Grundbegriffe der Statistik

Quelle:

Statistica (2003). ***Auszug aus dem elektronischen Handbuch des Statistikprogramms Statistica 6.1*****. Tula, OK: StatSoft, Inc.**

Inhaltsverzeichnis

Überblick über Grundbegriffe in der Statistik

In dieser Einführung werden kurz die für weitergehende Untersuchungen auf jedem Gebiet statistischer Datenanalyse notwendigen Grundbegriffe erläutert. Die ausgewählten Themen veranschaulichen die grundlegenden Annahmen der meisten statistischen Methoden und/oder haben sich in der Forschung als notwendig für ein allgemeines Verständnis der "quantitativen Natur" der Wirklichkeit erwiesen. (Nisbett u.a. 1987).

Aus Platzgründen werden hier schwerpunktmäßig die funktionellen Aspekte erläutert. Die Darstellung wird außerdem sehr knapp erfolgen. Weitere Informationen zu jedem dieser Begriffe enthalten die Abschnitte Introductory Overview und Examples des englischen Handbuchs und statistische Lehrbücher.

Variablen

Variablen sind diejenigen Gegenstände in der Forschung, die gemessen, kontrolliert oder manipuliert werden können. Sie unterscheiden sich in vielerlei Hinsicht, vor allem in der Rolle, die sie für ein Forschungsvorhaben spielen, und der Art ihrer Messung.

Korrelationsforschung und experimentelle Forschung.

Die meisten empirischen Untersuchungen können in eine dieser beiden Kategorien eingeordnet werden: In Korrelationsuntersuchungen werden keine Variablen beeinflusst; zumindest wird versucht, dies zu vermeiden. Die Variablen werden nur gemessen und es wird nach Korrelationen zwischen Variablen, wie z.B. Blutdruck und Cholesterinspiegel, gesucht. In experimentellen Untersuchungen werden einige Variablen manipuliert und dann die Wirkung auf andere Variablen gemessen: Der Blutdruck wird z.B. künstlich erhöht und dann der Cholesterinspiegel aufgezeichnet. Auch in der experimentellen Forschung werden "Korrelationen" zwischen Variablen (manipulierte Variablen und die durch diese Manipulation betroffenen Variablen) berechnet. Experimentell erhobene Daten besitzen jedoch einen qualitativ höherwertigen Informationsgehalt: Nur durch sie können kausale Beziehungen zwischen Variablen aufgezeigt werden. Wird z.B. festgestellt, dass jede Änderung von Variable A eine Änderung von Variable B hervorruft, kann gefolgert werden, dass B durch A beeinflusst wird. Daten aus Korrelationsuntersuchungen können nur bei Vorliegen theoretischer Anhaltspunkte kausal interpretiert werden. Korrelationsuntersuchungen können letztendlich jedoch keine Kausalität beweisen.

Abhängige und unabhängige Variablen.

Unabhängige Variablen sind diejenigen, die manipuliert, d.h. beeinflusst werden können. Abhängige Variablen werden dagegen nur gemessen. Diese Begriffsunterscheidung erscheint zunächst verwirrend, da oft behauptet wird, dass "alle Variablen von irgend etwas abhängen". Diese Unterscheidung ist jedoch außerordentlich wichtig: Die Begriffe abhängig und unabhängig werden meistens in experimentellen Untersuchungen gebraucht. Dort werden einige Variablen manipuliert. In diesem Sinne sind sie "unabhängig" von Reaktionsmuster, Eigenschaften, Absichten, usw. der untersuchten Objekte. Bei anderen Variablen wird dagegen

vermutet, dass sie von dieser Manipulation oder den Versuchsbedingungen "abhängig" sind. Sie beschreiben sozusagen die Reaktion des Objekts. Diese Begriffe werden auch dann gebraucht, wenn in einer Studie die unabhängigen Variablen nicht im eigentlichen Sinne manipuliert werden, sondern nur Objekte nach ihren Eigenschaften in verschiedene "Versuchsgruppen" eingeteilt werden. Beispielsweise wird in einem Versuch die Anzahl weißer Blutkörperchen (englisch: White Cell Count oder WCC) zwischen Männern und Frauen verglichen. Geschlecht wäre hier die unabhängige, WCC die abhängige Variable.

Skalenniveaus

Variablen unterscheiden sich auch durch den Informationsgehalt ihrer Messung, d.h. ihr "Skalenniveau". Jede Messung ist zunächst mit Messfehlern behaftet. Diese legen den maximalen, überhaupt messbaren Informationsumfang fest. Ein weiterer Bestimmungsfaktor für den Informationsumfang einer Variable ist ihr "Skalenniveau": Variablen werden unterschieden in (a) nominal-, (b) ordinal-, (c) intervall- oder (d) verhältnisskaliert.

(a) Nominalskalierte Variablen ermöglichen nur eine qualitative Klassifikation. Es kann nur festgestellt werden, ob ein einzelnes Objekt in eine Kategorie gehört oder nicht. Diese Kategorien sind jedoch nicht quantifizierbar oder durch eine natürliche Reihenfolge gekennzeichnet. Zwei Personen können sich bzgl. einer Variable A unterscheiden (z.B. verschiedene Rassenzugehörigkeit). Es kann jedoch nicht bestimmt werden, welche dieser Personen "mehr" von der Eigenschaft der entsprechenden Variablen besitzt. Typische Beispiele für nominalskalierte Variablen sind Geschlecht, Rasse, Hautfarbe, Stadtzugehörigkeit usw.

(b) Ordinalskalierte Variablen ermöglichen die Bildung einer natürlichen Rangfolge der gemessenen Objekte. Das Mehr oder Weniger an Vorhandensein einer Eigenschaft ist bestimmbar. Es ist jedoch nicht möglich, zu sagen, "um wie viel" mehr diese Eigenschaft vorhanden ist. Ein typisches Beispiel für ordinalskalierte Variablen ist der sozioökonomische Status von Familien. Der Status der oberen Mittelklasse ist höher als derjenige der Mittelklasse. Es ist aber unmöglich zu behaupten, dass er um 18% höher ist. Auch die Unterscheidung zwischen Nominal-, Ordinal- und Intervallskalen selbst ist ein Beispiel für eine ordinalskalierte Variable: Eine Nominalskala misst weniger Information als eine Ordinalskala. Die Differenz des Informationsumfangs im Vergleich zur Differenz des Informationsumfangs zwischen Ordinal- und Intervallskala ist jedoch nicht quantifizierbar.

(c) Intervallskalierte Variablen ermöglichen nicht nur die Bildung einer Reihenfolge für die gemessenen Objekte, sondern auch eine Quantifizierung und einen Vergleich der Differenz zwischen den Objekten. Die Temperaturmessung in Grad Fahrenheit oder Celsius ist ein Beispiel für eine Intervallskala: Eine Temperatur von 40 Grad ist höher als eine von 30 Grad. Ein Temperaturanstieg von 20 auf 40 Grad ist doppelt so hoch wie ein Anstieg von 30 auf 40 Grad.

(d) Verhältnisskalierte Variablen besitzen alle Eigenschaften von intervallskalierten Variablen und haben darüber hinaus noch einen absoluten (natürlichen) Nullpunkt. Sie ermöglichen da-

mit Aussagen wie: "x ist das Zweifache von y". Typische Beispiele für Verhältnisskalen sind Raum- und Zeitmessungen. Auch die Kelvin-Temperaturskala ist z.B. verhältnisskaliert: Eine Temperatur von 200 Grad ist nicht nur höher als eine von 100 Grad; sie ist doppelt so hoch. Intervallskalen haben diese Verhältnis- oder Quotienteneigenschaft nicht. Die meisten statistischen Verfahren unterscheiden nicht zwischen Intervall- und Verhältniseigenschaft von Skalenniveaus.

