

Erfahrung bei der Beurteilung der Wirksamkeit von Arzneimitteln

B. Schneider

Einleitung

Immanuel Kant beginnt seine 'Kritik der reinen Vernunft' [6] mit den Sätzen: "Daß alle unsere Erkenntnis mit der Erfahrung anfangt, daran ist gar kein Zweifel; denn wodurch sollte das Erkenntnisvermögen sonst zur Ausübung erweckt werden, geschähe es nicht durch Gegenstände, die unsere Sinne rühren und teils von selbst Vorstellungen bewirken, teils unsere Verstandestätigkeit in Bewegung bringen, diese zu vergleichen, sie zu verknüpfen oder zu trennen, und so den rohen Stoff sinnlicher Eindrücke zu einer Erkenntnis der Gegenstände zu verarbeiten, die Erfahrung heißt? Der Zeit nach geht also keine Erkenntnis in uns vor der Erfahrung vorher, und mit dieser fängt alle an.

Wenn aber gleich alle unsere Erkenntnis mit der Erfahrung anhebt, so entspringt sie darum doch nicht eben alle aus der Erfahrung. Denn es könnte wohl sein, daß selbst unsere Erfahrungserkenntnis ein Zusammengesetztes aus dem sei, was wir durch Eindrücke empfangen, und dem, was unser eigenes Erkenntnisvermögen ... aus sich selbst hergibt...

Es ist also wenigstens eine der näheren Untersuchung noch benötigte und nicht auf den ersten Anschein sogleich abzufertigende Frage: ob es ein dergleichen von der Erfahrung und selbst von allen Eindrücken der Sinne unabhängiges Erkenntnis gebe. Man nennt solche Erkenntnisse *a priori*, und unterscheidet sie von den *empirischen*, die ihre Quellen *a posteriori*, nämlich in der Erfahrung, haben."

Hier wird die Dualität der Erkenntnis angesprochen, die schon immer die Philosophen, die sich mit der Erkenntnisfrage beschäftigt haben, bewegt hat: Einerseits erhalten wir über die Welt nur durch Sinneseindrücke Kenntnis, andererseits werden diese Eindrücke aber nur durch das Medium unseres Verstandes zu Erkenntnissen. Wieviel vom 'wahren' Zustand der Welt wird durch unseren Verstand verschleiert oder verändert? Können wir überhaupt über die Sinneseindrücke ein 'wahres' Bild von der Welt erhalten oder sind die uns vom Verstand übermittelten Erkenntnisse nur 'Schatten', wie es Platon in seinem Höhlenbeispiel (im Buch 'Politeia') ausgedrückt hat, oder 'Schein', wie Parmenides, der in der 2. Hälfte des 6. Jahrhunderts v. Chr. in Elea in Unteritalien lebte, in einem uns nur bruchstückhaft überlieferten Lehrgedicht meinte? Es ist nicht meine Absicht, ausführlich auf das Erkenntnisproblem einzugehen. Hierüber sei auf die einschlägige Literatur, vor allem auf das Standardwerk von Ernst Cassirer [2] und (aus neuerer Zeit) auf die Grundzüge der Erkenntnislehre von Werner Flach [4] verwiesen, der die Rolle des 'Geltungsanspruchs' bei der Erkenntnis betont. Die Frage, die hier behandelt werden soll, ist die Frage nach der Zuverlässigkeit der Erkenntnisse, die wir aus der Erfahrung gewinnen. Diese Frage erhält eine neue Dimension, wenn die Erkenntnisse nicht nur passiv registriert, sondern aktiv zur Lebensgestaltung genutzt werden sollen. Es geht dann nicht nur um die Zuverlässigkeit der Erkenntnis als solcher, sondern um die Frage, ob und wie die aus Erfahrungen in bestimmten Situationen gewonnenen Erkenntnisse auf andere und vor allem zukünftige Situationen übertragen und so 'verallgemeinert' werden können. Eine solche Übertragung heißt **Induktion** und die gestellte Frage das **Induktionsproblem**. Diese Frage ist essentiell für jede empirische Wissenschaft und somit auch für eine wissenschaftliche Beurteilung der Wirksamkeit eines Arzneimittels. Ein Arzneimittel ist bei einer bestimmten Krankheit wirksam, wenn mehr Menschen, die an

dieser Krankheit leiden, mit dem Mittel eine Heilung oder Linderung erfahren als ohne das Mittel. Die Beurteilung der Wirksamkeit verlangt somit eine Verallgemeinerung der Erfahrungen, die bei der Behandlung der Krankheit mit dem Mittel und ohne das Mittel oder mit anderen Mitteln gemacht wurden, auf zukünftige Situationen und fällt so voll unter das Induktionsproblem. Eine erste, umfassende Formulierung des Induktionsproblems und erste Ansätze zu einer befriedigenden Lösung stammen von zwei Engländern und sind inzwischen mehr als 250 Jahre alt. Die Formulierung ist in der Abhandlung 'Enquiry concerning Human Understanding' von David Hume enthalten, deren erste Ausgabe 1742 erschien. Der erste Lösungsansatz stammt von Thomas Bayes und wurde 2 Jahre nach seinem Tod unter dem Titel 'An Essay towards solving a Problem in the Doctrine of Chances' 1763 von seinem Freunde Richard Price der Royal Society in London vorgelegt. Im folgenden sollen zunächst diese beiden Ansätze gebracht und kommentiert werden. Anschließend werden die Konsequenzen erörtert, die sich daraus für die Beurteilung der Wirksamkeit eines Arzneimittels ergeben.

