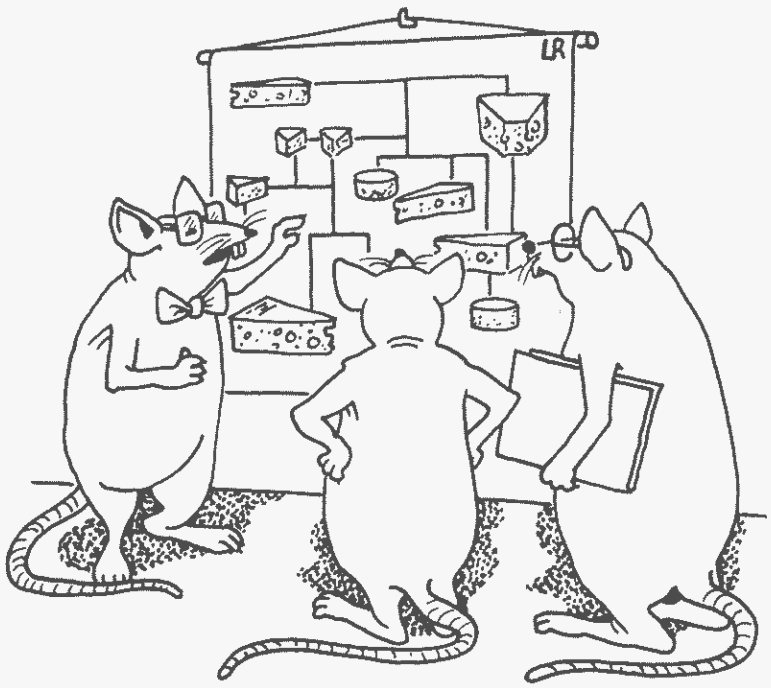
Frage ist nicht ganz einfach, da es zur Lösung dieser wissenschaftstheoretischen Fragen keine „Patentrezepte" gibt. Vielmehr liegen eine Reihe nuancierter Lösungsansätze vor, deren Vertreter sich gegenseitig zum Teil sehr heftig kritisieren.

Zweifellos ist der *kritische Rationalismus* das bekannteste wissenschaftstheoretische Rahmenmodell; unabhängig davon hat sich die bereits zu Beginn dieses Jahrhunderts entwickelte statistische Hypothesenprüfung mittels *Signifikanztests* in der modernen empirischen Forschung weitgehend durchgesetzt. Die Ursprünge einer großen Gruppe von Signifikanztests, nämlich der Korrelations- und Regressionsanalysen, gehen auf K. Pearson (1896) zurück, die erste Varianzanalyse legte Fisher zusammen mit MacKenzie im Jahr 1923 vor, deutlich vor der Publikation von Poppers „Logik der Forschung" (1934). Gemeinsames Ursprungsland all dieser Entwicklungen ist interessanterweise England.

Wie lassen sich nun die Forschungslogik des kritischen Rationalismus bzw. Falsifikationismus und der statistische Signifikanztest als Methode der Hypothesenprüfung miteinander verbinden? Fisher (1925, 1935, 1956), der mit seinen Arbeiten der empirisch-statistischen Forschung ganz wesentliche Impulse gab, war selbst Anhänger eines induktiven Modells; er sprach von „induktiver Inferenz" und meinte damit die Schlußfolgerung von Stichproben auf Populationen und von Beobachtungsdaten auf Hypothesen. Erkenntnisfortschritt kommt seiner Auffassung nach durch ein wiederholtes *Widerlegen von Nullhypothesen* (s. u.) zustande bzw. durch den indirekten Nachweis von Effekten. Ausgangspunkt der Hypothesenprüfung waren bei dem Biologen und Statistiker Fisher stets *statistische Hypothesen* (also Wahrscheinlichkeitsmodelle, vgl. S. 27 ff.).

Popper (1934) dagegen thematisierte die Bildung von Theoriesystemen im größeren Rahmen, ihn interessierten auch die unterschiedlichen Abstraktionsebenen von Aussagen und die Ableitungsbeziehungen zwischen Hypothesen. Die Falsifikation von Theorien bezieht sich bei Popper auf die Relation zwischen Hypothesen, Randbedingungen und Theorien, während sich Fisher (1925) primär mit der Relation von Daten und statistischen Hypothesen befaßte.



Für inhaltliche Hypothesen müssen die passenden statistischen Hypothesen formuliert werden. (Zeichnung: R. Löffler, Dinkelsbühl)

Statistische Hypothesen bzw. Wahrscheinlichkeitsaussagen sind weder falsifizierbar (es gibt keine logisch falsifizierenden Ereignisse, da eine Wahrscheinlichkeitsaussage grundsätzlich alle Ereignisse zuläßt, ihnen lediglich unterschiedliche Auftretenshäufigkeiten zuschreibt) noch verifizierbar (es lassen sich nicht alle Elemente der Population, über die Aussagen getroffen werden sollen, untersuchen). Bei statistischen Hypothesen läßt sich Falsifizierbarkeit jedoch durch die Festlegung von Falsifikationskriterien *herstellen*. Genau dies schlägt Fisher (1925) vor und steht damit in Einklang mit den Vorstellungen von Popper (1989, S. 208): „Nach unserer Auffassung sind Wahrscheinlichkeitsaussagen, wenn man sich nicht *entschließt*, sie durch Einführung einer methodologischen Regel *falsifizierbar zu machen*, eben wegen ihrer völligen Unentscheidbarkeit metaphysisch." Die auf Fisher (1925) zurückgehende Festlegung eines *Signifikanzniveaus* (s. u.) ist gleichbedeutend mit der Vereinbarung einer Falsifikationsregel; diese Parallele werden wir unten näher ausführen.