Zusammenhänge zwischen Variablen

Ungeachtet ihres Typs spricht man von einer Beziehung oder einem Zusammenhang zwischen Variablen in einer Stichprobe, wenn die Werte dieser Variablen gleichartig verteilt sind, d.h. ihre Werte für diese Beobachtungen systematisch miteinander korrespondieren. Geschlecht und WWC würden z.B. zusammenhängen, wenn Männer generell höhere WWC-Werte als Frauen hätten oder umgekehrt; Größe und Gewicht hängen zusammen, weil größere Personen i. d. R. schwerer als kleinere sind; Intelligenzquotient (IQ) und Anzahl der Fehler in einem Test hängen dann zusammen, wenn Personen mit höherem IQ weniger Fehler machen.

Bedeutung von Zusammenhängen zwischen Variablen

Das Ziel jeder wissenschaftlichen Analyse oder Forschung ist im allgemeinen die Aufdeckung von Zusammenhängen oder Beziehungen zwischen Variablen. Die (philosophische) Wissenschaftstheorie besagt, dass das "Wesen der Dinge" nur über qualitative und quantitative Beziehungen dargestellt werden kann. Dies beinhaltet auf jeden Fall Zusammenhänge zwischen Variablen. Der Fortschritt der Wissenschaft besteht nun darin, immer neue Beziehungen zwischen Variablen zu finden. Korrelationsuntersuchungen versuchen, solche Beziehungen unmittelbar zu messen. Hierin unterscheiden sie sich allerdings nicht von der experimentellen Forschung. Der o.g. Vergleich der WWC zwischen Männern und Frauen kann als Untersuchung der Korrelation zwischen den beiden Variablen WWC und Geschlecht aufgefasst werden. Die Aufgabe der Statistik besteht darin, bei der Bewertung dieser Zusammenhänge zwischen Variablen behilflich zu sein. Sämtliche im (elektronischen) Handbuch beschriebenen Verfahren können auf die eine oder andere Weise als Bewertung von Zusammenhängen zwischen Variablen aufgefasst werden.

Zwei Haupteigenschaften von Zusammenhängen zwischen Variablen

Die beiden grundlegenden formalen Eigenschaften jedes Zusammenhangs zwischen Variablen sind (a) seine Stärke (oder "Größe") und (b) seine Zuverlässigkeit (oder "Wahrhaftigkeit").

(a) Stärke (oder "Größe"). Die Stärke ist leichter zu verstehen und zu messen als die Zuverlässigkeit. Hätte z.B. jeder Mann in der o.g. Stichprobe einen höheren WCC-Wert als jede Frau, könnte man sagen, dass die Stärke des Zusammenhangs zwischen den beiden Variablen (Geschlecht und WCC) in der Stichprobe sehr groß ist. Man kann die Werte der einen aufgrund der Werte der anderen vorhersagen (zumindest für diese Stichprobe).

(b) Zuverlässigkeit (oder "Wahrhaftigkeit"). Der Begriff der Zuverlässigkeit ist weniger intuitiv, aber dennoch sehr wichtig. Er beschreibt die "Repräsentativität" der Ergebnisse in einer Stichprobe für die Grundgesamtheit. Die Zuverlässigkeit misst die Wahrscheinlichkeit dafür, dass ein ähnlicher Zusammenhang gefunden würde, wenn der Versuch mit weiteren Stichproben aus der gleichen Grundgesamtheit durchgeführt würde. Man ist letztendlich an der Stichprobe nur in dem Maße interessiert, wie sie Informationen über die entsprechende Grundgesamtheit enthält. Sofern eine Studie bestimmte Kriterien erfüllt (siehe weiter unten), kann die in der Stichprobe beobachtete Zuverlässigkeit des Zusammenhangs quantitativ abgeschätzt und mittels einer Kennziffer dargestellt werden. (sogenanntes p-Niveau oder statistisches Signifikanzniveau, siehe "Statistische Signifikanz" (p-Niveau)?).

"Statistische Signifikanz" (p-Niveau)

Die statistische Signifikanz ist ein geschätztes Maß dafür, inwieweit ein gefundenes Ergebnis "wahr" (i. S. v. repräsentativ für die Grundgesamtheit) ist. Das p-Niveau (der Begriff wurde erstmalig von Brownlee, 1960, verwendet.) wird durch einen Index abnehmender Werte für die Zuverlässigkeit eines Ergebnisses dargestellt. Je höher das p-Niveau ist, desto weniger kann man annehmen, dass die Beziehung zwischen den Variablen in der Stichprobe ein zuverlässiger Indikator für den Zusammenhang der entsprechenden Variablen in der Grundgesamtheit ist. Im eigentlichen Sinne misst das p-Niveau die Wahrscheinlichkeit für einen Fehler bei der Akzeptanz eines beobachteten Ergebnisses als gültig (repräsentativ für die Grundgesamtheit). Ein p-Niveau von 0,05 (oder 1/20) besagt z. B., dass die Wahrscheinlichkeit für das Auftreten eines "scheinbaren" oder "falschen" Zusammenhangs in einer Stichprobe 5 % beträgt. Man kann dies auch anders darstellen: Angenommen, in der Grundgesamtheit gäbe es überhaupt keinen Zusammenhang und man würde den o. g. Versuch beliebig oft wiederholen: Bei jeweils einer von 20 Wiederholungen würde man dann in etwa einen gleichstarken oder sogar stärkeren Zusammenhang zwischen den Variablen feststellen. Auf vielen Forschungsgebieten wird ein p-Niveau von 0,05 üblicherweise als "Grenzwert" für das Fehlerniveau akzeptiert.

Bestimmung der "wirklichen" Signifikanz eines Ergebnisses

Bei der Festlegung eines Niveaus, ab dem man von "wirklicher" Signifikanz spricht, lässt sich ein bestimmtes Ausmaß an Willkür nicht vermeiden. Die Auswahl eines konkreten Signifikanzniveaus, bis zu dem man die Ergebnisse als ungültig ablehnt, ist also willkürlich. In der Praxis hängt die Entscheidung darüber gewöhnlich davon ab, ob der Ausgang der Untersuchung a priori vorhergesagt wurde oder nur post hoc ("im nachhinein") im Laufe vieler Analysen und Vergleiche gefunden wurde. Sie hängt u.a. auch vom Ausmaß an unterstützenden, konsistenten Beweisen für den Datensatz und von "Traditionen" im entsprechenden Forschungsbereich ab. In vielen Wissenschaften werden Ergebnisse mit $p < 0,05$ als "gerade noch statistisch signifikant" akzeptiert. Dieser Grenzwert für statistische Signifikanz schließt jedoch immer noch eine ziemlich hohe Fehlerwahrscheinlichkeit (5 %) ein. Ergebnisse mit $p < 0,01$ werden gewöhnlich als statistisch signifikant, solche mit $p< 0,005$ oder $p < 0,001$ als "hoch" signifikant bezeichnet. Diese Einteilungen stellen aber - wie bereits gesagt - nur will-

kürliche Konventionen dar, die auf allgemeinen Erfahrungen aus der Forschungspraxis basieren.

Statistische Signifikanz und Anzahl durchgeführter Analysen

Je mehr Analysen mit einem Datensatz durchgeführt werden, desto mehr Ergebnisse werden tendenziell das vorgeschriebene Signifikanzniveau "zufällig" einhalten. Bei der Berechnung von Korrelationen zwischen 10 Variablen (d.h. 45 verschiedene Korrelationskoeffizienten) werden ca. 2 (d. h. einer von 20) Korrelationskoeffizienten für $p < 0,05$ durch Zufall signifikant sein. Dies passiert selbst dann, wenn die Werte dieser Variablen völlig zufällig ausgewählt wurden und diese Variablen in der Grundgesamtheit unkorreliert sind. Einige statistische Verfahren, die viele Vergleiche beinhalten und damit für solche Fehler sehr anfällig sind, korrigieren dies über die Gesamtanzahl der Vergleiche. Viele statistische Verfahren (insbesondere einfache explorative Datenanalysen) bieten jedoch keine unmittelbare Möglichkeit zur Behebung dieses Problems. Man sollte daher die Zuverlässigkeit unerwarteter Versuchsergebnisse stets vorsichtig bewerten. Viele Beispiele im elektronischen Handbuch geben Hinweise dafür. Auch die meisten Lehrbücher über Forschungsmethoden beinhalten entsprechende Informationen.