1. Das Induktionsproblem

1.1 Die Rolle der Logik

Die Suche nach Verfahren oder Regeln, mit denen aus als wahr erkannten Sätzen die Wahrheit anderer Sätze hergeleitet werden kann, steht am Anfang der abendländischen Philosophie. Das von Aristoteles im 4. Jahrhundert v. Chr. entwickelte System der logischen Schlußweisen (Syllogismen) ist bis heute Grundlage rationalen (d.h. vernünftigen) Argumentierens. Dieses System gibt an, welche Sätze aus zwei gegebenen Prämissen (d.s. als wahr erkannte oder angenommene Sätze) als wahr gefolgert werden können. Als ein Beispiel für die erste und elementarste der insgesamt 19 gültigen Schlußweisen (modi) sei folgender Schluß (der im Mittelalter mit dem Kunstwort BARBARA belegt wurde) genannt. Die Prämissen sind: "Alle depressiven Menschen, die Johanniskraut einnehmen, werden geheilt" und "Frau X ist depressiv und hat Johanniskraut eingenommen". Daraus folgt die Konklusion: "Frau X ist geheilt". Zu beachten ist, daß zumindest eine der Prämissen ein genereller Satz (d.h. ein Satz, der mit dem Quantor 'Alle' oder 'Einige' gebildet ist) sein muß, dessen Subjekt oder Prädikat als Subjekt oder Prädikat in der anderen Prämisse vorkommt und in der Konklusion fehlt (sog. Mittelbegriff). Im Beispiel ist dies der Begriff 'Johanniskraut einnehmen'. Die Syllogismen gestatten somit nur die Deduktion der Wahrheit einer weniger allgemeinen Aussage aus der Wahrheit einer allgemeineren Aussage bzw. die Falschheit einer allgemeineren Aussage aus der Falschheit einer weniger allgemeinen Aussage. Ersteres ist die Grundlage der wissenschaftlichen Deduktion, die in der Form eines hypothetischen Schlusses nach dem modus ponens dargestellt werden kann; z.B.: Hypothese: "Wenn ein depressiver Mensch Johanniskraut einnimmt, wird er geheilt."; Feststellung: "Frau X war depressiv und hat Johanniskraut eingenommen"; Konklusion: "Frau X ist geheilt". Die Hypothese ist eine allgemeine Aussage bzw. ein als allgemeingültig angesehenes Gesetz (Naturgesetz) und beinhaltet eine Kausalbeziehung. Im Beispiel ist die Einnahme von Johanniskraut die Ursache, die stets die Heilung von der Depression als Wirkung zur Folge haben soll. Der modus ponens gestattet die Voraussage auf ein singuläres Ereignis (Heilung der Frau X), wenn bei diesem Ereignis die Ursache (Einnahme von Johanniskraut) vorliegt und die Hypothese wahr ist. Gerade darin liegt aber das Problem. Wie kann man feststellen, daß eine Hypothese, d.h. eine empirische Kausalbeziehung, allgemeingültig und damit ein Naturgesetz ist? Nach dem oben gebrachten Zitat von Kant fängt alle Erkenntnis mit der Erfahrung an, auch die Erkenntnis von

Naturgesetzen. Die Erfahrung liefert uns aber nur Kenntnis von singulären Ereignissen, d.h. einzelnen Tatsachen. Wie kann man von diesen Tatsachen zu einem allgemeinen Gesetz kommen?

1.2 Das Induktionsproblem bei David Hume

Mit dieser Frage hat sich vor mehr als 250 Jahren David Hume in der Schrift 'Enquiry concerning Human Understanding' ausführlich beschäftigt. Ich möchte seine Gedanken hier kurz skizzieren, trotz der Warnung Humes, daß die Beschäftigung mit diesem Thema (er spricht von der "genauen und unzugänglichen" Philosophie) bei "der Mehrzahl aller Menschen" nicht "den Vorzug behaupten" wird wie "die leichte und einleuchtende Philosophie" (die Zitate sind der deutschen Übersetzung bei Felix Meiner, Hamburg, 1973 [5] entnommen).

Zunächst stellt er fest, daß Tatsachen "der menschlichen Vernunft ... nicht in gleicher Weise als gewiß verbürgt" sind (wie die Sätze der Geometrie, Algebra und Arithmetik); "ebensowenig ist unsere Evidenz von ihrer Wahrheit, wenn auch noch so stark, von der gleichen Art wie bei der vorhergehenden. Das Gegenteil jeder Tatsache bleibt immer möglich, denn es kann niemals einen Widerspruch in sich schließen und wird vom Geist mit derselben Leichtigkeit und Deutlichkeit vorgestellt, als wenn es noch so sehr mit der Wirklichkeit übereinstimmte. Daß die Sonne morgen nicht aufgehen wird, ist ein nicht minder verständlicher Satz und nicht widerspruchsvoller, als die Behauptung, daß sie aufgehen wird...." (S. 35/36).). Wenn wir trotzdem aus der Beobachtung, daß zwei oder mehr selbständige Tatsachen häufig sich in zeitlicher und räumlicher Nähe ereignen, einen Zusammenhang zwischen diesen Tatsachen herstellen, eine oder mehrere als Ursachen und die übrigen als Wirkungen bezeichnen und diese Beziehung als 'Kausalbeziehung' auch bei künftigen Ereignissen erwarten, so muß es dafür ein anderes als das logische Prinzip geben. "Dies Prinzip ist **Gewohnheit** oder **Übung**. Wo immer die Wiederholung einer bestimmten Handlung oder Tätigkeit die Neigung hervorruft, dieselbe Handlung oder Tätigkeit ohne irgend einen Anstoß durch einen Denkkakt oder Verstandesvorgang, zu erneuern: da sagen wir stets, diese Neigung sei die Wirkung der *Gewohnheit*"... (S.55). "Was ist nun das Schlußergebnis von alledem? Ein einfaches – wenn auch allerdings recht weit ab von den gewöhnlichen Theorien der Philosophie. Aller Glaube an Tatsachen oder wirkliches Sein stammt lediglich von irgend einem Gegenstand, der dem Gedächtnis oder den Sinnen gegenwärtig ist, und von einem gewohnheitsmäßigen Zusammenhang zwischen diesem und einem anderen Gegenstande..." (S. 59).

1.3 Das Falsifikationsprinzip bei Karl Raimund Popper

Viele wollen sich mit diesem Schlußergebnis, daß die Naturgesetze nichts anderes als das Ergebnis von Gewohnheit oder Übung seien, nicht abfinden. Sie glauben, daß diesen Gesetzen "Erkenntnisse a priori" (wie es Kant ausgedrückt hat [6]) zugrunde liegen, die gewissermaßen angeboren und deshalb stets wahr sind. Manche verfallen in einen primitiven Induktionismus und glauben, wenn wiederholte Erfahrungen auf einen Zusammenhang zwischen Tatsachen hinweisen, daß damit schon eine allgemeingültige Kausalbeziehung bewiesen sei. Gegen diesen Aberglauben hat sich vehement Karl Raimund Popper in seinem Buch 'Logik der Forschung' [14] gewandt. Auch wenn millionenfach die Erfahrung einen Zusammenhang zwischen Tatsachen aufzeigt (z.B. jeden Morgen die Sonne aufgegangen ist), kann daraus nicht mit Sicherheit logisch geschlossen werden, daß beim nächsten Mal dieser Zusammenhang wieder vorliegen muß (am nächsten Morgen wieder die Sonne aufgeht). Allerdings wird in den Naturwissenschaften zur Herleitung und Begründung