Ganz wesentlich bei der wissenschaftstheoretischen Interpretation statistischer Hypothesenprüfung ist der Gedanke, daß die Daten einem nicht „sagen", ob eine Hypothese „stimmt", sondern daß die Daten nur die Grundlage einer *Entscheidung* für oder gegen eine Hypothese darstellen. Die Möglichkeit, sich dabei falsch zu entscheiden, soll möglichst minimiert werden; sie ist jedoch niemals gänzlich auszuschalten.

An dieser Stelle sei noch einmal ausdrücklich darauf verwiesen, daß es neben dem sog. *klassischen Signifikanztest* bzw. *Nullhypothesen-Test* in der Tradition von Fisher noch weitere Varianten der statistischen Hypothesenprüfung gibt, nämlich den Signifikanztest nach Neyman und E. Pearson (1928), den Sequentialtest nach Wald (1947) und die Bayes'sche Statistik (Edwards et al., 1963). Auf Entstehungszusammenhänge und Unterschiede dieser Ansätze gehen z. B. Cowles (1989), Gigerenzer und Murray (1987), Ostmann und Wutke (1994) sowie Willmes (1996) ein. Um eine methodische Brücke zwischen *inhaltlichen* Hypothesen und Theorien einerseits und *statistischen* Hypothesen andererseits bemühen sich z. B. Erdfelder und Bredenkamp (1994) und Hager (1992), die auch Poppers Konzept der „Strenge" einer Theorieprüfung umsetzen. Kritische Anmerkungen zur Praxis des Signifikanztestens sind z. B. Morrison und Henkel (1970), Ostmann und Wutke (1994) sowie Witte (1989) zu entnehmen (vgl. hierzu auch Kap. 8.1.3).

Wenden wir uns nun dem Funktionsprinzip des klassischen Signifikanztests und seiner Bedeutung für den wissenschaftlichen Erkenntnisgewinn zu. Unter Verzicht auf technische Details und Präzisierungen (vgl. hierzu Abschnitt 8.1.2 bzw. Bortz, 1999, Kap. 4) wollen wir zunächst das Grundprinzip der heute gängigen statistischen Hypothesenprüfung darstellen. Dieses Signifikanztestmodell steht in der Tradition von Fisher (1925), übernimmt aber auch einige Elemente (z. B. die Idee einer *Alternativhypothese*, s. u.) aus der Theorie von Neyman und Pearson (1928) und wird deswegen zuweilen auch als „Hybrid-Modell" (Gigerenzer, 1993) oder als „Testen von Nullhypothesen nach Neyman und Pearson" bezeichnet (Ostmann und Wutke, 1994, S. 695). Das Hybrid-Modell ist in Deutschland seit den 50er/60er Jahren bekannt geworden.

1.3.1 Statistische Hypothesenprüfung

Ausgangspunkt der statistischen Hypothesenprüfung ist idealerweise eine Theorie (bzw. ersatzweise eine gut begründete Überzeugung), aus der unter Festlegung von Randbedingungen eine inhaltliche Hypothese abgeleitet wird, die ihrerseits in eine statistische Hypothese umzuformulieren ist. Die statistische Hypothese sagt das Ergebnis einer empirischen Untersuchung vorher (Prognose) und gibt durch ihren theoretischen Hintergrund gleichzeitig eine Erklärung des untersuchten Effektes.

Untersuchungsplanung

Greifen wir das Beispiel der Arousal-Theorie wieder auf: Aus dieser Theorie läßt sich die *Forschungshypothese* ableiten, daß Harmoniefolgen mit mittlerem Erregungspotential positiv bewertet werden. Diese Aussage ist noch sehr allgemein gehalten. Um ein empirisches Ergebnis vorherzusagen, muß zunächst die Zielpopulation bestimmt werden (z. B. alle erwachsenen Personen aus dem westeuropäischen Kulturkreis). Die Hypothesenprüfung wird später nicht am Einzelfall, sondern

an einer Stichprobe von Personen aus der Zielpopulation erfolgen. Weiterhin müssen wir uns Gedanken darüber machen, welche Art von Harmoniefolgen die Untersuchungsteilnehmer bewerten sollen, und wie sie ihre Urteile äußern. Man könnte sich z. B. für ein Zwei-Gruppen-Design entscheiden, bei dem eine Gruppe eine zufällige Auswahl von Popmusikstücken hört (Kontrollgruppe) und einer anderen Gruppe nur Popmusik mit mittlerem Erregungspotential präsentiert wird (Experimentalgruppe).