Stärke und Zuverlässigkeit eines Zusammenhangs zwischen Variablen

Wie gesagt umschreiben Stärke und Zuverlässigkeit zwei verschiedene Eigenschaften von Zusammenhängen zwischen Variablen. Sie sind jedoch nicht völlig unabhängig voneinander. Allgemein gilt für eine Stichprobe mit gegebenem Umfang: Je stärker der Zusammenhang, desto größer die Zuverlässigkeit des Zusammenhangs (siehe Stärkere Zusammenhänge sind signifikanter).

Stärkere Zusammenhänge sind signifikanter

Wird angenommen, dass kein Zusammenhang zwischen Variablen in der Grundgesamtheit besteht, würde man mit größter Wahrscheinlichkeit auch keinen Zusammenhang zwischen den korrespondierenden Variablen in der Stichprobe finden. Je stärker daher der Zusammenhang in der Stichprobe ist, desto unwahrscheinlicher ist es, dass kein entsprechender Zusammenhang in der Grundgesamtheit existiert. Stärke und Zuverlässigkeit hängen offensichtlich stark voneinander ab. Man kann daher die Signifikanz aus der Stärke berechnen und umgekehrt. Dies gilt jedoch nur, wenn der Stichprobenumfang konstant gehalten wird. Ein Zusammenhang mit gegebener Stärke kann nämlich je nach Stichprobenumfang hoch oder überhaupt nicht signifikant sein (siehe Abhängigkeit der Stärke einer Beziehung vom Stichprobenumfang).

Abhängigkeit der Stärke einer Beziehung vom Stichprobenumfang

Bei Vorliegen nur weniger Beobachtungen gibt es auch nur entsprechend wenige Kombinationsmöglichkeiten für die Werte der Variablen. Die Wahrscheinlichkeit, dabei eine Kombination zu erhalten, die zufällig einen starken Zusammenhang anzeigt, ist relativ hoch. Dies kann man wie folgt veranschaulichen: Gegeben seien zwei Variablen (Geschlecht: männlich/weiblich und WCC: hoch/niedrig) und eine Stichprobe von 4 Personen (2 Männer und 2

Frauen). Die Wahrscheinlichkeit, rein zufällig einen 100%igen Zusammenhang zwischen diesen beiden Variablen zu finden, beträgt dann ein Achtel. Die Chance, dass beide Männer einen hohen und beide Frauen einen niedrigen WCC-Wert oder umgekehrt haben, ist 1 zu 8. Die Wahrscheinlichkeit dafür, zufällig einen solchen perfekten Zusammenhang bei einer Stichprobe von 100 Personen zu finden, ist dagegen fast Null. Dazu sei nun ein allgemeineres Beispiel betrachtet: Gegeben sei eine theoretische Grundgesamtheit, in der der durchschnittliche WCC-Wert für Männer und Frauen gleich ist. Würde man jetzt in einem einfachen Versuch wiederholt paarweise Stichproben (jeweils Männer und Frauen) mit gegebenem Umfang entnehmen und die Differenz zwischen durchschnittlichem WCC-Wert für jedes Paar der Stichprobe berechnen, wäre das Ergebnis bei den meisten Versuchsausgängen praktisch 0. Bei einigen Stichprobenpaaren würde das Ergebnis jedoch deutlich von 0 abweichen. Wie oft kann so etwas passieren? Je kleiner der Stichprobenumfang pro Versuch ist, desto wahrscheinlicher ist ein solch "irrtümliches" Ergebnis. In diesem Fall würde das bedeuten, dass die Existenz eines Zusammenhangs zwischen Geschlecht und WCC angezeigt wird, obwohl in der Grundgesamtheit kein solcher Zusammenhang existiert.

Beispiel: Verhältnis von Jungengeburten zu Mädchengeburten

Im folgenden Beispiel aus der Forschung über statistische Aussagen (Nisbett u.a. 1987) werden zwei Krankenhäuser betrachtet: Im ersten gibt es pro Tag 120 Geburten, im zweiten dagegen nur 12. Im Mittel ist das Verhältnis von Jungen- zu Mädchengeburten in beiden Krankenhäuser 50 : 50. An einem bestimmten Tag werden nun in einem der Krankenhäuser doppelt so viele Mädchen wie Jungen geboren. In welchem Krankenhaus ist dies wahrscheinlicher? Wie die Forschung belegt ist, die Antwort für einen Statistiker offensichtlich, nicht jedoch für einen statistischen Laien: Dieses Ereignis ist im kleinen Krankenhaus viel wahrscheinlicher. Die Begründung dafür ist, dass die Wahrscheinlichkeit für eine zufällige Abweichung von einer bestimmten Größe (Mittelwert der Grundgesamtheit) mit zunehmendem Stichprobenumfang abnimmt.

Schwache Zusammenhänge sind nur in großen Stichproben als signifikant beweisbar

Das Beispiel im vorigen Abschnitt (siehe Beispiel: Verhältnis von Jungengeburten zu Mädchengeburten) deutet bereits an, dass ein "objektiv" (in der Grundgesamtheit) schwacher Zusammenhang zwischen zwei Variablen in einer Untersuchung nur durch eine entsprechend große Stichprobe aufgezeigt werden kann. Selbst wenn die Stichprobe "völlig repräsentativ" wäre, ist der Effekt statistisch nicht signifikant, wenn die Stichprobe klein ist. Analog dazu kann bei einem "objektiv" (in der Grundgesamtheit) sehr starken Zusammenhang das Ergebnis einer Untersuchung hoch signifikant sein, selbst wenn die Studie nur auf einer sehr kleinen Stichprobe basiert. Ein weiteres Beispiel soll dies verdeutlichen: Beim Wurf einer nichtidealen Münze tritt das Ereignis Kopf öfter als Zahl ein. (z.B. 60% zu 40%). Zehn Münzwürfe würden dann nicht ausreichen, jemanden von der Asymmetrie dieser Münze zu überzeugen. Dies wäre selbst dann der Fall, wenn der Ausgang dieses Versuchs völlig repräsentativ für die Münze wäre (sechsmal Kopf, viermal Zahl). Kann man also behaupten, dass 10 Münzwürfe nicht ausreichen, um irgend etwas zu beweisen? Dies ist offenbar nicht der Fall, denn wäre

der zu untersuchende Effekt nur stark genug, würden 10 Würfe sicherlich ausreichen. Wenn die Münze z.B. so asymmetrisch ist, dass sie bei jedem Wurf Kopf anzeigt, würden die meisten Leute das als Beweis dafür ansehen, dass diese Münze nicht ideal ist. Dies würde als Beleg dafür akzeptiert, dass in der (unendlichen) Grundgesamtheit aller Würfe das Ergebnis Kopf häufiger als das Ergebnis Zahl vorhanden ist. Ist also ein Zusammenhang stark, kann er auch in kleinen Stichproben als signifikant belegt werden.

"Kein Zusammenhang" als signifikantes Ergebnis?

Je schwächer der Zusammenhang zwischen Variablen ist, desto höher ist der für den Beweis seiner Signifikanz notwendige Stichprobenumfang. Man stelle sich vor, wie groß z.B. der Stichprobenumfang sein müsste, um die Asymmetrie einer Münze zu belegen, wenn diese Asymmetrie nur 0,000001 % betragen würde! Der notwendige Stichprobenumfang nimmt daher in dem Maße zu, wie die Stärke des aufzuzeigenden Effekts abnimmt. Wenn die Stärke des Effekts sich dem Wert 0 nähert, würde der für den Nachweis notwendige Stichprobenumfang gegen unendlich gehen. Wenn es also fast keinen Zusammenhang zwischen zwei Variablen gibt, muss der Stichprobenumfang fast gleich der Grundgesamtheit sein, welche als unendlich groß angenommen wird. Statistische Signifikanz stellt die Wahrscheinlichkeit für ein ähnliches Ergebnis bei Test der ganzen Grundgesamtheit dar. Jedes Ergebnis einer Untersuchung der ganzen Grundgesamtheit ist daher definitionsgemäß in höchstem Maße signifikant. Dies gilt auch für alle Ergebnisse, die "keinen Zusammenhang" nachweisen.