von Kausalgesetzen im allgemeinen nicht der primitive Induktionismus, sondern ein verfeinerter Induktionismus angewandt, der darin besteht, daß die beobachteten Zusammenhänge in ein Theoriensystem eingebettet und im Rahmen dieses Systems aus einfachen Grundgesetzen hergeleitet und so erklärt werden. Beim Beispiel des Sonnenaufgangs wird die Zuversicht, daß morgen die Sonne wieder aufgehen wird, nicht einfach mit der wiederholten Erfahrung begründet, sondern durch die Rotation der Erde um sich selbst und die Bewegung der Erde um die Sonne auf einer schwach elliptischen Bahn. Diese Bewegungen werden zurückgeführt auf die Bewegungsgesetze der Mechanik und auf das Massenwirkungsgesetz. Diese Gesetze lassen sich aber nicht weiter logisch und mathematisch deduzieren. Sie müssen entweder als 'Axiome' oder 'a priorische Erkenntnisse' ohne weitere Erklärung geglaubt oder induktionistisch mit der wiederholten Erfahrung (nicht nur bei der Bewegung der Erde und Sonne, sondern bei allen beobachteten Bewegungen von Körpern) begründet werden. Letztlich kann also auch der verfeinerte Induktionismus keine logische und damit allgemeingültige Begründung für Kausalgesetze liefern, sondern ist auf 'Gewohnheit und Übung' angewiesen. Einstein spricht in diesem Zusammenhang von der "auf Einfühlung in die Erfahrung sich stützende Intuition" (zitiert nach [14], S.7).

Karl Raimund Popper hat in seiner 'Logik der Forschung' versucht, den Induktionismus durch das Prinzip der "deduktiven Überprüfung von Theorien" zu überwinden bzw. zu ersetzen: "Die Methode der kritischen Nachprüfung, der Auslese der Theorien, ist nach unserer Auffassung immer die folgende: Aus der vorläufig unbegründeten Antizipation, dem Einfall, der Hypothese, dem theoretischen System, werden auf logisch-deduktivem Weg Folgerungen abgeleitet; diese werden untereinander und mit anderen Sätzen verglichen, indem man feststellt, welche logischen Beziehungen (z.B. Äquivalenz, Ableitbarkeit, Vereinbarkeit, Widerspruch) zwischen ihnen bestehen. Dabei lassen sich insbesondere vier Richtungen unterscheiden, nach denen die Prüfung durchgeführt wird: der logische Vergleich der Folgerungen untereinander, durch den das System auf seine innere Widerspruchslosigkeit hin zu untersuchen ist; eine Untersuchung der logischen Form der Theorie mit dem Ziel, festzustellen, ob es den Charakter einer empirisch-wissenschaftlichen Theorie hat, also z.B. nicht tautologisch ist; der Vergleich mit anderen Theorien, um unter anderem festzustellen, ob die zu prüfende Theorie, falls sie sich in den verschiedenen Prüfungen bewähren sollte, als wissenschaftlicher Fortschritt zu bewerten wäre; schließlich die Prüfung durch 'empirische Anwendung' der abgeleiteten Folgerungen. Diese letzte Prüfung soll feststellen, ob sich das Neue, das die Theorie behauptet, auch praktisch bewährt, etwa in wissenschaftlichen Experimenten oder in der technisch-praktischen Anwendung. Auch hier ist das Prüfungsverfahren ein deduktives: Aus dem System werden (unter Verwendung bereits anerkannter Sätze) empirisch möglichst leicht nachprüfbar bzw. anwendbare singuläre Folgerungen ('Prognosen') deduziert und aus diesen insbesondere jene ausgewählt, die aus bekannten Systemen nicht ableitbar sind bzw. mit ihnen in Widerspruch stehen. Über diese – und andere – Folgerungen wird nun im Zusammenhang mit der praktischen Anwendung, den Experimenten usw., entschieden. Fällt die Entscheidung positiv aus, werden die singulären Folgerungen anerkannt, *verifiziert*, so hat das System die Prüfung vorläufig bestanden; wir haben keinen Anlaß, es zu verwerfen. Fällt eine Entscheidung negativ aus, werden Folgerungen *falsifiziert*, so trifft ihre Falsifikation auch das System, aus dem sie deduziert wurden" ([14], S. 7/8).

Das Falsifikationskriterium Poppers stützt sich auf den hypothetischen Schluß 'modus tollens' der klassischen Logik, dessen Schema lautet: Hypothese: 'Alle A haben die Eigenschaft E'. Feststellung: 'Ein a aus A hat nicht die Eigenschaft E'. Konklusion: 'Nicht alle A haben die Eigenschaft E'. Damit ist ein logisches Kriterium angegeben, mit dem eine Theorie oder ein allgemeiner Satz als nicht gültig erwiesen werden kann. Wenn Popper aber meint, daß er damit das Induktionsproblem gelöst hat ("Of course, I may be mistaken; but I think that I have solved a major philosophical problem: the problem of induction... This solution has been extremely fruitful, and it has enabled me to solve a good number of other philosophical problems" ([15], S.1)), dann hat er den Mund etwas zu voll genommen. Dieses Problem besteht ja nicht darin, logisch die Falschheit einer Theorie nachzuweisen, sondern ihre Wahrheit, d.h. Allgemeingültigkeit; und das kann logisch nicht nachgewiesen werden.

