\bullet}$ die absolute Häufigkeit in der i-ten Zeile über alle Spalten (also die Randhäufigkeit der i-ten Zeile), $n_{\bullet j}$ die absolute Häufigkeit in der j-ten Spalte über alle Zeilen (also die Randhäufigkeit der j-ten Spalte) und n der Gesamtumfang der Stichprobe.

In der Beispielstabelle beim Stichwort Kreuztabelle ergibt sich ein Chi² von 135,58 bei 4 Freiheitsgraden. Der kritische Wert einer Chi-Quadrat-Verteilung mit 4 Freiheitsgraden bei einem Signifikanzniveau von 5 Prozent beträgt 9,49; er wird von der Teststatistik deutlich überschritten. 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. überzufällig großen Zellhäufigkeiten zeigen oder nicht, sie besagt aber nichts darüber, welche dies sind. In solchen Fällen 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.

2. Likelihood-Verhältnis-Test

Es sei *abschließend* noch einmal darauf hingewiesen, dass die Einsatzmöglichkeiten des Chi-Quadrat-Tests bzw. allgemein der Chi-Quadrat-Verteilung wesentlich vielfältiger sind und sich keineswegs auf die genannten zwei Möglichkeiten beschränken. Nicht zuletzt können beobachtete Häufigkeiten diskreter Variablen im Prinzip gegen beliebige erwartete Werte getestet werden.

t-Test (engl.: t-Test)

Ein Verfahren zur statistischen Hypothesenprüfung, bei dem geprüft wird, ob eine Teststatistik im Ablehnungsbereich der T-Verteilung liegt.

Das wichtigste Anwendungsgebiet ist der Vergleich der Mittelwerte zweier Gruppen (Stichproben). Dabei werden zwei Fälle unterschieden:

1. Mittelwertvergleich bei unabhängigen Stichproben

Da die Varianzen der beiden Grundgesamtheiten, aus denen die beiden Stichproben gezogen wurden, meistens unbekannt sind, werden die Varianzen aus der Stichprobe geschätzt. Sind die Varianzen der beiden Untergruppen gleich, so wird folgende Formel herangezogen:

$$T = \frac{\bar{x} - \bar{y} - \mu}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \frac{(n_1 - 1) \cdot \hat{\sigma}_1^2 + (n_2 - 1) \cdot \hat{\sigma}_2^2}{n_1 + n_2 - 2}}},$$

andernfalls

$$T = \frac{\bar{x} - \bar{y} - \mu}{\sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}}}.$$

Dabei ist

$\bar{x}$ das arithmetische Mittel der ersten Gruppe,

$\bar{y}$ das arithmetische Mittel der zweiten Gruppe,

n_1 der Stichprobenumfang der ersten Gruppe,

n_2 der Stichprobenumfang der zweiten Gruppe,

$\hat{\sigma}_1^2$ die Varianz (Schätzwert für die Grundgesamtheit) der ersten Gruppe,

$\hat{\sigma}_2^2$ die Varianz (Schätzwert für die Grundgesamtheit) der zweiten Gruppe, und

μ eine Größe, deren Betrag der zu prüfenden Nullhypothese entspricht. Diese Größe wird meist gleich Null gesetzt, d.h. es wird geprüft, ob die Differenz der beiden Mittelwerte größer als Null ist.

Die Teststatistik T folgt einer t-Verteilung mit k Freiheitsgraden. Im Falle gleicher Varianzen hat k den Betrag $n_1 + n_2 - 2$. Bei ungleichen Varianzen wird k geschätzt durch die Welch-Satterthwaite-Approximation (das Ergebnis ist gegebenenfalls auf die nächste ganze Zahl zu runden:

$$k = \frac{\left(\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{\hat{\sigma}_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{\hat{\sigma}_2^2}{n_2}\right)^2}$$

Die Gleichheit der beiden Varianzen kann z.B. durch den Levene-Test geprüft werden. Wird die Nullhypothese der Varianzgleichheit nicht abgelehnt, so ist die erste der oben angeführten Formeln für den t-Test zu verwenden, andernfalls die zweite.

Hinweis: Sind die Varianzen σ der Grundgesamtheiten bekannt, aus denen die beiden Gruppen stammen, so ist die Teststatistik

$$Z = \frac{\bar{x} - \bar{y} - \mu}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

normalverteilt.