Messung der Stärke von Zusammenhängen zwischen Variablen

In der Statistik gibt es viele Kennziffern zur Beschreibung der Stärke eines Zusammenhangs zwischen Variablen. Die Auswahl eines konkreten Maßes hängt von der Anzahl der involvierten Variablen, den verwendeten Skalenniveaus, der Natur der Zusammenhänge usw. ab. Fast alle diese Kennziffern folgen jedoch einem allgemeinen Prinzip: Sie bewerten den beobachteten Zusammenhang, indem sie ihn mit dem "maximal vorstellbaren Zusammenhang" zwischen den betreffenden Variablen vergleichen. Ein üblicher Weg für eine solche Bewertung besteht darin, die Streuung der Variablenwerte zu betrachten und zu berechnen, welcher Anteil dieser "gesamten verfügbaren Streuung" erklärt wird, wenn die Streuung den zu untersuchenden Variablen gemeinsam ist. Man vergleicht also den Grad an Gemeinsamkeit bei beiden Variablen mit demjenigen, den man bei einem perfekten Zusammenhang zwischen den Variablen erhalten würde. Dies kann man wie folgt veranschaulichen: In einer Stichprobe ist der mittlere WCC-Index bei Männern 100 und bei Frauen 102. Im Mittel beinhaltet die Abweichung jedes einzelnen Werts vom Gesamtmittel (101) eine Komponente, die auf das Geschlecht der Person zurückzuführen ist; die Größe dieser Komponente beträgt 1. Dieser Wert misst gewissermaßen den Zusammenhang zwischen Geschlecht und WCC. Es ist jedoch nur ein recht grobes Maß, das nichts über die relative Größe im Vergleich zur gesamten Streuung der WCC-Werte aussagt. Zwei extreme Fälle sind denkbar:

(a) Wären alle WCC-Werte bei Männern exakt 100 und bei Frauen exakt 102, würden die Abweichungen vom Gesamtmittel in der Stichprobe vollständig durch das Geschlecht erklärt.

In der Stichprobe sind dann Geschlecht und WCC perfekt miteinander korreliert, d.h. 100% der Streuung einer Personen bzgl. ihres WCC-Wertes wird auf ihr Geschlecht zurückgeführt.

(b) Würden die WCC-Werte im Bereich von 0 bis 1000 liegen, würde dieselbe Differenz (von 2) zwischen mittleren WCC-Werten bei Männern und Frauen nur einen vernachlässigbar geringen Anteil an der gesamten Streuung der Werte erklären. Eine weitere, in die Stichprobe aufgenommene Person könnte die Differenz verändern oder ihre Richtung sogar umkehren. Daher muss jede gute Kennziffer von Variablen-Zusammenhängen die gesamte Streuung der individuellen Werte in der Stichprobe einbeziehen und beurteilen, wie viel dieser Streuung durch den zu untersuchenden Zusammenhang erklärt wird.

"Allgemeine Form" der meisten statistischen Tests

Da das oberste Ziel der meisten statistischen Tests in der Bewertung von Zusammenhängen zwischen Variablen besteht, folgen die meisten dieser Tests der im vorigen Abschnitt Messung der Stärke von Zusammenhängen zwischen Variablen) beschriebenen Form. Sie bilden dabei das Verhältnis von der gemessenen Streuung, die den Variablen gemeinsam ist, zu der gesamten gemessenen Streuung für diese Variablen. Sie stellen z.B. das Verhältnis der durch das Geschlecht erklärten Streuung der WCC-Werte zur Gesamtstreuung der WCC-Werte dar. Dieses Verhältnis wird üblicherweise mit erklärter Streuung (Varianz) zur gesamten Streuung (Varianz) bezeichnet. Der Begriff erklärte Streuung meint in der Statistik nicht notwendigerweise, dass man diese Streuung im eigentlichen Sinne intellektuell "versteht". Er wird nur verwendet, um die gemeinsame Streuung der Variablen, d.h. den Anteil der Streuung der Werte einer Variablen, der durch die Werte einer anderen Variablen "erklärt" wird und umgekehrt, zu kennzeichnen.

Berechnung des "statistischen Signifikanzniveaus"

Angenommen, der Zusammenhang zwischen zwei Variablen ist bereits irgendwie berechnet und gemessen worden. (wie oben erklärt). Es stellt sich nun die Frage, "wie signifikant dieser Zusammenhang ist". Ist z.B. eine erklärte Varianz von 40% ausreichend, um von einem signifikanten Zusammenhang zu sprechen? Die Antwort lautet: "Je nachdem". Insbesondere hängt die Signifikanz hauptsächlich vom Stichprobenumfang ab. Wie bereits erläutert sind in großen Stichproben oftmals selbst schwache Zusammenhänge signifikant, während in kleinen Stichproben selbst starke Zusammenhänge nicht als zuverlässig (signifikant) angesehen werden können. Um das Niveau der statistischen Signifikanz bestimmen zu können, benötigt man eine Funktion, die die Beziehung zwischen "Stärke" und "Zuverlässigkeit" von Zusammenhängen in Abhängigkeit vom Stichprobenumfang misst. Diese Funktion besagt exakt, "wie wahrscheinlich ein Zusammenhang mit gegebener Stärke bei einem bestimmten Stichprobenumfang ist, wenn annahmegemäß kein solcher Zusammenhang zwischen den Variablen in der Grundgesamtheit besteht." Diese Funktion liefert das Signifikanzniveau (p). Sie misst die Wahrscheinlichkeit für den Fehler, der mit einer Ablehnung der Behauptung, dass der zu untersuchende Zusammenhang in der Grundgesamtheit nicht existiert, verbunden ist. Diese "alternative" Hypothese (kein Zusammenhang in der Grundgesamtheit) wird üblicherweise als Nullhypothese bezeichnet. Es wäre ideal, wenn diese Wahrscheinlichkeitsfunktion linear wäre

und z.B. für verschiedene Stichprobenumfänge nur verschiedene Steigungen hätte. Unglücklicherweise ist die Funktion komplizierter und nicht immer gleich. In den meisten Fällen ist ihre Form jedoch bekannt, und man kann mit ihr die Signifikanzniveaus für Stichproben mit jeweils gegebener Größe bestimmen. Die meisten dieser Funktionen können auf einen einzigen Typ zurückgeführt werden, den man (vor allem im Englischen) als normal bezeichnet.

Bedeutung der "Normalverteilung"

Die "Normalverteilung" ist bedeutend, weil sie meistens der im vorigen Abschnitt (Berechnung des "statistischen Signifikanzniveaus") beschriebenen Funktion entspricht. Die Verteilung vieler statistischer Tests basiert auf der Normalverteilung oder kann aus der Normalverteilung abgeleitet werden. Philosophisch betrachtet stellt die Normalverteilung eine der empirisch verifizierten elementaren "Wahrheiten" über das allgemeine Wesen der Welt dar. Ihr Status ist mit demjenigen der elementaren Gesetze in den Naturwissenschaften vergleichbar. Die exakte Form der Normalverteilung (charakteristische "Glockenkurve") wird durch eine Funktion beschrieben, die nur zwei Parameter besitzt: Mittelwert und Standardabweichung.

Eine typische Eigenschaft der Normalverteilung ist, dass 68 % aller ihrer Beobachtungen in den symmetrischen Bereich des ±1-fachen und 95 % in den Bereich des ±2-fachen der Standardabweichung vom Mittelwert fallen. In der Normalverteilung besitzen Beobachtungen, die einen standardisierten Wert von -2 oder weniger und +2 oder mehr haben, eine relative Häufigkeit von 5 % oder weniger. (Standardisieren bedeutet, dass die Differenz vom Mittelwert gebildet wird und dann durch die Standardabweichung dividiert wird). Sie können die exakten Werte der Wahrscheinlichkeiten für verschiedene Werte der Normalverteilung mit dem Wahrscheinlichkeitsrechner in Elementare Statistik berechnen. Wenn Sie z.B. einen Z-Wert (d. h. standardisierten Wert) von 4 eingeben, beträgt die von STATISTICA berechnete Wahrscheinlichkeit weniger als 0,0001, da in der Normalverteilung fast alle Beobachtungen (d.h. mehr als 99.99 %) in den Bereich des ±4-fachen der Standardabweichung fallen.