Das Falsifikationskriterium ist trotzdem für die empirische Wissenschaft äußerst wertvoll und notwendig; und zwar nicht als Lösung des Induktionsproblems sondern als **Abgrenzungskriterium**, mit dem sich die empirische Wissenschaft von anderen Formen der Welterklärung wie Dichtung, Mythen, Dogmen oder Aberglauben (Popper faßt dies alles unter der Bezeichnung 'Metaphysik' zusammen) abgrenzen läßt. Eine empirische Theorie soll nur dann als wissenschaftlich gelten, wenn sie mit der Erfahrung nachprüfbar ist und gegebenenfalls durch die Erfahrung falsifiziert werden kann. Es stellt sich dann die Frage nach dem Charakter dieser Erfahrung, die zur Nachprüfung von Theorien verwendet werden soll. Popper nennt diese Erfahrung die 'empirische Basis' oder 'Basissätze' einer wissenschaftlichen Theorie. Wesentliche Bedingung für Basissätze ist die Objektivität; d.h. die Sätze "müssen grundsätzlich von jedermann nachgeprüft und eingesehen werden können", sie müssen "*intersubjektiv* nachprüfbar sein" ([14], S.18). Dies bedeutet, daß die Basissätze, bzw. genauer die Erfahrungsgrundlage (Experimente oder Beobachtungen) der Basissätze, zur Überprüfung wissenschaftlicher Theorien **wiederholbar** sein müssen und genaue Vorschriften für die Durchführung der Experimente und Beobachtungen vorliegen. Dies schließt rein subjektive Überzeugungserlebnisse oder 'Intuition' als empirische Basis zur Nachprüfung wissenschaftlicher Theorien aus, so wichtig diese auch bei der Aufstellung und Akzeptanz von Theorien sind. Bei der Forderung nach Objektivität der Basissätze stellt sich aber sofort das Problem der eingangs geschilderten Dualität zwischen Sinneseindrücken (ich bezeichne sie als 'Realität') und Erkenntnisinhalten (ich bezeichne sie als 'Welt'). Nur über das Medium der subjektiven und damit auch psychologischen Verstandeskräfte können wir aus den Sinneseindrücken, die wir von der Realität empfangen, zu Erkenntnissen über die Welt kommen. Diese subjektiven, psychologischen Verstandeskräfte sind nicht nur angeboren, sondern haben sich auch (z.B. durch Lernen) entwickelt, wobei sie sowohl von Überzeugungserlebnissen oder Intuitionen als auch von Theorien und insbesondere auch von der zu überprüfenden Theorie beeinflusst sind. Popper hat diese Auffassung so ausgedrückt, "daß Beobachtungen und erst recht Sätze über Beobachtungen und über Versuchsergebnisse immer *Interpretationen* der beobachteten Tatsachen sind und daß sie *Interpretationen im Lichte von Theorien* sind" (S. 72). Wie kann man da noch von Objektivität sprechen?

Popper versucht diesem Dilemma durch die Unterscheidung zwischen "der objektiven Wissenschaft" und "unserem Wissen" zu begegnen: "Sicher kann uns nur Beobachtung 'ein Wissen über die Tatsachen liefern', können wir [wie Hahn sagt] 'Tatsachen ... nur durch Beobachtung erfassen'. Aber dieses unser Wissen, unser Erfassen begründet nicht die Geltung von Sätzen. Die Fragestellung der Erkenntnistheorie

kann daher nicht sein: '... worauf geht *unser Wissen* zurück? ..., genauer: womit kann ich, wenn ich das *Erlebnis* S gehabt habe, meine ... Erkenntnis ... begründen, gegen Zweifel rechtfertigen?'... ; sondern wir werden fragen: Durch welche intersubjektiv nachprüfbar Folgerungen sind die wissenschaftlichen Sätze überprüfbar?" (S. 64) Die Mittel zur Überprüfung wissenschaftlicher Sätze sind nach Popper die Basissätze; d.h. singuläre 'Es-gibt-Sätze', die eine **beobachtbare** Aussage ausdrücken. Der Subjektivität des Beobachtens (d.h. die subjektive Interpretation der Beobachtung) und der Auswahl der Basissätze versucht Popper dadurch zu entgehen, daß er intersubjektive Festsetzungen fordert, mit denen festgelegt wird, welche Basissätze anerkannt werden und wie die Beobachtungen durchzuführen und zu interpretieren sind: "Die Basissätze werden durch Beschluß, durch Konvention anerkannt, sie sind *Festsetzungen*.... Die Festsetzung der Basissätze erfolgt anlässlich einer *Anwendung* der Theorie und ist ein Teil dieser Anwendung, durch die wir die Theorie *erproben*; wie die Anwendung überhaupt, so ist die Festsetzung ein durch theoretische Überlegungen geleitetes planmäßiges Handeln" (S. 71), sie entspricht "... einer (methodisch geregelten) *Beschlußfassung* ..." (S. 74).... "So ist die empirische Basis der objektiven Wissenschaft nichts 'Absolutes'; die Wissenschaft baut nicht auf Felsengrund. Es ist eher ein Sumpfland, über dem sich die kühne Konstruktion ihrer Theorien erhebt; sie ist ein Pfeilerbau, dessen Pfeiler sich von oben her in den Sumpf senken – aber nicht bis zu einem natürlichen, 'gegebenen' Grund: Denn nicht deshalb hört man auf, die Pfeiler tiefer hineinzutreiben, weil man auf eine feste Schicht gestoßen ist: wenn man hofft, daß sie das Gebäude tragen werden, beschließt man, sich vorläufig mit der Festigkeit der Pfeiler zu begnügen." (S. 75/76)

1.4 Die Wissenschaftsauffassung von Thomas Kuhn

Objektivität der Basissätze soll also dadurch erreicht werden, daß die Subjektivität des Erkennens durch die Intersubjektivität der Gemeinschaft ersetzt wird. Mit dieser Auffassung nähert sich Popper der Wissenschaftssicht von Thomas S. Kuhn ([7], [8]), der das **Paradigma** als Kennzeichen der Wissenschaft eingeführt hat. Der Ausdruck Paradigma "wird in enger Nachbarschaft (örtlicher wie logischer) zum Ausdruck **wissenschaftliche Gemeinschaft** eingeführt. Ein Paradigma ist das, was den Mitgliedern einer wissenschaftlichen Gemeinschaft, und nur ihnen, gemeinsam ist" ([8] S.390). Das Paradigma ist die **disziplinäre Matrix** der wissenschaftlichen Gemeinschaft. "Zu den Bestandteilen der disziplinären Matrix gehören alle oder die meisten Gegenstände von Gruppenfestlegungen" (S.392), also auch die Festsetzungen der Basissätze. Diese Sicht fragt zunächst nicht danach, was Wissenschaft ist, sondern wie Wissenschaft betrieben wird. Der Wissenschaftsbetrieb unterscheidet sich formal nicht von anderen sozialen Gebilden wie Wirtschafts- oder Handelsgemeinschaften oder Vereinen und wird im wesentlichen durch Handlungsnormen und Verhaltensregeln bestimmt, die sich die Gemeinschaft (genauer die Wortführer der Gemeinschaft, die 'Päpste' oder 'Muhammed Alis' mit 'the biggest mouth and strongest fists') gegeben hat. Die Einhaltung dieser Normen wird von der Gemeinschaft eifersüchtig überwacht, Verstöße dagegen mit Sanktionen bestraft. Solange diese Normen eingehalten werden und die Führungsrolle der Wortführer nicht in Frage gestellt wird, handelt es sich um eine 'ruhende Wissenschaft' (resting science). Es kommt aber immer wieder vor, daß von Einzelnen die Normen in Frage gestellt und neue Normen eingeführt werden. Wenn dies eindringlich genug geschieht, entsteht eine 'wissenschaftliche Revolution' (scientific revolution) und wenn sich die Revolution durchsetzt, kommt es zu einem Paradigmawechsel. Die Gemeinschaft etabliert sich mit einem neuen Paradigma und verharrt dort, bis es wieder zur Revolution kommt. Man könnte dieses Bild von der Wissenschaft etwas despektierlich als 'Jahrmarkt der Eitelkeiten' bezeichnen,