2. Mittelwertvergleich bei abhängigen (oder verbundenen) Stichproben

Dieser Fall liegt z.B. vor, wenn bei den Untersuchungseinheiten einer Stichprobe ein Merkmal zweimal gemessen und geprüft wird, ob der zweite Messwert höher (oder niedriger) als der erste liegt, oder wenn bei Ehepaaren untersucht wird, ob die Einkommen der Frauen unter denen der Männer liegen (bzw. ob die Differenz einen Betrag μ überschreitet). Die T-Statistik wird hier berechnet als

$$T = \frac{\bar{x}_d - \mu_d}{\frac{s_d}{\sqrt{n}}}$$,

mit

$\bar{x}_d$ als dem arithmetischen Mittel der Differenzen der Messwerte der einzelnen Paare, und

s_d als der Standardabweichung dieser Differenzen, die gemäß der Formel

$$\sqrt{\frac{\sum_{i-1}^{n}(d_i - \bar{x}_d)^2}{n-1}}$$

berechnet wird. (d_i sind hier die Differenzen der Messwerte der einzelnen Paare.)

Die Größe T folgt einer t-Verteilung mit n-1 Freiheitsgraden; n ist hier die Anzahl der Messwertpaare (also nicht der einzelnen Messungen!).

Anwendungsvoraussetzungen: Der t-Test kann eingesetzt werden, wenn die zu untersuchende abhängige Variable mindestens (mehr oder weniger) intervallskaliert ist (siehe Messniveau). Außerdem sollte das Merkmal in den untersuchten Populationen normalverteilt sein; vor allem bei großen Stichproben und nicht zu unterschiedlichem Umfang der beiden Gruppen ist das Verfahren jedoch relativ robust gegen die Verletzung dieser Annahme. Eine genauere Diskussion findet sich bei Bortz.

Korrelation und Assoziation (engl.: Correlation, Association)

Mit K. und A. werden durch statistische Kennzahlen ausdrückbare Zusammenhänge zwischen zwei Variablen bezeichnet. Zumeist wird der Begriff "Korrelation" auf die Bezeichnung des Zusammenhangs zweier metrischer (siehe Messniveau) Merkmale beschränkt und der Begriff "Assoziation" zur Kennzeichnung anderer Zusammenhänge (etwa zwischen nominal- und/oder ordinalskalierten Merkmalen) verwendet; in der Forschungspraxis wird diese Terminologie nicht immer strikt eingehalten. Werden Zusammenhänge zwischen zwei Variablen unter Berücksichtigung ihrer Zusammenhänge mit weiteren Variablen berechnet, so spricht man von Partialkorrelation oder partieller Korrelation.

Maße für die Stärke der K. werden meist als Korrelationskoeffizienten bezeichnet. Im allgemeinen können diese Werte zwischen minimal -1 und maximal +1 annehmen, wobei -1 einen perfekten negativen ("je größer X, desto kleiner Y") und +1 einen perfekten positiven ("je größer X, desto größer Y") Zusammenhang bezeichnet. Unter den Assoziationsmaßen (also Maßzahlen für die Stärke des Zusammenhangs vor allem zwischen nominal- oder ordinalskalierten Variablen) finden sich gelegentlich auch solche, die die Richtung des Zusammenhangs nicht durch unterschiedliche Vorzeichen ausdrücken, oder die nicht das Maximum von +1 bzw. das Minimum von -1 erreichen können.

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

Die Wahl des Korrelationskoeffizienten bzw. Assoziationsmaßes hängt vom Messniveau der Variablen ab. Die wichtigste Koeffizienten 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, Spearmans Korrelationskoeffizient (auch Spearmans Rho genannt) oder u. U. der ->polychorische Korrelationskoeffizient (bei dem angenommen wird, dass den ordinal gemessenen Merkmalen latente metrische und normalverteilte Merkmale zugrunde liegen).
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.