Die Zusammenstellung von Musikstücken mit mittlerem Erregungspotential wird Fachleuten (z. B. Personen mit musikwissenschaftlicher Ausbildung) überlassen; die Zuordnung der Untersuchungsteilnehmer zu beiden Gruppen sollte zufällig erfolgen (Randomisierung). Zur Erfassung der Einschätzung der Musikstücke werden Ratingskalen eingesetzt (gefällt mir gar nicht – wenig – teils-teils – ziemlich – völlig), d. h. jeder Proband schätzt eine Serie von z. B. 20 Musikstücken ein, bewertet jedes Stück auf der Ratingskala und erhält eine entsprechende Punktzahl („gefällt mir gar nicht" = 1 Punkt bis „gefällt mir völlig" = 5 Punkte). Je positiver die Bewertung, um so höher ist die Gesamtpunktzahl pro Person.

Statistisches Hypothesenpaar

Nach diesen designtechnischen Vorüberlegungen läßt sich das Untersuchungsergebnis laut Forschungshypothese prognostizieren: Die Experimentalgruppe sollte die Musik positiver einschätzen als die Kontrollgruppe. Diese inhaltliche Unterschiedshypothese (Experimental- und Kontrollgruppe sollten sich *unterscheiden*) ist, wie auf S. 12 erläutert, in eine statistische Mittelwertshypothese zu überführen, die ausdrückt, daß der Mittelwert der Musikbewertungen in der Experimentalgruppe (bzw. genauer: in der Population westeuropäischer Personen, die Musikstücke mittleren Erregungspotentials hören) größer ist als in der Kontrollgruppe: $\mu_1 > \mu_2$.

Eine Besonderheit der statistischen Hypothesenprüfung besteht darin, daß sie stets von einem *Hypothesenpaar*, bestehend aus einer sog. **Alternativhypothese** (H_1) und einer **Nullhypothese** (H_0), ausgeht. Die Forschungshypothese entspricht üblicherweise der Alternativhypothese, während die Nullhypothese der Alternativhypothese genau widerspricht. Besagt die gerichtete Alternativhypothese wie oben, daß der Mittelwert unter den Bedingungen der Experimentalgruppe größer ist als der Mittelwert unter den Bedingungen der Kontrollgruppe, so behauptet die Nullhypothese, daß sich beide Gruppen *nicht* unterscheiden oder der Mittelwert der Experimentalgruppe sogar kleiner ist. In Symbolen:

H_1: $\mu_1 > \mu_2$
H_0: $\mu_1 \leq \mu_2$

Die Nullhypothese drückt inhaltlich immer aus, daß Unterschiede, Zusammenhänge, Veränderungen oder besondere Effekte in der interessierenden Population *überhaupt nicht* und/oder *nicht in der erwarteten Richtung* auftreten. Im Falle einer *ungerichteten* Forschungs- bzw. Alternativhypothese postuliert die Nullhypothese keinerlei Effekt. Im Falle einer *gerichteten* Alternativhypothese wie im obigen Beispiel geht die Nullhypothese von keinem oder einem gegengerichteten Effekt aus (zur Richtung von Hypothesen siehe S. 11).

Beispiele für Nullhypothesen sind: „Linkshänder und Rechtshänder unterscheiden sich *nicht* in ihrer manuellen Geschicklichkeit"; „Es gibt *keinen* Zusammenhang zwischen Stimmung und Wetterlage"; „Die Depressivitätsneigung ändert sich im Laufe einer Therapie *nicht*"; „Aufklärungskampagnen über AIDS-Risiken haben *keinen* Einfluß auf die Kondomverwendung." Alternativhypothesen bzw. Forschungshypothesen handeln demgegenüber gerade vom Vorliegen besonderer Unterschiede, Zusammenhänge oder Veränderungen, da man Untersuchungen typischerweise durchführt, um interessante oder praktisch bedeutsame Effekte nachzuweisen und nicht etwa, um sie zu negieren.

Fassen wir zusammen: Vor jeder Hypothesenprüfung muß ein **statistisches Hypothesenpaar**, bestehend aus H_1 und H_0, in der Weise formuliert werden, daß alle möglichen Ausgänge der Untersuchung abgedeckt sind. Die Nullhypothese „Der Mittelwert unter Experimentalbedingungen ist kleiner oder gleich dem Mittelwert unter Kontrollbedingungen" und die Alternativhypothese „Der Mittelwert unter Experimentalbedingungen ist größer als der der Mittelwert unter Kontrollbedingungen" bilden ein solches Hypothesenpaar.

! Eine statistische Hypothese wird stets als *statistisches Hypothesenpaar*, bestehend aus *Nullhypothese* (H_0) und *Alternativhypothese* (H_1), formuliert. Die Alternativhypothese postuliert dabei einen bestimmten Effekt, den die Nullhypothese negiert.