Verwendung der Normalverteilung bei statistischen Begründungen/Aussagen (Einführung)

Es sei noch einmal das o. g. Beispiel der paarweisen Stichproben von Männern und Frauen, bei denen der mittlere WCC-Wert in beiden Geschlechtern gleich ist, betrachtet: Obwohl der wahrscheinlichste Ausgang für diese Versuche (ein Stichprobenpaar pro Versuch) eine Differenz zwischen den mittleren WCC-Werten bei Männern und Frauen nahe bei Null ist, könnte sich bei einigen Stichproben eine Differenz ergeben, die stark von 0 abweicht. Wie oft wird dies passieren? Wenn der Stichprobenumfang groß genug ist, werden die Ergebnisse solcher Versuchswiederholungen "normalverteilt" sein. Da die Form der Normalverteilungskurve bekannt ist, können die Wahrscheinlichkeiten für verschiedene, zufällige Abweichungsniveaus vom hypothetischen Mittel der Grundgesamtheit, d.h. von 0, exakt berechnet werden. Wenn eine so berechnete Wahrscheinlichkeit ein vorgegebenes Kriterium für statistische Signifikanz erfüllt, kann man daraus schließen, dass das Versuchsergebnis die Grundgesamtheit besser beschreibt als die "Nullhypothese". Die Nullhypothese fungiert dabei nur als Maßstab, an dem die empirischen Ergebnisse gemessen werden.

Sind alle Teststatistiken normalverteilt?

Nicht alle Prüfgrößen sind normalverteilt, aber die meisten von ihnen basieren entweder direkt auf der Normalverteilung oder auf einer daraus abgeleiteten Verteilung, wie z.B. t, F, oder Chi-Quadrat Typischerweise setzen diese Tests voraus, dass die analysierten Variablen in der Grundgesamtheit selbst normalverteilt sind, d.h. dass sie die sogenannte Normalverteilungsvoraussetzung erfüllen. Viele beobachtete Variablen sind tatsächlich normalverteilt, was ein weiterer Beleg für die Allgegenwärtigkeit der Normalverteilung in der empirischen Realität ist. Probleme entstehen dann, wenn Tests, die auf der Normalverteilung basieren, auf Variablen, die ihrerseits nicht normalverteilt sind, angewendet werden. (siehe Tests auf Normalverteilung in Nichtparametrische Verfahren oder Elementare Statistik). In diesen Fällen gibt es zwei generelle Möglichkeiten. Erstens können "nichtparametrische" Tests" (oder sogenannte "verteilungsfreie Tests", siehe Nichtparametrische Verfahren) angewendet werden. Dies ist jedoch oft unvorteilhaft, da solche Tests gewöhnlich weniger mächtig und weniger flexibel hinsichtlich ihrer Aussagen sind. Alternativ dazu kann weiterhin die Normalverteilung verwendet werden, sofern der Stichprobenumfang groß genug ist. Diese letztere Möglichkeit basiert auf einer wichtigen Regel, welche für die Beliebtheit vieler auf der Normalverteilung basierender Tests verantwortlich ist. Sie besagt, dass mit zunehmendem Stichprobenumfang die Form der Stichprobenverteilung (d. h. der Verteilung der Statistik aus der Stichprobe; der Begriff wurde erstmals von Fisher, 1928a, verwendet) sich der Normalform annähert, selbst wenn die ursprüngliche Verteilung nicht normal ist. Dieses Prinzp wird in der nachfolgenden Animation illustriert, die eine Serie von Stichprobenverteilungen (durch graduell wachsenden Stichprobenumfang von 2, 5, 10, 15, und 30 erzeugt) mit einer in der Grundgesamtheit nichtnormalverteilten Variablen zeigt, deren Werteverteilung also deutlich asymmetrisch ist.

Mit wachsendem Stichprobenumfang (von Stichproben zur Erzeugung der Stichprobenverteilung des Mittelwertes) nähert sich die Verteilungsform für die Mittelwerte einer Normalverteilung. Für n=30 ist die Verteilungsform "beinahe" normal (siehe enge Anpassung). Diese Regel bezeichnet man als Zentralen Grenzwertsatz. Diese Bezeichnung wurde erstmals von Pólya (1920) gebraucht.

Konsequenzen bei Verletzung der Normalverteilungsvoraussetzung

Obwohl viele der Aussagen der vorigen Abschnitte mathematisch bewiesen worden sind, gibt es einige, für die kein theoretischer Beweis vorliegt und die daher nur empirisch über sogenannte Monte-Carlo-Versuche gezeigt werden können. In diesen Versuchen werden von Computern mit vorgegebenen Anweisungen große Anzahlen von Stichproben generiert. Die Ergebnisse solcher Stichproben werden dann mit einer Reihe von Tests analysiert. Auf diesem Weg können Typ und Ausmaß von Fehlern bzw. Verzerrungen, die bei Nichterfüllung von Testvoraussetzungen durch einen bestimmten Datensatz auftreten, beurteilt werden. Monte-Carlo-Studien wurden insbesondere verwendet, um zu prüfen, wie auf der Normalverteilung basierende Tests auf Verletzung der Normalverteilungsvoraussetzung für die analysierten Variablen in der Grundgesamtheit reagieren. Die allgemeine Schlussfolgerung aus diesen Studien lautet, dass Konsequenzen solcher Verletzungen oftmals weniger bedeutsam sind, als zunächst vermutet. Diese Schlussfolgerungen sollten sicherlich niemanden davon abhalten,

sich bei einem Problem über die Erfüllung der Normalverteilungsvoraussetzung Gedanken zu machen. Dennoch haben diese Monte-Carlo-Studien die allgemeine Vorliebe für verteilungsabhängige Tests in allen Bereichen der Forschung erhöht.

Statistische Verfahren

Einführung

Ein kurzer Überblick zur Idee des Signifikanztests.

Um das Konzept der nichtparametrischen Verfahren verstehen zu können, wird zunächst ein grundlegendes Verständnis für die parametrischen Verfahren der Statistik benötigt. Im Abschnitt Grundbegriffe wird das Vorgehen bei einem statistischen Signifikanztest erläutert, das auf der Kenntnis der Verteilung der Teststatistik beruht. Kurz gesagt, falls die zugrundeliegende Verteilung einer Variablen bekannt ist, können wir Voraussagen darüber treffen, wie sich die Teststatistik bei wiederholten Stichproben gleicher Größe "verhalten" wird, d. h. wie sie verteilt ist. Falls wir z. B. 100 Stichproben von je 100 Erwachsenen der Bevölkerung erheben und die mittlere Körpergröße jeder Stichprobe bestimmen, dann wird die Verteilung der standardisierten Mittelwerte über die Stichproben wahrscheinlich näherungsweise die Normalverteilung (genauer gesagt, die Studentsche t-Verteilung mit 99 Freiheitsgraden, siehe unten) sein. Stellen wir uns vor, dass wir eine weitere Stichprobe aus einer bestimmten Stadt ziehen, von der wir annehmen, dass die Leute dort größer sind als die mittlere Bevölkerung. Falls die mittlere Körpergröße in dieser Stichprobe über dem oberen 95 % Bereich der t-Verteilung liegt, könnten wir tatsächlich schlussfolgern, dass die Leute dieser Stadt größer sind als der Durchschnitt der Bevölkerung.

Sind alle Variablen normalverteilt?

In dem obigen Beispiel verließen wir uns auf unser Wissen, dass in wiederholten Stichproben gleicher Größe die (standardisierten) Mittelwerte (für die Körpergröße) t-verteilt sind (mit einem gewissen Mittelwert und einer gewissen Varianz). Dies ist jedoch nur dann richtig, wenn die Variable in der Grundgesamtheit (in unserem Beispiel die Körpergröße) normalverteilt ist. Für viele Variablen, die uns interessieren, wissen wir aber nicht, ob das wirklich der Fall ist. Ist z.B. die Höhe des Einkommens in der Bevölkerung normalverteilt? Wahrscheinlich nicht. Die Häufigkeiten des Auftretens seltener Krankheiten sowie die Anzahl der Autounfälle sind z.B. nicht normalverteilt. Es gibt eine ganze Reihe weiterer Variablen, an denen man interessiert ist, die ebenfalls nicht normalverteilt sind.