Faktorenanalyse (engl.: Factor Analysis)

Überblick

Die F. ist ein Verfahren der Datenreduktion. Die grundlegende Idee ist folgende: Es wird angenommen, dass hinter einer Reihe von Messwerten – z. B. Ergebnissen eines psychometrischen Tests oder einer Befragung zu Meinungen über bestimmte Sachverhalte – eine grundlegende, nicht direkt messbare, hypothetische (oder auch latente) Variable steht, etwa eine Eigenschaft oder eine Einstellung (Attitüde). Eine solche hypothetische Variable wird als »Faktor« bezeichnet. Wenn angenommen wird, dass die erzielten Messergebnisse auf einen einzigen Faktor zurückgehen, so bedeutet das, dass die betreffenden Variablen untereinander in hohem Maße korrelieren müssten. Die entsprechende Korrelationsmatrix ist daher Ausgangspunkt der F. Häufig wird die F. auch eingesetzt, wenn man annimmt, dass eine Serie von Messwerten verschiedene Variablen repräsentiert; dann soll die F. festlegen, welcher Messwert zu welchem Faktor gehört bzw. entsprechende Vorab-Hypothesen testen.

Die *explorative F.* versucht, aus der Korrelationsmatrix Faktoren zu »extrahieren«; diese Faktoren sollen voneinander unabhängig sein. Nach bestimmten Verfahren werden solange Faktoren ermittelt, bis ein bestimmtes Stop-Kriterium erreicht ist (es stehen mehrere solcher Kriterien zur Verfügung), nach dem die Annahme weiterer Faktoren keinen Erklärungsgewinn mehr verspricht. Im Allgemeinen sucht man anschließend die Faktoren so mit den Messwerten »abzugleichen«, dass alle Messwerte mit einem der Faktoren sehr hoch zusammenhängen (oder hoch auf ihm »laden«) und mit allen anderen nicht oder nur äußerst niedrig. Zu diesem Zweck werden in einem zweiten Schritt die ermittelten Faktoren »rotiert« (*Faktorenrotation*). Die so ermittelten Faktoren müssen nunmehr »interpretiert« werden, d. h. man inspiziert die Zusammenhänge zwischen den einzelnen Messwerten und den Faktoren darauf, ob sich sinnvolle Ergebnisse gezeigt haben. Diese substanzwissenschaftliche Interpretation ist sehr wichtig, da eine Faktorenanalyse nahezu immer im rein mathematisch-statistischen Sinn »brauchbare« Ergebnisse liefert.

Die *konfirmatorische* (manchmal auch: konfirmative) F.hingegen legt vorab Hypothesen über die Faktorstruktur fest und prüft dann, ob diese Hypothesen verworfen werden müssen oder beibehalten werden können. Die konfirmatorische F. wird im allgemeinem im Rahmen von Strukturgleichungsmodellen durchgeführt.

Näheres zur explorativen Faktorenanalyse

Formal lässt sich die Idee der Faktorenanalyse wie folgt ausdrücken: Die Messwerte ergeben sich als Linearkombination aus den zugrundeliegenden Faktoren. Ganz ähnlich wie in einer Regressionsgleichung werden die Messwerte also auf andere Variablen zurückgeführt, nur dass diese bei der F. unbeobachtete, hypothetische Konstrukte sind:

$z_{mi} = a_{i1} \cdot f_{m1} + a_{i2} \cdot f_{m2} + ... + a_{ij} \cdot f_{mj}$

Das bedeutet: Für jede von m Personen liegen i Messwerte vor (die z_{mi}; es handelt sich um standardisierte Variablen). Diese Messwerte lassen sich erklären aus den »Faktorwerten« der Personen (den f_{mj}, d. h. den Werten, den sie in insgesamt j verschiedenen unbeobachteten Merkmalen [den Faktoren] aufweisen), die mit den »Gewichten« dieser Faktoren für das jeweilige Merkmal, den a_{ij}, multipliziert werden. Diese »Gewichte« heißen *Faktorladungen*. Für jede Variable existieren also so viele Faktorladungen, wie es Faktoren gibt; insgesamt gibt es also bei i Variablen und j Faktoren i · j Faktorladungen. Die Summen aller quadrierten Faktorladungen pro Variablen heißen Kommunalitäten; diese geben an, wie gut die manifesten Messwerte durch die hypothetischen Faktoren erklärt werden (bei einer Kommunalität von 1 lassen sich die Messwerte perfekt durch die Faktoren erklären).