Das komplementäre Verhältnis von H_0 und H_1 stellt sicher, daß bei einer Zurückweisung der H_0 „automatisch" auf die Gültigkeit der H_1 geschlossen werden kann, denn andere Möglichkeiten gibt es ja nicht.

Auswahl eines Signifikanztests

Nach Untersuchungsplanung und Formulierung des statistischen Hypothesenpaares wird ein geeigneter Signifikanztest ausgewählt. Kriterien für die Auswahl von Signifikanztests sind etwa die Anzahl der Untersuchungsgruppen, der Versuchspersonen, der abhängigen und unabhängigen Variablen oder die Qualität der Daten. In unserem Beispiel ist es der sog. *t-Test*, der genau für den Mittelwertsvergleich zwischen zwei Gruppen konstruiert ist. Für Mittelwertsvergleiche zwischen mehreren Gruppen wäre dagegen z. B. eine *Varianzanalyse* indiziert, für eine Zusammenhangshypothese würde man Korrelationstests heranziehen usw. All diese Signifikanztests beruhen jedoch auf demselben Funktionsprinzip, das wir unten darstellen. An dieser Stelle sei noch einmal betont, daß die bisher geschilderten Arbeitsschritte *vor* der Erhebung der Daten durchzuführen sind. Erst nach einer detaillierten Untersuchungsplanung kann die Untersuchung praktisch durchgeführt werden, indem man wie geplant eine Stichprobe zieht, in Kontroll- und Experimentalgruppe aufteilt, Musikstücke bewerten läßt, für die Bewertungen Punkte vergibt und anschließend die Mittelwerte beider Gruppen berechnet.

Das Stichprobenergebnis

Das Stichprobenergebnis (z. B. $\bar{x}_1 = 3{,}4$ und $\bar{x}_2 = 2{,}9$) gibt „per Augenschein" erste Hinweise über die empirische Haltbarkeit der Hypothesen. In Übereinstimmung mit der Alternativhypothese hat die Experimentalgruppe, wie am höheren Punktwert erkennbar, die präsentierten Musikstücke tatsächlich positiver eingeschätzt als die Kontrollgruppe. Diese Augenscheinbeurteilung des Stichprobenergebnisses (*deskriptives Ergebnis*) läßt jedoch keine Einschätzung darüber zu, ob das Ergebnis auf die Population zu generalisieren ist (dann sollte man sich für die H_1 entscheiden) oder ob der Befund zufällig aus den Besonderheiten der Stichproben resultiert und sich bei anderen Stichproben gar nicht gezeigt hätte, so daß eine Entscheidung für die H_0 angemessen wäre. Diese Entscheidung wird nicht nach subjektivem Empfinden, sondern auf der Basis eines Signifikanztestergebnisses gefällt.

Berechnung der Irrtumswahrscheinlichkeit mittels Signifikanztest

Bei einem Signifikanztest wird zunächst gefragt, ob das Untersuchungsergebnis durch die Nullhypothese erklärt werden kann. Kurz formuliert ermittelt man hierfür über ein Wahrscheinlichkeitsmodell einen Wert (sog. **Irrtumswahrscheinlichkeit**), der angibt, mit welcher bedingten Wahrscheinlichkeit das gefundene Untersuchungsergebnis auftritt, wenn in der Population die Nullhypothese gilt.

! Die *Irrtumswahrscheinlichkeit* ist die bedingte Wahrscheinlichkeit, daß das empirisch gefundene Stichprobenergebnis zustandekommt, wenn in der Population die Nullhypothese gilt.

In unserem Beispiel würden wir also zunächst „probeweise" annehmen, daß bei erwachsenen Personen aus dem westlichen Kulturkreis (Population) Musikstücke mit mittlerem Erregungspotential *nicht* besonders positiv bewertet werden (Nullhypothese). Wäre dies der Fall, müßte man für das Stichprobenergebnis mit hoher Wahrscheinlichkeit erwarten, daß die Experimentalgruppe in etwa denselben Mittelwert erreicht wie die Kontrollgruppe.

Signifikante und nicht-signifikante Ergebnisse

Ein vernachlässigbar geringer Unterschied zwischen Experimental- und Kontrollgruppe schlägt sich in einer *hohen Irrtumswahrscheinlichkeit* im Signifikanztest nieder und wird als *nicht-signifikantes Ergebnis* bezeichnet. Bei einem nicht-signifikanten Ergebnis gilt die Alternativhypothese als nicht bestätigt. Würde man bei dieser Datenlage dennoch auf der Alternativhypothese beharren, ginge man ein hohes Risiko ein, sich zu irren

(hohe Irrtumswahrscheinlichkeit!). Der Irrtum bestünde darin, daß man zu Unrecht davon ausgeht, der empirisch gefundene Stichprobeneffekt (Unterschied der Stichprobenmittelwerte) würde analog auch in der Population gelten (Unterschied der Populationsmittelwerte). Die wichtigste Funktion des Signifikanztests liegt also in der Bestimmung der Irrtumswahrscheinlichkeit. Beim t-Test gehen die beiden Gruppenmittelwerte in eine Formel ein, aus der sich die Irrtumswahrscheinlichkeit berechnen läßt.