wobei sich die verschiedenen Paradigmen durch verschiedene Eitelkeiten unterscheiden. Man muß aber hinzusetzen, daß es Kuhn nicht nur bei dieser oberflächlichen Beschreibung des Wissenschaftsbetriebs beläßt, sondern nach Kennzeichen der Paradigmen (resp. wissenschaftlichen Eitelkeiten) fragt. Er führt drei solcher Kennzeichen auf: **symbolische Verallgemeinerungen**, das "sind diejenigen Ausdrücke, die von der Gruppe ohne Zögern angewandt werden und sich leicht auf eine logische Form wie $(x) (y) (z) \phi(x,y,z)$ bringen lassen. Es sind die formalen und leicht formalisierbaren Bestandteile der disziplinären Matrix"; **Modelle**, die "der Gruppe bevorzugte Analogien oder, wenn sie von großer Überzeugungskraft getragen sind, eine Ontologie" liefern, und **Musterbeispiele**, das "sind konkrete Problemlösungen, die von der Gruppe in einem ganz gewöhnlichen Sinn als paradigmatisch anerkannt sind" ([8] S. 392/393). Die symbolischen Verallgemeinerungen und Modelle kann man als 'Theorie' und die Musterbeispiele als 'Basissätze' im Sinne Poppers identifizieren. Auch in dieser Hinsicht unterscheidet sich Kuhns Wissenschaftsauffassung nicht prinzipiell von der Poppers.

Am Schluß dieser wissenschaftstheoretischen Überlegungen möchte ich kurz die Frage streifen, ob und wann die Naturheilverfahren oder sonstigen alternativen Therapieformen (Homöopathie, Akupunktur u.ä.) als Wissenschaft angesehen werden können. Legt man oberflächlich Kuhns Wissenschaftsauffassung zugrunde, so ist die Frage auf jeden Fall zu bejahen; denn diejenigen, die diese Verfahren betreiben, haben sich zu Gemeinschaften etabliert, die sich wissenschaftlich nennen und entsprechende Paradigmen besitzen. Legt man Poppers Wissenschaftsauffassung zugrunde, so kommt es darauf an, inwieweit die alternativen Therapieverfahren auf logisch konsistenten Theorien beruhen, die durch beobachtbare Basissätze überprüft werden können; inwieweit also Poppers Abgrenzungskriterium erfüllbar ist. Die Festsetzung der Basissätze und ihre Überprüfung können intersubjektiv durch die Gruppe selbst erfolgen. Wenn sie aber auch von den Angehörigen der 'etablierten' Medizin anerkannt werden wollen, ist es empfehlenswert, diese Festsetzungen gemeinsam zu treffen.

2. Wahrscheinlichkeit

Seit David Hume ist bekannt, daß es für die Verallgemeinerung der Erfahrung auf künftige Ereignisse kein logisches, d.h. immer gültiges Verfahren gibt, sondern nur Gewohnheit und Übung. Der Wirksamkeitsnachweis für Arzneimittel verlangt eine solche Verallgemeinerung. Müssen wir uns beim Wirksamkeitsnachweis nur mit dem Hinweis auf die Gewohnheit, auf die 'traditionelle Anwendung' begnügen, wie dies noch oft in der Phytotherapie und bei den alternativen Therapieverfahren geschieht? Die Frage ist zu verneinen, wenn die Forderung der Allgemeingültigkeit abgeschwächt wird. Popper hat in der 'Logik der Forschung' auch auf diese Möglichkeit hingewiesen, wenn er in Bezug auf das "Humesche Problem der Induktion" schreibt: "Die Wurzel dieses Problems ist der scheinbare Widerspruch zwischen der 'Grundthese jedes Empirismus' – der These, daß nur 'Erfahrung' über empirisch-wissenschaftliche Aussagen entscheiden kann – und der Humeschen Einsicht in die Unzulänglichkeit induktiver Beweisführungen. Dieser Widerspruch besteht nur dann, wenn man postuliert, daß alle empirisch-wissenschaftlichen Sätze 'vollentscheidbar', d.h. verifizierbar *und* falsifizierbar sein müssen. Hebt man dieses Postulat auf, läßt man als empirisch auch 'teilentscheidbare', einseitig falsifizierbare Sätze zu, die durch methodische Falsifikationsversuche überprüft werden können, so verschwindet der Widerspruch" ([14] S.16). Die Methode der Falsifikation kann aber – wie wir ge-

sehen haben – auch bei Reduktion des Anspruchs der Allgemeingültigkeit das Humesche Problem nicht lösen. Benötigt wird vielmehr eine Methode, die auch bei Versagen der Falsifikation eine objektive Aussage über den Grad der Allgemeingültigkeit der Theorie gestattet. Auf diese Methode hat Hume in der 'Enquiry' selbst hingewiesen, indem er schreibt: "Werden wir also durch Begründungen veranlaßt, vergangener Erfahrung zu vertrauen und sie zum Maßstab unserer künftigen Urteile zu nehmen, so können diese Begründungen nur wahrscheinliche, d.h. solche sein, welche nach der obigen Einteilung Tatsachen und wirkliches Dasein betreffen" ([5], S. 46). Es ist also die **Wahrscheinlichkeit**, die statt der Allgemeingültigkeit der Theorie zu fordern ist, um das Humesche Problem der Induktion zu lösen.