Von den Kommunalitäten sind die *Eigenwerte* der Werte der Faktoren zu unterscheiden. Diese geben an, wie viel Varianz in den Daten *insgesamt* (also nicht nur in einer einzelnen Variablen) durch den jeweiligen Faktor erklärt werden kann. Die Summe der Eigenwerte entspricht der Anzahl der Variablen, denn da es sich um standardisierte Variablen handelt, hat jede Variable die Varianz 1. Der Eigenwert eines jeden Faktors ist die Summe aller quadrierten Faktorladungen, die zu ihm gehören. (Man beachte also: Die Summe der quadrierten Faktorladungen über jede Variable ergibt die Kommunalität der Variablen; die Summe dieser Ladungen über den Faktor ergibt dessen Eigenwert.)

Vor der eigentlichen Durchführung der Faktorenanalyse ist es sinnvoll, die Korrelationsmatrix der Variablen, die in die Analyse eingehen, auf ihre *Eignung für die Faktorenanalyse* zu prüfen. Gemeint ist damit folgendes: Wenn die erhobenen Merkmale allesamt gar nicht oder nur schwach untereinander zusammenhängen, so ist es unwahrscheinlich, dass sich Faktoren finden lassen, durch die sich die Vielzahl der Variablen auf einer geringere Zahl von Dimensionen (eben die Faktoren) reduzieren lässt. Für die Prüfung der Eignung der Korrelationsmatrix wurden mehrere Verfahren vorgeschlagen: Der *Bartlett-Test auf Sphärizität* prüft, ob die Daten aus einer Grundgesamtheit stammen, in der die Variablen untereinander allesamt unkorreliert sind; da dieser Test bei größeren Stichproben fast immer zu einem positiven (d. h. signifikanten) Ergebnis führt (das liegt in der Natur großer Stichproben), gilt er nur bei kleinen Fallzahlen als sinnvoll. Besser ist eine Betrachtung der »Anti-Image-Korrelationsmatrix«. Diese gibt an, wie groß der Anteil der Varianz der einzelnen Variablen ist, der sich *nicht* durch die anderen Variablen erklären lässt. Die Diagonalelemente dieser Matrix sollten möglichst groß sein (Variablen mit Werten unter 0,5 gelten als gänzlich ungeeignet, ab 0,6 gelten sie als brauchbar, über 0,8 als recht gut); auch ein Gesamtmaß – das Kaiser-Meyer-Olkin-Kriterium der Stichprobeneignung – wird hieraus abgeleitet, für das das die gleichen Zahlenwerte als Eignungskriterien gelten.

Für die Extraktion der Faktoren wurden unterschiedliche Verfahren entwickelt; ebenso gibt es verschiedene Kriterien für die Zahl der zu extrahierenden Faktoren. Unter den Extraktionsverfahren dürften die wichtigsten die Hauptkomponenten-Methode und die Hauptfaktoren- (oder Hauptachsen)-Methode sein. Der wesentliche Unterschied ist in folgendem zu sehen: Man wird üblicherweise unterstellen können bzw. müssen, dass die Faktoren nicht komplett die Varianz der einzelnen Variablen erklären können, sondern dass ein unerklärter Rest an Varianz bleibt (Einzelrestvarianz, »spezifische Varianz«). Die Hauptkomponenten-Methode interessiert sich aber nicht für letztere; daher führt sie dann zu Kommunalitäten von 1 (bzw. zu einer erklärten Gesamtvarianz von 100 Prozent), wenn die Zahl der Faktoren der Zahl der Variablen entspricht. Die Hauptachsenmethode berücksichtigt dagegen explizit, dass unerklärte spezifische Varianz vorhanden ist, dass also die einzelnen Variablen nicht komplett aufeinander bzw. die zugrundeliegenden Faktoren zurückgeführt werden können. Daher führt sie im allgemeinen zu niedrigeren Kommunalitäten bzw. Faktorladungen. Allerdings ist jedenfalls dann, wenn den beobachteten Daten wirklich klare Faktoren entsprechen, die grundsätzliche Faktorenlösung (welche Items laden hoch auf welchem Faktor?) bei beiden Verfahren in der Regel identisch (eine ausführlichere Diskussion mit dem gleichen Ergebnis findet sich bei Velicer/Jackson 1990).

Für alle Verfahren der Faktorenextraktion gilt: Der erste extrahierte Faktor soll einen so großen Varianzanteil wie möglich erklären. Der zweite Faktor (*falls* mehr als ein Faktor extrahiert wird) ist von dem ersten Faktor völlig unabhängig (in der Fachterminologie: er ist orthogonal zum ersten Faktor); er erklärt den maximalen Anteil der Restvarianz. Soweit weitere Faktoren zu extrahieren sind, erfüllen diese wiederum die Bedingung der Orthogonalität zu den übrigen Faktoren und der Extraktion von so viel Varianz wie jeweils möglich.