Je größer der Mittelwertunterschied zwischen Experimental- und Kontrollgruppe, desto schlechter ist er mit der Nullhypothese zu vereinbaren. Es ist äußerst unwahrscheinlich, daß in den geprüften Stichproben ein solcher Unterschied „zufällig" auftaucht, wenn in den Populationen kein Unterschied besteht (H_0), zumal man dafür Sorge getragen hat (oder haben sollte), daß keine untypischen Probanden mit ungewöhnlichen Musikwahrnehmungen befragt wurden. Weicht das Stichprobenergebnis deutlich von den Annahmen der Nullhypothese ab, wertet man dies nicht als Indiz dafür, eine ganz außergewöhnliche Stichprobe gezogen zu haben, sondern interpretiert dieses unwahrscheinliche Ergebnis als Hinweis darauf, daß man die Nullhypothese verwerfen und sich lieber für die Alternativhypothese entscheiden sollte, d.h. für die Annahme, daß auch in der Population Musikstücke mit mittlerem Erregungspotential positiver bewertet werden.

Läßt sich das Stichprobenergebnis schlecht mit der Nullhypothese vereinbaren, berechnet der Signifikanztest eine *geringe Irrtumswahrscheinlichkeit.* In diesem Fall spricht man von einem **signifikanten Ergebnis**, d.h. die Nullhypothese wird zurückgewiesen und die Alternativhypothese angenommen. Da die Datenlage gegen die Nullhypothese spricht, geht man bei Annahme der Forschungshypothese nur ein geringes Risiko ein, sich zu irren (geringe Irrtumswahrscheinlichkeit).

! Ein *signifikantes Ergebnis* liegt vor, wenn ein Signifikanztest eine *sehr geringe* Irrtumswahrscheinlichkeit ermittelt. Dies bedeutet, daß sich das gefundene Stichprobenergebnis nicht gut mit der Annahme vereinbaren läßt, daß in der Population die Nullhypothese gilt. Man lehnt deshalb die Nullhypothese ab und akzeptiert die Alternativhypothese.

Ein Restrisiko bleibt jedoch bestehen, weil es ganz selten doch vorkommt, daß „in Wirklichkeit" die Nullhypothese in der Population gilt und die in der Stichprobe vorgefundenen Effekte reine Zufallsprodukte aufgrund untypischer Probanden darstellen und somit die Nullhypothese zu unrecht verworfen wird.

Das Signifikanzniveau

Um solche Irrtümer möglichst zu vermeiden, wurden für die Annahme der Alternativhypothese bzw. für die Ablehnung der Nullhypothese strenge Kriterien vereinbart: Nur wenn die Irrtumswahrscheinlichkeit wirklich sehr klein ist, nämlich unter 5% liegt, ist die Annahme der Alternativhypothese akzeptabel. Man beachte, daß es sich bei der Irrtumswahrscheinlichkeit um eine (bedingte) *Datenwahrscheinlichkeit* handelt und nicht um eine *Hypothesenwahrscheinlichkeit.* Bei einer Irrtumswahrscheinlichkeit von z.B. 3% zu behaupten, die Alternativhypothese träfe mit 97%iger Wahrscheinlichkeit zu, wäre also vollkommen falsch. Die richtige Interpretation lautet, daß die Wahrscheinlichkeit für das Untersuchungsergebnis (und aller Ergebnisse, die noch deutlicher für die Richtigkeit der H_1 sprechen) für den Fall, daß die H_0 gilt, nur 3% beträgt.

Die 5%-Hürde für die Irrtumswahrscheinlichkeit nennt man *Signifikanzniveau* oder Signifikanzschwelle; sie stellt ein willkürlich festgelegtes Kriterium dar und geht auf Fisher (1925) zurück. In besonderen Fällen wird noch strenger geprüft, d.h. man orientiert sich an einer 1%- oder 0,1%-Grenze. Dies ist insbesondere dann erforderlich, wenn von einem Ergebnis praktische Konsequenzen abhängen und ein Irrtum gravierende Folgen hätte. In der Grundlagenforschung ist dagegen ein Signifikanzniveau von 5% üblich.

Zusammenfassend kann man sagen, daß der Signifikanztest eine standardisierte statistische Methode darstellt, um auf der Basis von empirisch-quantitativen Stichprobendaten zu entscheiden, ob die Alternativhypothese anzunehmen ist oder nicht. Da die Alternativhypothese, die stets das Vorliegen von Effekten postuliert, in der Regel der Forschungshypothese entspricht, die der Wissenschaftler bestätigen will, soll die Entscheidung für die Alternativhypothese nicht vorschnell und irrtümlich erfolgen.