2.1 Interpretation und Definition der Wahrscheinlichkeit

Die dafür notwendige Interpretation und Präzisierung des Wahrscheinlichkeitsbegriffs erfolgte ca. 100 Jahre vor der Erstpublikation der 'Enquiry'; und zwar 1654 im Briefwechsel zwischen Blaise Pascal und Pierre de Fermat über ein Problem des Chevalier de Méré bezüglich der Verteilung des Spielgewinns bei vorzeitigem Abbruch eines Glücksspiels. In diesem Briefwechsel wurde das Wort 'Wahrscheinlichkeit' (Anschein der Wahrheit) von dem negativen Beigeschmack befreit, der ihm im Altertum und Mittelalter als 'Vortäuschen der Wahrheit' anhaftete, und quantitativ als 'Grad der Erwartung' (*l'esperance*) für das Eintreffen zukünftiger Ereignisse aufgefaßt. Es wurden Rechenregeln für den Umgang mit diesem quantitativen Begriff angegeben. Die damit eingeführte Wahrscheinlichkeitsrechnung wurde rasch weiterentwickelt und ihre Bedeutung für alle Bereiche des Lebens (nicht nur für Glücksspiele) erkannt. In dem 1713 in Basel posthum erschienenen Buch 'Ars conjectandi' von Jakob Bernoulli (J. Bernoulli war 1708 verstorben) werden im vierten Teil bereits ausführlich "Anwendungen der vorhergehenden Lehre auf bürgerlich, sittliche und wirtschaftliche Verhältnisse" erörtert. Die Wahrscheinlichkeit wird dabei als "ein Grad der Gewissheit" definiert, der "sich von ihr wie ein Theil vom Ganzen" unterscheidet (zitiert nach der Übersetzung von R. Haussner, 1899, in Ostwald's Klassiker der exakten Wissenschaften Nr. 108). Eine spätere Definition von Pierre Simon Laplace (im *Essai Philosophique sur les Probabilités*, 1812) erklärt die Wahrscheinlichkeit für ein Ereignis als 'das Verhältnis der für das Ereignis günstigen zu den gleichmöglichen Fällen'. Diese Definition wird heute noch gelegentlich benutzt. An ihr ist aber zu kritisieren, daß der Begriff 'gleichmöglich' nur sinnvoll als 'gleichwahrscheinlich' interpretiert werden kann und somit eine Definition der Wahrscheinlichkeit bereits voraussetzt und zusätzlich die Wahrscheinlichkeit auf gleichwahrscheinliche Fälle einschränkt. Diese Kritik wurde vor allem von Richard von Mises (in dem Artikel 'Grundlagen der Wahrscheinlichkeitsrechnung', *Math. Zeitschr.*, 4, 1919) vorgebracht, der darauf hinwies, daß zwar bei den Glücksspielen die Gleichwahrscheinlichkeit der primären Spielergebnisse (z.B. die zu würfelnde Augenzahl) angenommen werden kann, in der Naturwissenschaft aber meistens nicht. Er führte daher den Begriff des 'Kollektivs' ein als die Gesamtheit der möglichen Ergebnisse, die in der Folge eines unendlich oft wiederholten Experiments oder einer Beobachtung vorkommen können, und definierte die Wahrscheinlichkeit für ein bestimmtes Ergebnis als den Grenzwert der relativen Häufigkeiten dieses Ergebnisses in der Folge der Beobachtungen. Diese Definition setzt aber voraus, daß die Folge der relativen Häufigkeiten konvergiert und daher die wiederholten Experimente oder Beobachtungen bestimmten Bedingungen genügen müssen, die R. von Mises in einem 'Auswahlaxiom' zu präzisieren versuchte. Dieses Auswahlaxiom ist aber logisch nicht widerspruchsfrei, so daß auch die Definition von R. von Mises nicht befriedigt. Eine Lösung des Problems der Wahrscheinlichkeitsdefinition wird erreicht, wenn man zwischen der *formalen axiomatischen Definition*, die

nur die Form festlegt, wie mit Wahrscheinlichkeiten zu rechnen ist, den Inhalt aber offenläßt, der *Interpretation*, die keiner axiomatischen Strenge genügen muß, und der *praktischen Anwendung (Pragmatik)* unterscheidet. Eine voll befriedigende axiomatische Definition der Wahrscheinlichkeit wurde 1933 von A. Kolmogoroff (in dem Buch: 'Grundbegriffe der Wahrscheinlichkeitsrechnung') gegeben. Dabei werden die Ereignisse oder Ergebnisse, denen eine Wahrscheinlichkeit zugeordnet werden soll, als Teilmengen einer umfassenden Menge (dem Wahrscheinlichkeitsraum Ω) aufgefaßt, die ein meßbares Mengenfeld F bilden (d.h. mit je zwei Teilmengen A und B gehört auch die Vereinigung $A \cup B$ zum Mengenfeld). Die Wahrscheinlichkeit wird definiert als ein Maß $P(\cdot)$ über die Mengen von F , das folgenden Axiomen genügt: a) $0 \leq P \leq 1$, b) $P(\Omega) = 1$ und c) für zwei disjunkte Mengen A und B aus F , die keine gemeinsamen Elemente haben, gilt: $P(A \cup B) = P(A) + P(B)$. Aus diesen Axiomen können alle Rechenregeln für Wahrscheinlichkeiten widerspruchsfrei hergeleitet werden. Die allgemeinste Interpretation der Wahrscheinlichkeit wurde bereits erwähnt: Wahrscheinlichkeit ist zu interpretieren als der *Grad der Gewißheit*, der unbekanntem Aussagen oder Ereignissen aufgrund bekannter Tatsachen (Erfahrung) zukommt. Dieser Grad der Gewißheit kann entweder 'subjektiv' interpretiert werden, als persönliche Gewißheit, die z.B. durch die Wettquote für das unbekanntem Ereignis ausgedrückt werden kann, oder 'objektiv' (bzw. 'frequentistisch') als die Häufigkeit, mit der das Ereignis oder Ergebnis in der hypothetischen 'Grundgesamtheit' aller möglichen Ereignisse oder Ergebnisse vorkommt. Da es sich nur um eine Interpretation handelt, muß die 'Wettquote' oder 'Grundgesamtheit' nicht weiter präzisiert werden. Entscheidend ist, daß die axiomatisch festgelegten Rechenregeln eingehalten werden und pragmatische Regeln vorgegeben sind, nach denen die Wahrscheinlichkeit mit beobachteten Daten verknüpft werden kann. Darauf wird im folgenden näher eingegangen.

2.2 Anwendung der Wahrscheinlichkeit (Statistik)

Bei der Anwendung der Wahrscheinlichkeit zur Lösung des Humeschen Problems ist der allgemeine Satz (die Theorie) als Wahrscheinlichkeitssatz zu formulieren; z.B. (um auf das bei der Demonstration eines Schlusses verwendete Johanniskraut zurückzukommen): 'Nach Einnahme von Johanniskraut bei Depression ist die Wahrscheinlichkeit einer Heilung mindestens 60%' oder 'Nach Einnahme von Johanniskraut bei Depression ist die Wahrscheinlichkeit einer Heilung größer als ohne diese Therapie'. Diese Hypothesen sind mit den Erfahrungen zu vergleichen, die an depressiven Patienten bei der Behandlung mit oder ohne Johanniskraut gemacht wurden. Das Ergebnis dieses Vergleichs kann aber nicht einfach das Diktum: 'die Hypothese ist wahr' oder 'sie ist falsch' sein, da der allgemeine Satz keine Wahrheit, sondern nur den 'Anschein der Wahrheit' beansprucht. Es wird vielmehr auch eine Wahrscheinlichkeitsaussage sein, die die Zuverlässigkeit der Hypothese präzisiert.