Stichprobenumfang

Ein weiterer Faktor, der die Anwendbarkeit von Tests einschränkt, die auf der Voraussetzung der Normalverteilung basieren, ist der Stichprobenumfang n. Wir können die Stichprobenverteilung als Normalverteilung annehmen, wenn der Stichprobenumfang groß genug ist (z.B. 100 oder mehr Beobachtungen). Ist jedoch die Stichprobe klein, dann dürfen diese Tests nur angewendet werden, wenn wir sicher sind, dass die Variable normalverteilt ist. Es gibt aber gerade dann keine Möglichkeit, diese Annahme zu testen.

Probleme bei Messungen

Anwendungen von Tests, die auf Normalverteilungsannahmen basieren, sind weiterhin dadurch begrenzt, dass die Genauigkeit der Messungen (das Niveau der Messskala) nicht ausrei-

chend ist. Wir betrachten z.B. eine Studie, in der Schulnoten als Variable gemessen werden. Ist die Note 1 doppelt so gut wie die Note 3? Ist die Differenz zwischen 2 und 1 mit der Differenz von 4 und 3 vergleichbar? Diese Variable "Schulnote" erlaubt uns nur, eine Rangordnung der Schüler aufzustellen, die von "sehr guten" bis zu "sehr schlechten" Noten reicht. Dieses allgemeine Problem wird in Statistiklehrbüchern unter der Überschrift Datentypen oder Skalenniveau behandelt. Die Mehrzahl der statistischen Verfahren wie Varianzanalyse (und t-Tests), Regression usw. setzen voraus, dass die Variablen zumindest intervall-skaliert sind, was bedeutet, dass gleich große Intervalle der Skala in sinnvoller Weise miteinander verglichen werden können (z. B. B minus A ist gleich D minus C). Wie in unserem Beispiel ist diese Annahme sehr oft nicht zu vertreten, und die Daten stellen nur eine Rangfolge von Beobachtungen (ordinale Skala) dar.

Parametrische und nichtparametrische Verfahren

Nach dieser etwas ausführlichen Einleitung sollte der Bedarf an statistischen Verfahren ausreichend begründet sein, die es uns erlauben, Daten "minderer Qualität", Daten aus kleinen Stichproben und Daten, über deren Verteilung nichts bekannt ist, zu verarbeiten. Nichtparametrische Verfahren wurden insbesondere entwickelt, um in den Fällen verwendet zu werden, in denen man nichts über die Parameter der interessierenden Variablen der Grundgesamtheit weiß (daher der Name nichtparametrisch). Präziser formuliert stützen sich die nichtparametrischen Verfahren nicht auf die Schätzung von Parametern (wie Mittelwert und Standardabweichung) für die Beschreibung der Verteilung der interessierenden Variablen der Grundgesamtheit. Diese Verfahren werden deshalb auch oft parameterfreie oder verteilungsfreie Verfahren genannt.

Nichtparametrische Verfahren

Grundsätzlich gibt es zu jedem allgemeinen Typ eines parametrischen Testverfahrens mindestens ein nichtparametrisches Verfahren. Diese Tests lassen sich im wesentlichen in die folgenden Kategorien einteilen:

- Tests auf Differenzen zwischen Gruppen (unabhängige Stichproben);
- Tests auf Differenzen zwischen Variablen (gepaarte Stichproben);
- Tests auf Beziehungen zwischen Variablen.

Differenzen zwischen unabhängigen Gruppen

Angenommen, wir haben zwei Stichproben, die wir im Hinblick auf die Mittelwerte einer bestimmten Variablen vergleichen wollen, so würde man gewöhnlich den t-Test für unabhängige Stichproben (in Elementare Statistik) heranziehen. Nichtparametrische Alternativen für diesen Test sind der Wald-Wolfowitz-Test, der Mann-Whitney U-Test und der Kolmogorov-Smirnov-Zweistichprobentest. Bei mehreren Gruppen würde man die Varianzanalyse verwenden (siehe ANOVA/MANOVA); die analogen nichtparametrischen Verfahren sind die Kruskal-Wallis-ANOVA und der Median-Test.

Differenzen zwischen verbundenen Gruppen

Falls wir zwei Variablen derselben Stichprobe vergleichen wollen, würde gewöhnlich der t-Test für gepaarte Stichproben (in Elementare Statistik) herangezogen. (Als Beispiel könnte der Vergleich mathematischer Fertigkeiten von Schülern zu Beginn und zum Ende eines Schuljahres dienen.) Nichtparametrische Alternativen für diesen Test sind der Vorzeichentest und der Wilcoxon-Test für gepaarte Stichproben. Sind die interessierenden Variablen ihrer Natur nach kategorial mit nur zwei Ausprägungen (dichotom, d.h. "bestanden" - "nicht bestanden", "männlich" - "weiblich" usw.), ist der McNemar-Test geeignet. Bei mehr als zwei Variablen, die an derselben Stichprobe beobachtet wurden, würde man gewöhnlich die ANOVA für Messwiederholungen anwenden. Die nichtparametrischen Alternativen dafür sind Friedmans ANOVA und Cochrans Q-Test (falls die Variablen kategorial sind mit nur zwei Ausprägungen, d.h. dichotom, z.B. "bestanden" - "nicht bestanden"). Cochrans Q ist besonders dann von Nutzen, wenn Veränderungen in Häufigkeiten (Anteilen) über die Zeit gemessen werden sollen.

Beziehungen zwischen Variablen

Um eine Beziehung zwischen zwei Variablen zu beschreiben, wird gewöhnlich der Korrelationskoeffizient berechnet. Analoge nichtparametrische Maße zum Pearsonschen Korrelationskoeffizienten sind Spearmans R, Kendalls Tau und der Koeffizient Gamma. Sind zwei Variablen ihrer Natur nach dichotom (d.h. sie besitzen nur zwei Ausprägungen, wie z.B. "bestanden" - "nicht bestanden" und "männlich" - "weiblich"), dann sind geeignete nichtparametrische Verfahren für den Test der Beziehungen zwischen den beiden Variablen der Chi-Quadrat-Test, die Berechnung des Phi-Koeffizienten und der exakte Test von Fisher.). Außerdem ist ein simultaner Test für Beziehungen zwischen mehreren Variablen verfügbar: Kendalls Konkordanzkoeffizient. Dieser Test wird oftmals verwendet, um die Übereinstimmung von Testpersonen (Gutachtern) einzuschätzen, die die gleichen Objekte bewerten müssen.

Deskriptive Statistik

Sind die zu untersuchenden Daten nicht normalverteilt und enthalten die Messungen bestenfalls Informationen über die Rangfolge, dann ist die Berechnung der üblichen deskriptiven Statistiken (z.B. Mittelwert, Standardabweichung) zuweilen nicht der beste Weg, die Information über die Daten zusammenzufassen. In der Psychometrie ist z.B. wohlbekannt, dass die wahrgenommene Intensität eines Stimulus (z.B. wahrgenommene Helligkeit einer Lampe) oftmals eine logarithmische Funktion des Stimulus (Helligkeit der Lampe, objektiv gemessen, z.B. in Lux) ist. In diesem Beispiel ist der einfache Mittelwert der Wahrnehmung (Summe der Wahrnehmungen durch die Anzahl der Stimuli) keine adäquate Zusammenfassung der tatsächlichen mittleren Intensität der Stimuli. (In diesem Beispiel würde man wahrscheinlich eher das geometrische Mittel berechnen.) Mit den Nichtparametrischen Verfahren können eine Vielzahl an Lagemaßen (Mittelwert, Median, Modus usw.) und Streuungen (Varianz, mittlere Abweichung, Quartilsabstand usw.) berechnet werden, um ein vollständiges Bild der Daten zu liefern (siehe Deskriptive Statistik).