Der Signifikanztest berechnet als Entscheidungsgrundlage eine Irrtumswahrscheinlichkeit, die angibt, wie gut sich das Stichprobenergebnis mit den in der Nullhypothese postulierten Populationsverhältnissen vereinbaren läßt. Passen die Stichprobendaten gut zur Nullhypothese, wird eine hohe Irrtumswahrscheinlichkeit berechnet und die Nullhypothese beibehalten (*nicht-signifikantes Ergebnis*). Lassen sich die Stichprobendaten dagegen nur schwer mit der Nullhypothese vereinbaren, wird eine niedrige Irrtumswahrscheinlichkeit berechnet. Ist die Irrtumswahrscheinlichkeit extrem klein (kleiner als das Signifikanzniveau von 5%), wird die Nullhypothese verworfen und die Alternativhypothese angenommen (*signifikantes Ergebnis*).

Eine noch bessere Entscheidungsgrundlage hätte man freilich, wenn nicht nur geprüft würde, wie gut die Daten zur Nullhypothese passen (*α-Fehler-Wahrscheinlichkeit* bzw. Irrtumswahrscheinlichkeit), sondern auch, wie gut sie sich mit den in der Alternativhypothese formulierten Populationsverhältnissen vereinbaren lassen (*β-Fehler-Wahrscheinlichkeit*). Während im Signifikanztest-Ansatz von Fisher (1925) nur mit Nullhypothesen (und somit auch nur mit α-Fehler-Wahrscheinlichkeiten) operiert wurde, entwickelten Neyman und Pearson (1928) etwa zeitgleich ein Signifikanztest-Modell, das auch Alternativhypothesen und β-Fehler-Wahrscheinlichkeiten berücksichtigt. Im heute gängigen „Hybrid-Modell" (vgl. S. 27) werden Alternativhypothesen explizit formuliert, so daß – unter bestimmten Voraussetzungen – auch β-Fehler-Wahrscheinlichkeiten berechnet und in die Entscheidung für oder gegen eine Hypothese einbezogen werden können (vgl. Abschnitt 8.1.3).

1.3.2 Erkenntnisgewinn durch statistische Hypothesentests

Was leistet nun das Konzept der statistischen Hypothesenprüfung (dessen Erweiterung wir in Abschnitt 9.1 kennenlernen) für das Falsifikationsprinzip des kritischen Rationalismus? Erkenntniszugewinn – so lautet die zentrale Aussage – entsteht durch die Eliminierung falscher Theorien bzw. durch deren Falsifikation, d.h. also durch einen empirischen Ausleseprozeß, den nur bewährte Erklärungsmuster der aktuellen Realität „überleben", ohne dadurch das Zertifikat „wahr" oder „bewiesen" zu erlangen.

Falsifikation bedeutet, durch kritische Empirie die Untauglichkeit einer Theorie nachzuweisen. Dem entspricht im Kontext der statistischen Hypothesenprüfung ein nicht-signifikantes Ergebnis, also ein Ergebnis, bei dem konventionsgemäß die aus einer Theorie abgeleitete Forschungshypothese als nicht bestätigt gilt. Falsifizierende Untersuchungen sind damit Untersuchungen mit nicht-signifikanten Ergebnissen.

Wie auf S. 24f. ausgeführt, erfordert eine empirische Falsifikation jedoch nicht, die gesamte, der Hypothese zugrundeliegende Theorie abzulehnen. Bevor die „Kerntheorie" verworfen wird, sollte geprüft werden, ob die Ursache für den negativen Ausgang der Hypothesenprüfung möglicherweise im „Schutzgürtel der Hilfstheorien" zu finden ist, ob also Untersuchungsfehler wie z.B. ungeeignete operationale Indikatoren oder ungenaue Meßvorschriften für das nicht-signifikante Ergebnis verantwortlich sind.

Auch wenn Untersuchungsfehler dieser Art auszuschließen sind, besteht noch keine Notwendigkeit, die Theorie als ganze aufzugeben. Eine Reanalyse der Untersuchung könnte darauf aufmerksam machen, daß Teile der Untersuchungsergebnisse durchaus hypothesenkonform sind (einige Personen könnten hypothesenkonform reagiert haben), so daß sich evtl. die Möglichkeit zur exhaurierenden Erweiterung des Wenn-Teils der Theorie anbietet. Sollten jedoch weitere Untersuchungen (Replikationen; vgl. S. 41) erneut zu nicht-signifikanten Ergebnissen führen, dürfte die Theorie allmählich so stark belastet sein, daß sie letztlich aufgegeben werden muß.

Die möglicherweise verblüffende Behauptung nachzuvollziehen, daß gerade nicht-signifikante Ergebnisse unser Wissen erweitern, wird durch die Vorstellung erleichtert, daß falsche Theorien und Überzeugungen durch das Falsifikationsprinzip gezielt „entlarvt" werden, was als ein gewichtiger Beitrag zur Verhinderung wissenschaftlicher Fehlentwicklungen angesehen werden muß. Falsifikation führt freilich nicht nur zum Aussondern von Theorien, wodurch sich der Theorienfundus ja ständig verringern würde, sondern sie regt eben auch zur

Modifikation (z. B. durch Exhaustion, vgl. S. 25) sowie zur Neubildung von Theorien an.