2.2.1 Bayesianische Statistik

Ein exaktes Verfahren, mit dem aus Beobachtungen auf den Grad der Zuverlässigkeit von Wahrscheinlichkeitsaussagen geschlossen werden kann, hat Thomas Bayes in der Schrift 'An Essay towards solving a Problem in the Doctrine of Chances' dargelegt, die zwei Jahre nach seinem Tod von seinem Freund Richard Price der Royal Society vorgelegt und dort am 23. Dezember 1763 vorgelesen wurde [1]. Das Problem, das in dieser Schrift behandelt wird, stellt sich am Beispiel des Johanniskrauts folgendermaßen dar: Es wurde eine Anzahl n (z.B. 10) von depressiven Patienten mit Johanniskraut behandelt, wovon x (z.B. 8) geheilt wurden. Was läßt sich aus dieser Erfahrung über die Heilungswahrscheinlichkeit aussagen? Das umgekehrte Problem,

bei Kenntnis der Heilungswahrscheinlichkeit (die mit π bezeichnet werden soll) die Wahrscheinlichkeit zu berechnen, mit der bei n Anwendungen x -mal eine Heilung zu erwarten ist, wurde bereits im Briefwechsel zwischen Pascal und Fermat gelöst. Die Formel dafür, die heute als Binomialwahrscheinlichkeit bezeichnet wird und bereits von Isaak Newton hergeleitet worden war, lautet: $p(x|\pi, n) = {}_x C_n \pi^x (1-\pi)^{n-x}$, wobei ${}_x C_n$ der Binomialkoeffizient ist, der die Anzahl der Möglichkeiten angibt, mit der x aus n Objekten (ohne Berücksichtigung der Anordnung) ausgewählt werden können. Beim Problem von Bayes bekommt diese Formel eine neue Bedeutung. Bei diesem Problem ist die Zahl x bekannt; die Beobachtungen wurden ja bereits durchgeführt. Unbekannt ist aber die Wahrscheinlichkeit π , die diesen Beobachtungen zugrunde liegt. Der Ausdruck $p(x|\pi, n)$ stellt daher bei diesem Problem keine Wahrscheinlichkeit für x dar (Bekanntem kann man keine Wahrscheinlichkeit zuordnen), sondern eine Funktion, mit der das bekannte x (ein 'Datum', d.h. etwas 'Gegebenes') mit der unbekannt Wahrscheinlichkeit π verknüpft wird. Diese Funktion ist das gesuchte Bindeglied zwischen beobachteten singulären Tatsachen (x) und einem allgemeinen Wahrscheinlichkeitsgesetz (das in dem hier betrachteten einfachen Fall durch die Wahrscheinlichkeit π gegeben ist). Alle induktiven Schlüsse von Beobachtungen auf Wahrscheinlichkeitsgesetze gründen sich auf diese Funktion. Ronald Aimler Fisher hat ihr 1912 den Namen **Likelihood** gegeben, der für sie heute allgemein verwendet wird. Die Likelihood für verschiedene Werte von π ist in Abbildung 1 für den Fall dargestellt, daß von 10 Patienten 8 geheilt wurden. Die Likelihood ist für den Wert π , der der beobachteten Häufigkeit 0,8 ($=x/n$) entspricht, am größten.

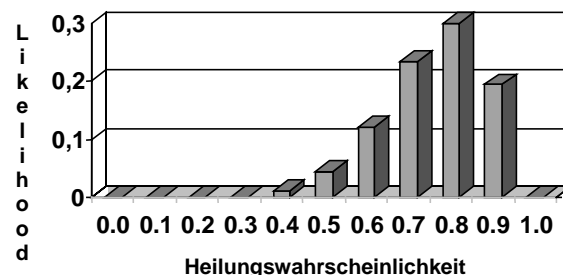


Abbildung 1
Likelihood für die Heilungswahrscheinlichkeit
bei 8 Heilungen in 10 Behandlungen

Bayes Lösung des Induktionsproblems bestand darin, daß er die Wahrscheinlichkeit π als Zufallsgröße ansah (da π ja nicht bekannt ist) und vor Durchführung der Beobachtung ihren möglichen Werten 'a priori' Wahrscheinlichkeiten $p(\pi)$ zuordnete, die den Grad der Gewißheit vor der Beobachtung präsentieren. Durch die Beobachtung wird dieser Grad zur 'a posteriori' Wahrscheinlichkeiten $p(\pi|x, n)$ verändert, wobei das Verhältnis von a posteriori zur a priori Wahrscheinlichkeit proportional zur Likelihood $p(x|\pi, n)$ ist: $p(\pi|x, n) \propto p(x|\pi, n)p(\pi)$ (Theorem von Bayes).

Thomas Bayes und sein Freund Richard Price waren sich der Bedeutung dieses Theorems als Lösung des Induktionsproblems in wahrscheinlichkeitstheoretischer Formulierung durchaus bewußt. Im Begleitbrief schreibt Price: "Now I send you an essay which I have found among papers of our deceased friend Mr. Bayes In an introduction which he has writ to this Essay, he says, that his design at first in thinking on the subject of it was, to find out a method by which we might judge concerning the probability that an event has to happen, in given circumstances, upon supposition that we know nothing concerning it but that, under the same circumstances,

it has happened a certain number of times, and failed a certain other number of times. He adds, that he soon perceived that it would not be very difficult to do this, provided some rule could be found according to which we ought to estimate the chance that the probability for the happening of an event perfectly unknown, should lie between any two named degrees of probability, antecedently to any experiments made about it; Every judicious person will be sensible that the problem now mentioned is by no means merely a curious speculation in the doctrine of chances, but necessary to be solved in order to sure foundations for all our reasonings concerning past facts, and what is likely to be hereafter" ([1], S. 370-371).