Die *Zahl der zu extrahierenden Faktoren* liegt bei der explorativen Faktorenanalyse nicht a priori fest. Aus dem Vorstehenden wird klar, dass irgendwann einmal die Restvarianz, die ein weiterer Faktor erklärt, so klein ist, dass sich die Extraktion dieses Faktors nicht mehr »lohnt«. Meist wird als Kriterium herangezogen, dass der Eigenwert eines jeden Faktors größer als 1 sein soll (Eigenwerte kleiner oder gleich 1 bedeuten, dass ein Faktor keinen größeren Varianzanteil als eine einzelne Variable erklärt). Allerdings kann unter Umständen schon dieses Kriterium zu einer zu großen Anzahl von Faktoren führen; häufig wird daher der sog. Scree-Plot herangezogen, in dem die Eigenwerte der Faktoren als Liniendiagramm abgetragen werden. Meist ist in diesem Diagramm ein deutlicher Knick zu sehen, was bedeutet, dass die Faktoren ab diesem Knick keinen nennenswerten Erklärungsbeitrag liefern. Ist der Knick also beispielsweise beim vierten Faktor zu sehen, so dürfte die beste Lösung bei drei Faktoren liegen. Letztlich müssen aber auch immer Kriterien der inhaltlichen bzw. theoretischen Adäquatheit herangezogen werden.

Wie schon geschildert, werden die Faktoren jeweils so extrahiert, dass sie möglichst viel (Rest-)Varianz in *allen* Merkmalen erklären. Man ist aber an Faktoren interessiert, die möglichst unabhängig voneinander sind (man sagt auch, dass die Faktoren orthogonal – also rechtwinklig – zueinander sein sollten); einzelne Variablen sollen hoch auf einem Faktor und möglichst niedrig auf den anderen Faktoren laden. Dazu wird die durch die Extraktion gewonnene Faktorenlösung »rotiert« (dieser Begriff geht auf die geometrische Veranschaulichung von Korrelationen zurück, die hier nicht dargestellt werden kann). Auch für diese Rotation gibt es verschiedene Kriterien, u. a. solche, die auch ein gewisses Ausmaß an Korrelation zwischen den Faktoren zulassen (»schiefwinklige« oder »oblique Rotation«). Die Qualität des Ergebnisses einer Faktorenanalyse kann man auch daran erkennen, dass sie für jede Variable eine hohe Ladung (Absolutbetrag mindestens 0,6) auf einem Faktor und niedrige Ladungen (Absolutbetrag möglichst unter 0,1) auf allen anderen Faktoren ergibt (sog. »Einfachstruktur«). (Stellt sich dieses Ergebnis nicht ein, so liegt dies meist nicht an der Faktorenanalyse, sondern daran, dass die Daten etwas komplexer als vermutet sind.)

Den letzten Schritt der F. stellt die Interpretation der gewonnenen Faktoren dar. Die Tatsache, dass bestimmte Merkmale zu einem gemeinsamen Faktor gehören, muss durch inhaltliche Überlegungen plausibel gemacht werden können. (Würde man beispielsweise annehmen, dass der Spruch »Männer sind Schweine« sich auch als Eigenschaftspaar ausdrücken lässt, so sollte die »Männlichkeit« von Untersuchungspersonen mit ihrer »Schweinlichkeit« hoch zusammenhängen und beide Merkmale sollten dementsprechend in einer F. auch auf einem gemeinsamen Faktor laden.)

Das Fundamentaltheorem der Faktorenanalyse

Wie können die Gewichte a_{ij}, die sog. »Faktorladungen«, bestimmt werden? Das Fundamentaltheorem der Faktorenanalyse beschreibt die zentrale Bedingung, der die Faktorladungen entsprechen müssen. Diese lässt sich einfach und übersichtlich allerdings nur in Matrixschreibweise darstellen.