Allerdings kann diese Art von Erkenntnisfortschritt nur greifen, wenn die Scientific Community in ausreichendem Maße für die Bekanntmachung falsifizierender Untersuchungsbefunde sorgt – eine Forderung, die angesichts einer heute vorherrschenden Publikationspraxis, die empirische Untersuchungen mit positiven bzw. signifikanten Resultaten begünstigt, zu wenig Beachtung findet (vgl. hierzu auch Abschnitt 9.4 zum Stichwort Metaanalyse).

Ein **signifikantes Ergebnis** ist nichts anderes als eine *Entscheidungsgrundlage* für die vorläufige Annahme der Forschungshypothese bzw. der geprüften Theorie. Jede andere Interpretation, insbesondere die Annahme, die Forschungshypothese sei durch ein signifikantes Ergebnis endgültig bestätigt oder gar bewiesen, wäre falsch, denn sie liefe auf einen mit dem Verifikationsmodell verbundenen Induktionsschluß hinaus, bei dem unzulässigerweise aufgrund einer begrenzten Anzahl theoriekonformer Ereignisse auf uneingeschränkte Gültigkeit der Theorie geschlossen wird.

Daß ein signifikantes Ergebnis nicht als endgültiger Beleg für die Richtigkeit der Forschungshypothese gewertet werden darf, verdeutlicht auch die Tatsache, daß das Risiko einer fälschlichen Annahme der Forschungs- bzw. Alternativhypothese angesichts der empirischen Ergebnisse bei statistischen Hypothesentests niemals völlig ausgeschlossen ist. Anders formuliert: Die Behauptung, das empirische Ergebnis könne niemals resultieren, wenn die Forschungshypothese nicht zuträfe, ist immer mit einem gewissen, wenn auch gelegentlich sehr kleinen statistischen Restrisiko (von maximal 5%) verbunden.

> ! *Statistische Signifikanz* liegt vor, wenn die empirisch ermittelte Irrtumswahrscheinlichkeit das konventionell festgelegte Signifikanzniveau (z. B. 1% oder 5%) unterschreitet. *Statistische Signifikanz* ist ein per Konvention festgelegtes Entscheidungskriterium für die vorläufige Annahme von statistischen Populationshypothesen.

Zu erwähnen ist eine weitere Besonderheit der statistischen Hypothesenprüfung: Ein einziges, einem Wenn-dann-Satz widersprechendes Ereignis sollte bei naturwissenschaftlich-deterministischen Gesetzesaussagen Zweifel an der Richtigkeit der Aussage auslösen. Diese Forderung ist berechtigt, solange die Untersuchungsobjekte, an denen die Wenn-dann-Aussage geprüft wird, prinzipiell austauschbar sind (etwa Spiegel oder Drähte in unseren Beispielen; allerdings sind auch in den modernen Naturwissenschaften statistische Hypothesen und Stichprobenuntersuchungen gängig; zu Wahrscheinlichkeitshypothesen in der Physik s. z. B. Popper, 1989, S. 152 ff.). Hat man es hingegen mit heterogenen Untersuchungsobjekten zu tun (Menschen, Tiere, Schulklassen etc.), ist diese Austauschbarkeit nicht gegeben. Hier ist eine statistische Hypothesenprüfung zweckmäßiger, die sich nicht auf ein einzelnes Untersuchungsobjekt, sondern auf eine *Stichprobe* von Objekten bezieht, die für diejenige *Population* von Untersuchungsobjekten repräsentativ ist (oder sein sollte), für die die Hypothese Gültigkeit beansprucht.

In dieser Stichprobe können sich nun durchaus mehrere der Hypothese widersprechende Einzelereignisse befinden; eine Falsifikation wird erst erforderlich, wenn widersprechende Einzelergebnisse in der Stichprobe „zu häufig" vertreten bzw. wenn die Abweichungen der Ergebnisse von den theoretischen Erwartungen „zu gravierend" sind.

Um zu kennzeichnen, daß statistische Hypothesen Aussagen über die Tendenz von Gruppen (z. B. Gruppenmittelwerten) und nicht über jeden Einzelfall machen, spricht man auch von *Aggregathypothesen*, d. h. die individuellen Daten der einzelnen Untersuchungsteilnehmer werden zu einem Gesamtwert zusammengefaßt (aggregiert) und erst über diesen Gesamtwert (*Aggregatwert*) werden Prognosen gemacht. Stellt sich etwa in einer empirischen Untersuchung heraus, daß der durchschnittliche Intelligenzwert von Linkshändern höher ist als der von Rechtshändern, ist daraus keinesfalls zu schlußfolgern, daß in der Untersuchung alle Linkshänder intelligenter waren als die Rechtshänder. Es wäre durchaus möglich (und ist sogar wahrscheinlich), daß sich auch einige Rechtshänder in der untersuchten Gruppe befanden, die intelligenter waren als viele Linkshänder. Diese intelligenten Rechtshänder befanden sich jedoch offesichtlich in der Minderheit, so daß sich ihre guten Ergebnisse im Zusammenhang mit der Gesamtgruppe „ausgemittelt" bzw. ausgeglichen haben.