Parametrische Verfahren

t-Test für unabhängige Stichproben - Einführung

Der t-Test wird am häufigsten verwendet, um die Unterschiede zwischen Mittelwerten in zwei Gruppen zu überprüfen. Mit dem t-Test kann z. B. die Differenz zwischen den Testergebnissen einer mit einem Medikament behandelten Patientengruppen und einer Kontrollgruppe, welcher ein Placebo verabreicht wurde, getestet werden. Der t-Test kann theoretisch sogar dann eingesetzt werden, wenn die Stichprobenumfänge sehr klein sind (z. B. 10). Einige Autoren vertreten die Meinung, dass sogar noch kleinere Stichprobenumfänge möglich sind. Dies gilt, solange die Variablen normalverteilt sind und die Streuung der Werte in den beiden Gruppen nicht allzu verschieden ist (siehe auch Grundbegriffe). Wie bereits erwähnt, kann die Normalverteilungsvoraussetzung durch Untersuchung der Verteilung der Daten (über Histogramme) oder mittels Durchführung eines Normalverteilungstests (über die Option Deskriptive Statistik) überprüft werden. Die Voraussetzung gleicher Varianzen kann mit Hilfe des F-Tests verifiziert werden. Dieser ist auch in der Ergebnisausgabe zum t-Test enthalten. Sie können aber auch die Option für den robusteren Levene-Test (oder die Modifikation dieses Tests nach Brown-Forsythe) wählen. Wenn diese Bedingungen nicht erfüllt sind, können Sie die Mittelwertdifferenzen zwischen Gruppen unter Verwendung einer der nichtparametrischen Alternativen zum t-Test überprüfen (siehe Nichtparametrische Verfahren).

Die mit einem t-Test ausgegebenen p-Werte repräsentieren die Wahrscheinlichkeit für eine Fehlentscheidung bei Akzeptanz der Hypothese einer existierenden Differenz. Dies ist die Wahrscheinlichkeit für den Fehler, dass die Hypothese (keine Differenz zwischen den beiden Kategorien bzw. Gruppen von Beobachtungen in der Grundgesamtheit) abgelehnt wird, wenn diese Hypothese in Wirklichkeit zutrifft. Einige Autoren schlagen vor, bei Vorliegen einer Differenz in prognostizierter Richtung nur eine Hälfte (eine Seite) der Wahrscheinlichkeitsverteilung bzw. Dichte zu betrachten und daher den beim t-Test ("zweiseitige Wahrscheinlichkeit") angegebenen p-Wert durch 2 zu teilen. Andere sind dagegen der Meinung, dass immer die Wahrscheinlichkeit für den normalen, zweiseitigen t-Test angegeben werden sollte.

t-Test für gepaarte Stichproben - Grundidee

Bei einem t-Test für gepaarte Stichproben macht man sich das Vorhandensein eines bestimmten Design-Typs zunutze, in welchem eine wichtige Quelle der Inner-Gruppen-Streuung (der sogenannte Fehler) leicht identifiziert und aus der Analyse ausgeschlossen werden kann: Dabei stammen zwei Gruppen von (zu vergleichenden) Beobachtungen aus derselben Stichprobe von Personen, welche zweimal getestet werden (z.B. vor und nach einer Behandlung). Hierbei kann ein beträchtlicher Teil der Inner-Gruppen-Streuung der Werte in beiden Gruppen auf die anfänglichen Differenzen zwischen den Personen zurückgeführt werden. In gewissem Sinne unterscheidet sich diese Tatsache nicht von denjenigen Fällen, in denen die Gruppen völlig unabhängig sind (siehe t-Test für unabhängige Stichproben) und bei denen Unterschiede der Einzelbeobachtungen ebenfalls zur Fehlervarianz beitragen. Im Fall unabhängiger Stichproben kann man die auf die einzelnen Unterschiede zwischen den Personen zurückzuführende Streuung jedoch nicht identifizieren (oder "subtrahieren"). Wenn dieselbe Stichprobe jedoch

zweimal getestet wird, ist diese Identifikation (oder "Subtraktion") leicht durchführbar. Hierbei ist es möglich, statt einer getrennten Behandlung jeder Gruppe und einer Analyse der tatsächlichen Werte nur die Differenzen zwischen den beiden Messungen für jedes Objekt zu betrachten (z.B. "vor dem Test" und "nach dem Test"). Indem für jede Person der erste Wert vom zweiten abgezogen wird und nur diese "reine", paarweise Differenz analysiert wird, ist die auf die ungleiche Ausgangsbasis der einzelnen Personen zurückzuführende Streuung von der Analyse ausgeschlossen. Dies ist genau die Vorgehensweise beim t-Test für gepaarte Stichproben. Im Vergleich zum t-Test für unabhängige Stichproben liefert er "bessere" Ergebnisse (d. h. er ist stets sensitiver bzw. trennschärfer).

ANOVA/MANOVA

Der Zweck der Varianzanalyse. Im allgemeinen besteht der Zweck der Varianzanalyse (ANOVA) darin, die Signifikanz von Mittelwertdifferenzen zu testen. Der Abschnitt Grundbegriffe beinhaltet eine kurze Einführung in die Grundlagen statistischer Signifikanztests. Werden nur zwei Mittelwerte verglichen, liefert die ANOVA dieselben Ergebnisse wie der t-Test für unabhängige Stichproben (bei Vergleich von zwei verschiedenen Gruppen von Fällen oder Beobachtungen) oder der t-Test für gepaarte Stichproben (bei Vergleich von zwei Variablen für dieselben Fälle oder Beobachtungen). Wenn Sie mit diesen Tests nicht vertraut sind, sollten Sie an diesem Punkt durch Lesen des Abschnitts Elementare Statistik und Tabellen Ihre Kenntnisse "auffrischen".

Warum der Name Varianzanalyse? Es erscheint auf den ersten Blick möglicherweise seltsam, dass ein Verfahren zum Vergleich von Mittelwerten Varianzanalyse genannt wird. Diese Bezeichnung leitet sich jedoch aus der Tatsache ab, dass beim Testen der statistischen Signifikanz von Mittelwertdifferenzen eigentlich Varianzen (Streuungen) verglichen (d.h. analysiert) werden.

ANOVA/MANOVA Einführung - Grundidee

Der Zweck der Varianzanalyse. Im allgemeinen besteht der Zweck der Varianzanalyse (ANOVA) darin, die Signifikanz von Mittelwertdifferenzen zu testen. Der Abschnitt Grundbegriffe beinhaltet eine kurze Einführung in die Grundlagen statistischer Signifikanztests. Werden nur zwei Mittelwerte verglichen, liefert die ANOVA dieselben Ergebnisse wie der t-Test für unabhängige Stichproben (bei Vergleich von zwei verschiedenen Gruppen von Fällen oder Beobachtungen) oder der t-Test für gepaarte Stichproben (bei Vergleich von zwei Variablen für dieselben Fälle oder Beobachtungen). Wenn Sie mit diesen Tests nicht vertraut sind, sollten Sie an diesem Punkt durch Lesen des Abschnitts Elementare Statistik und Tabellen Ihre Kenntnisse "auffrischen".

Warum der Name Varianzanalyse? Es erscheint auf den ersten Blick möglicherweise seltsam, dass ein Verfahren zum Vergleich von Mittelwerten Varianzanalyse genannt wird. Diese Bezeichnung leitet sich jedoch aus der Tatsache ab, dass beim Testen der statistischen Signifikanz von Mittelwertdifferenzen eigentlich Varianzen (Streuungen) verglichen (d.h. analysiert) werden.

Im o.g. Beispiel hätte man auch mit Hilfe des Moduls Elementare Statistik und Tabellen einfach einen t-Test für unabhängige Stichproben durchführen können und wäre zur selben Schlussfolgerung gelangt. Tatsächlich erhält man das gleiche Ergebnis, wenn man diesen Test für den Vergleich der beiden Gruppen verwendet. Die ANOVA ist jedoch ein viel flexibleres und umfangreicheres Verfahren als der t-Test, welches auch bei komplizierteren Untersuchungsvorhaben angewendet werden kann.