Wir gehen aus von der Matrix der Korrelationen zwischen den Variablen. Hierbei handelt es sich bekanntlich um die standardisierten Kovarianzen; alternativ lassen sich die Korrelationen als die Kovarianzen der standardisierten Variablen schreiben. In Matrixschreibweise heißt dies:

$$\mathbf{R} = \frac{1}{N-1}(\mathbf{Z}'\mathbf{Z})$$

Die obige Grundgleichung der Faktorenanalyse lautet in Matrixschreibweise:

$$\mathbf{Z} = \mathbf{F} \cdot \mathbf{A}'$$

Setzen wir die rechte Seite dieses Ausdrucks in die vorherige Formel ein, so erhalten wir

$$\mathbf{R} = \frac{1}{N-1}(\mathbf{F} \cdot \mathbf{A}')' \cdot (\mathbf{F} \cdot \mathbf{A}')$$

Löst man die Klammern auf, so erhält man den Ausdruck

$$\mathbf{R} = \frac{1}{N-1} \cdot \mathbf{A} \cdot \mathbf{F}' \cdot \mathbf{F} \cdot \mathbf{A}' = \mathbf{A} \cdot \frac{1}{N-1} \cdot \mathbf{F}' \cdot \mathbf{F} \cdot \mathbf{A}'$$

Der Ausdruck in der Mitte

$$\frac{1}{N-1} \cdot \mathbf{F}' \cdot \mathbf{F}$$

ist nichts anderes als die Matrix der Korrelationen zwischen den Faktorwerten (diese sind ihrerseits standardisierte Werte mit Mittelwert 0 und Standardabweichung 1). Da die Faktoren untereinander unabhängig sind, entspricht diese Matrix aber einer Einheitsmatrix. Damit wird der vorstehende Ausdruck zu

$$R = A \cdot I \cdot A'$$

was (da Multiplikation mit der Einheitsmatrix der Multiplikation mit 1 in der »normalen« Mathematik entspricht) zu

$$R = A \cdot A'$$

führt.

Damit haben wir nun die Bedingung gefunden, der die Faktorladungen a_{ij} entsprechen müssen: Die Matrix der Ladungen A, multipliziert mit ihrer Transponierten, ergibt die Korrelationsmatrix der Variablen, die in die Faktorenanalyse eingingen. Natürlich wird diese Reproduktion der Korrelationsmatrix durch die Faktorladungen nicht ganz exakt sein (genauer gesagt: Sie ist es, wenn ebenso viele Faktoren extrahiert werden, wie Variablen vorhanden sind – aber dann ist die Faktorenanalyse überflüssig); aber es geht in der Faktorenanalyse auch nur darum, die Zusammenhänge der Variablen *im Wesentlichen* zu reproduzieren.

Die konkreten Verfahren, nach denen nun **A** bestimmt werden kann, sind komplex und liegen jenseits der Aufgaben dieses Artikels. Die modernen Extraktionsverfahren sind aber ohnehin nur mit Hilfe einschlägiger Statistiksoftware zu bewältigen.

Clusteranalyse (engl.: Cluster Analysis)

Unter C. versteht man eine Gruppe von Verfahren, Fälle (Untersuchungsobjekte) hinsichtlich ihrer Ähnlichkeit einzustufen und dann Gruppen (Cluster) von untereinander möglichst ähnlichen Fällen zu bilden.

Die Ähnlichkeit der Fälle wird durch Distanzmaße gemessen. Aus der Distanzmatrix werden dann die Cluster gebildet, wobei ebenso wie schon bei den Distanzmaßen eine Auswahl aus einer relativ großen Zahl von Verfahren getroffen werden muss. Die Verfahren sind teils partitionierend (sie gehen aus von der Gesamtheit der Fälle und bilden möglichst homogene Untergruppen), häufiger werden jedoch agglomerierende Verfahren verwendet, bei denen ausgehend von den einzelnen Fällen zunehmend ähnliche Fälle 'angelagert' werden, bis am Schluss einige wenige Cluster übrig bleiben. An dieser Stelle ist im allgemeinen noch festzulegen, wieviele Cluster man als 'sinnvolle' Einteilung akzeptiert.

Die Entscheidungen über die Wahl des Distanzmaßes, des Clusterungsverfahrens und der Zahl der Cluster sollte möglichst unter theoretischen Gesichtspunkten getroffen werden. Faktisch ist zumeist (mit Ausnahme vielleicht der Distanzmaße) das theoretische Vorwissen nicht ausreichend, so dass häufig ad hoc argumentiert werden muss. Dennoch haben sich C.n zu Recht als wichtiges explorativ es Verfahren etabliert.