Um Gruppenverhältnisse detailliert zu betrachten, ist der Mittelwert nur ein sehr grobes Maß. Bevor man Werte aggregiert, sollte man sich einen Eindruck von den Datenverhältnissen verschaffen (z.B. durch graphische Datenanalysen, vgl. Abschnitt 6.4.2), etwa um Verzerrungen von Mittelwerten durch Ausreißerwerte zu vermeiden. Unreflektiertes Aggregieren bzw. „Mitteln" von Werten ist einer der häufigsten methodischen Fehler (vgl. z. B. Sixtl, 1993) und liefert den Stoff für zahlreiche Statistiker-Witze der Art „Ein Jäger schießt auf einen Hasen. Der erste Schuß geht einen Meter links vorbei, der zweite Schuß geht einen Meter rechts vorbei. Statistisch ist der Hase tot." Erdfelder und Bredenkamp (1994) weisen darauf hin, daß es durchaus begründungsbedürftig ist, warum man einen Gruppenunterschied nur für den Aggregatwert und nicht für jedes einzelne Individuum prognostiziert, d.h. man sollte sich auch Gedanken darüber machen, wodurch hypothesenkonträre Einzelfälle zustandekommen könnten.

> **!** Bei einem *Aggregatwert* handelt es sich um die Zusammenfassung der Individualwerte einer Variablen über eine Gruppe von Untersuchungspersonen bzw. Untersuchungsobjekten hinweg. Obwohl die Bildung von Aggregatwerten statistisch immer möglich ist, sollte jeweils genau überprüft werden, ob Aggregatwerte im Kontext der konkreten Untersuchung a) inhaltlich sinnvoll interpretierbar sind und b) die Merkmalsverteilung innerhalb der Gruppe angemessen repräsentieren.

Im Unterschied zu einem Einzelergebnis, das der theoretischen Erwartung entweder entspricht oder widerspricht, haben wir es beim statistischen Hypothesentesten also mit einem Kontinuum zu tun, das unterschiedliche Grade der Hypothesenkonformität von Stichprobenergebnissen abbildet. Für eine genaue Bestimmung dessen, was unter „zu häufig" auftretenden, hypothesenkonträren Einzelfällen zu verstehen ist, hat die Scientific Community per Konventionsbeschluß eine Grenze festgelegt, die falsifizierende (d.h. nicht-signifikante) und vorläufig bestätigende (d.h. signifikante) Untersuchungsergebnisse voneinander trennt. Diese Grenze entspricht dem auf S. 30f. erwähnten *Signifikanzniveau*. Sie wurde so fixiert, daß unbegründete oder voreilige Schlußfolgerungen zugunsten der Forschungshypothese erheblich erschwert werden. Dies ist – wenn man so will – der Beitrag der statistischen Hypothesenprüfung zur Verhinderung wissenschaftlicher Fehlentwicklungen.

Allerdings kann die statistische Hypothesenprüfung – in verkürzter oder mißverstandener Form – auch Forschungsentwickungen begünstigen, die es eigentlich nicht wert sind, weiter verfolgt zu werden. Viele wissenschaftliche Hypothesen von der Art: „Es gibt einen Zusammenhang zwischen den Variablen X und Y" oder: „Zwei Populationen A und B unterscheiden sich bezüglich einer Variablen Z" sind sehr ungenau formuliert und gelten deshalb – bei sehr großen Stichproben – auch dann als bestätigt, wenn der Zusammenhang oder Unterschied äußerst gering ist.

Wir sprechen in diesem Zusammenhang von einer *Effektgröße*, die zwar statistisch signifikant, aber dennoch ohne *praktische Bedeutung* sein kann. Die Entwicklung einer Wissenschaft ausschließlich von signifikanten Ergebnissen abhängig zu machen, könnte also bedeuten, daß Theorieentwicklungen weiter verfolgt werden, die auf minimalen, wenngleich statistisch signifikanten Effekten beruhen, deren Erklärungswert für reale Sachverhalte eigentlich zu vernachlässigen ist. Wie die statistische Hypothesenprüfung mit dieser Problematik umgeht, werden wir ausführlicher in Kap. 9 behandeln.

Nachdem wir nun auf die Gefahr hingewiesen haben, kleine Effekte aufgrund ihrer Signifikanz *überzubewerten*, wollten wir aber auch noch auf die Gefahr hinweisen, kleine Effekte in ihrer Bedeutung zu *unterschätzen*. So werden beispielsweise empirische Studien, die parapsychologische Phänomene (z.B. Gedankenübertragung) nachweisen, oftmals mit dem Hinweis abgetan, es handele sich ja allenfalls um vernachlässigbar geringe Effekte. Diese Einschätzung ist sehr fragwürdig vor dem Hintergrund, daß etwa Aspirin als Mittel gegen Herzerkrankungen medizinisch verschrieben wird, obwohl der hierbei zugrundeliegende Effekt noch um den Faktor 10 geringer ist als der kleinste nachgewiesene parapsychologische Effekt (Utts, 1991).