Mehrere Faktoren. Die Welt ist naturgemäß kompliziert und multivariat. Fälle, in denen eine einzige Variable ein Phänomen vollständig erklärt, sind selten. Wenn man z.B. Möglichkeiten zum Züchten größerer Tomaten untersuchen möchte, muss man Faktoren wie den genetischen Aufbau der Pflanze, die Bodenbedingungen, die Lichtverhältnisse, die Temperatur usw. berücksichtigen. In einen typischen Versuch werden daher viele Faktoren einbezogen. Ein wichtiger Grund dafür, statt mehrerer Zwei-Gruppen-Untersuchungen mit Hilfe von t-Tests die ANOVA zu verwenden, besteht darin, dass die ANOVA das effizientere Verfahren ist und aus weniger Beobachtungen mehr Informationen gewonnen werden können. Dieses Argument soll nachfolgend vertieft werden.

Einfluss von Faktoren kontrollieren. Angenommen, im o.g. Beispiel mit zwei Gruppen wird ein weiterer Faktor Geschlecht eingeführt: Jede Gruppe besteht aus jeweils aus 3 Männern und 3 Frauen. Dieses Design kann in einer 2 x 2-Tabelle dargestellt werden:

	Versuchsgruppe 1	**Versuchsgruppe 1**
Jungen	2 3 1	6 7 5
Mittelwert	2	6
Mädchen	4 5 3	8 9 7
Mittelwert	4	8

Bereits bevor Berechnungen durchgeführt werden, kann die gesamte Varianz in mindestens drei Quellen zerlegt werden: (1) Fehlervarianz (innerhalb der Gruppen), (2) Varianz aufgrund der Versuchsgruppenzugehörigkeit und (3) Varianz aufgrund des Geschlechts. Es gibt darüber hinaus eine weitere Quelle, die Interaktion bzw. Wechselwirkung, welche an späterer Stelle erklärt wird. Was wäre passiert, wenn man den Faktor Geschlecht nicht in die Untersuchung mit aufgenommen hätte, sondern einen einfachen t-Test durchgeführt hätte? Wenn die Summen der Quadrate SQ ohne Berücksichtigung des Faktors Geschlecht berechnet wird (bei Verwendung der Mittelwerte innerhalb der Gruppen - unter Ignorieren des Faktors Geschlecht lautet das Ergebnis: SQ=10+10=20), zeigt sich, dass die Inner-Gruppen-SQ größer ist als bei Einschluss des Geschlechts. Bei Berechnung der Quadratsummen unter Verwendung der Mittelwerte innerhalb der Gruppen und innerhalb des Geschlechts lautet die Summe der Quadrate innerhalb der Gruppen nämlich: 2+2+2+2=8. Diese Differenz ist auf die Tatsache zurückzuführen, dass die Mittelwerte für Jungen systematisch niedriger als für Mädchen sind. Die Dif-

ferenz der Mittelwerte erhöht die Streuung, wenn dieser Faktor ignoriert wird. Die Einbeziehung dieser Quelle der Fehlervarianz erhöht die Sensitivität (Macht) eines Tests. Die Beispiele zeigen einen weiteren Vorzug der ANOVA gegenüber einfachen Zwei-Gruppen-Untersuchungen mittels t-Tests: In der ANOVA kann jeder einzelne Faktor getestet werden, während der Einfluss aller anderen Faktoren eliminiert wird. Dies ist der eigentliche Grund dafür, warum die ANOVA im statistischen Sinne mächtiger als der einfache t-Test ist, d.h. zum Nachweis eines signifikanten Effekts weniger Beobachtungen benötigt.

Grundidee - Interaktionseffekte

Es gibt einen weiteren Vorteil der ANOVA gegenüber einfachen t-Tests: ANOVA ermöglicht die Aufdeckung von Interaktionseffekten bzw. Wechselwirkungen zwischen Variablen. Damit können umfassendere Hypothesen über die Wirklichkeit getestet werden. Ein weiteres Beispiel soll diesen Punkt verdeutlichen.

Haupteffekte, zweifache Interaktion. Angenommen, es gibt eine Stichprobe mit sehr stark leistungsorientierten Schülern und eine weitere mit sogenannten "Leistungsverweigerern". Man kann nun diese Stichproben in zwei zufällige Hälften aufteilen und jeweils der einen Hälfte in jeder Stichprobe einen schweren Test und der anderen Hälfte einen leichten Test vorlegen. Es wird die Leistungsanstrengung der Schüler bei diesem Test gemessen. Die Mittelwerte dieser (fiktiven) Studie sind wie folgt:

	Leistungsorientiert	**Leistungsverweigerung**
Schwerer Test	10	5
Einfacher Test	5	10

Wie kann man diese Ergebnisse interpretieren? Ist es z.B. richtig, hieraus zu schließen, dass erstens schwere Tests zu größeren Anstrengungen führen oder zweitens leistungsorientierte Schüler sich stärker anstrengen als Leistungsverweigerer? Keine dieser Aussagen trifft den Kern dieses deutlich sichtbaren systematischen Musters der Mittelwerte. Die richtige Interpretation der Ergebnisse besteht darin, dass schwere Tests nur zu größeren Anstrengungen bei leistungsorientierten Schülern führen, während leichte Tests nur bei Leistungsverweigerern zu größeren Anstrengungen führen. Es liegt also eine Interaktion bzw. Wechselwirkung zwischen der Leistungsbereitschaft der Schüler und der Schwierigkeit des Tests vor. In diesem speziellen Beispiel handelt es sich um eine zweifache Interaktion zwischen Leistungsbereitschaft und Testschwierigkeit. Die oben gemachten Aussagen 1 und 2 beschreiben in diesem Zusammenhang die sogenannten Haupteffekte.

Interaktion höherer Ordnung. Während die beschriebene Interaktion leicht in Worte gefasst werden kann, ist die verbale Beschreibung von Interaktionen höherer Ordnung schwieriger. Es sei angenommen, dass ein weiterer Faktor Geschlecht in die Untersuchung zur Leistungsbereitschaft einbezogen und dabei folgendes Muster der Mittelwerte auftrat:

Mädchen	Leistungsorientiert	Leistungsverweigerung
Schwerer Test	10	5
Leichter Test	5	10
Jungen	**Leistungsorieniert**	**Leistuingsverweigerung**
Schwerer Test	1	6
Leichter Test	6	1

Wie können nun die Ergebnisse der Untersuchung interpretiert werden? Das Modul ANOVA/MANOVA ermöglicht die Erzeugung von Grafiken für Mittelwerte aller Effekte (Interaktionsplots), wobei im Prinzip nur ein Mausklick notwendig ist. Diese Grafiken erleichtern die Interpretation komplizierter Effekte. Das Muster in der obigen Tabelle stellt eine dreifache Interaktion bzw. Wechselwirkung zwischen Faktoren dar. Man kann dieses Muster interpretieren, indem man z.B. sagt, dass es für Mädchen eine zweifache Interaktion zwischen Leistungsbereitschaft und Testschwierigkeit gibt: Leistungsorientierte Schülerinnen strengen sich bei schweren Tests stärker an als bei leichten Tests, "leistungsverweigernde" Schülerinnen strengen sich dagegen bei leichten Tests stärker als bei schweren Tests an. Für männliche Schüler gilt diese Interaktion umgekehrt. Wie sich zeigt, ist die Beschreibung der Interaktion in diesem Fall weitaus schwieriger.

Eine allgemeine Beschreibung von Interaktionen. Eine allgemeine Möglichkeit zur Beschreibung aller Interaktionen besteht darin, zu sagen, dass ein Effekt durch einen anderen Effekt modifiziert wird. Dies soll nun mit Hilfe der o.g. zweifachen Interaktion dargestellt werden. Der Haupteffekt der Testschwierigkeit wird durch die Leistungsbereitschaft modifiziert. Für die dreifache Interaktion im Beispiel kann man sagen, dass die zweifache Interaktion zwischen Testschwierigkeit und Leistungsbereitschaft durch das Geschlecht modifiziert wird. Bei einer vierfachen Interaktion kann man sagen, dass die dreifache Interaktion durch die vierte Variable modifiziert wird, d.h. dass es unterschiedliche Typen von Interaktionen auf den verschiedenen Stufen dieser vierten Variablen gibt. Wie sich in der Praxis gezeigt hat, sind in vielen Forschungsbereichen Interaktionen von fünfter oder höherer Ordnung nicht ungewöhnlich.