Beachtenswert ist, daß bei dieser Lösung des Induktionsproblems der Begriff Wahrscheinlichkeit in zwei verschiedenen Situationen benutzt wird: zum einen zur Formulierung eines allgemeinen Gesetzes der Form: "Die Wahrscheinlichkeit, daß ein bestimmtes Ereignis eintritt, hat einen bestimmten Wert" und zum anderen, um den Grad der Gewißheit auszudrücken, mit der dieser Wert aus den Beobachtungen erschlossen werden kann. Bayes verwendet zur Unterscheidung dieser beiden Situationen im ersten Fall das Wort 'probability' und im zweiten das Wort 'chance' ("Required the chance that the probability of its happening in a single trial lies somewhere between any two degrees of probability that can be named." [1] S. 376), fügt aber hinzu: "By chance I mean the same as probability". In der oben gebrachten Formel wird in der ersten Situation die Wahrscheinlichkeit mit π symbolisiert, in der zweiten mit der a posteriori Wahrscheinlichkeitsdichte $p(\pi|x,n)$ (d.h. die Wahrscheinlichkeit für π -Werte im Intervall $(\pi, \pi+d\pi)$ ist bei kleinem $d\pi$ gleich $p(\pi|x,n)d\pi$).

Diese a posteriori Dichte präzisiert die Wahrscheinlichkeitsaussagen, die über das allgemeine Gesetz (Wahrscheinlichkeitsgesetz) aufgrund der Daten gemacht werden können. Nimmt man z.B. bei dem oben gebrachten Beispiel von 8 Heilungen bei 10 Patienten an, daß a priori über die Heilungswahrscheinlichkeit π nichts bekannt ist und somit jeder Wert zwischen 0 und 1 gleich wahrscheinlich ist (genauer: daß die Wahrscheinlichkeitsdichte für jeden Wert zwischen 0 und 1 gleich 1 ist), dann ist die a posteriori Verteilung zur Likelihood proportional und wird (bis auf einen Faktor) durch die Säulen der Abbildung 1 dargestellt. Man erkennt aus der Abbildung, daß die a posteriori Wahrscheinlichkeitsdichte für die Häufigkeit 8/10 am größten ist. Aber auch π -Werte zwischen 0.6 und 0.9 sind a posteriori noch plausibel. Aus der a posteriori Wahrscheinlichkeitsdichte können so Grenzen abgelesen werden, innerhalb derer der Grad der Gewißheit für die unbekannte Wahrscheinlichkeit π aufgrund der Daten noch als bedeutsam angesehen werden kann. Diese Aussage kann durch die Angabe eines 'Vertrauensintervalls' (credible interval) präzisiert werden. Dieses Intervall umfaßt den Bereich, in dem Werte der Heilungswahrscheinlichkeit π mit einer a posteriori Wahrscheinlichkeit von z.B. 95% zu erwarten sind (genauer: die a posteriori Wahrscheinlichkeit für π -Werte kleiner als die untere Grenze und größer als die obere Grenze des Vertrauensintervalls ist jeweils 2.5%). Bei 8 Heilungen in 10 Behandlungen und bei Annahme der nicht informativen Gleichverteilung als a priori Verteilung für π reicht das 95%-Vertrauensintervall von 0.48 bis 0.94. Die Breite dieses Intervalls nimmt mit zunehmender Beobachtungszahl ab und konzentriert sich auf die beobachtete Häufigkeit. Werden 50 Patienten behandelt und 40 davon geheilt, dann reicht das 95%-Vertrauensintervall von 0.60 bis 0.83; und bei 100 Behandlungen und 80 Heilungen von 0.67 bis 0.83. Mit zunehmender Beobachtungszahl kann somit immer zuverlässiger die beobachtete Häufigkeit als Wahrscheinlichkeit für das interessierende Ereignis angenommen werden. Dies ist der Inhalt vom 'Gesetz der großen Zahl', das sich bereits in J. Bernoullis Ars conjectandi findet. Aber

auch der Einfluß der a priori Wahrscheinlichkeit auf die a posteriori Verteilung nimmt mit zunehmender Beobachtungszahl ab. Hätten wir z.B. a priori angenommen, daß sich die Wahrscheinlichkeit π im Mittel um den Wert 0.25 verteilt (Beta-Verteilung mit den Parametern 2 und 6), dann geht bei 8 Heilungen unter 10 Behandlungen das a posteriori 95%-Vertrauensintervall von 0.33 bis 0.77, bei 40 Heilungen unter 50 Behandlungen von 0.60 bis 0.83, bei 80 Heilungen unter 100 Behandlungen von 0.67 bis 0.83 und bei 800 Heilungen unter 1000 Behandlungen von 0.77 bis 0.82.

2.2.2 Frequentistische Statistik

Gegen die Verwendung von a priori Wahrscheinlichkeiten haben sich in der ersten Hälfte dieses Jahrhunderts die englischen Statistiker um Karl Pearson und R.A. Fisher sowie R. von Mises gewandt. Sie sahen in der unbekanntem Wahrscheinlichkeit für ein Ereignis oder allgemein in den Kenngrößen (Parameter) von allgemeinen Wahrscheinlichkeitsgesetzen keine Zufallsgrößen, sondern unbekannte, feste Werte, für die aus den beobachteten Daten 'Schätzwerte' zu ermitteln oder über die 'Hypothesen' zu testen sind. Da die Likelihood die einzige Verbindung zwischen dem Wahrscheinlichkeitsgesetz und den Daten darstellt, sind auch alle empirischen Aussagen über die als feste Werte angenommenen Wahrscheinlichkeiten oder Parameter mit der Likelihood zu treffen. Es erscheint dabei als vernünftig, den Wert der Wahrscheinlichkeit oder der Parameter als 'Schätzwert' zu nehmen, für den bei den beobachteten Daten die Likelihood am größten ist. Dies ist das 'Maximum-Likelihood-Prinzip', das von Carl Friedrich Gauß (1777-1855) als 'Methode der kleinsten Quadrate' eingeführt, aber erst von R.A. Fisher voll entwickelt wurde (vor allem in seiner Arbeit: *On the mathematical foundations of theoretical statistics*, *Philosophical Transactions of the Royal Society*, vol. 222, 1921).

Ein Nachteil dieses Prinzips besteht darin, daß damit zwar ein Schätzwert hergeleitet, aber unmittelbar nichts über die Zuverlässigkeit dieses Schätzwertes ausgesagt werden kann. Um zu solchen Aussagen zu kommen, muß noch eine hypothetische Hilfskonstruktion eingeführt werden. Diese besteht darin, daß die beobachteten Daten (die 'Stichprobe') als Glied in einer unendlichen Folge von Wiederholungen des Experiments oder der Beobachtungen angesehen werden. Bei jeder Wiederholung werden zufällig andere Ergebnisse (Stichprobenwerte) und damit auch andere Schätzwerte erhalten. Im Rahmen dieser hypothetischen Folge kann somit den Schätzwerten eine Wahrscheinlichkeitsverteilung (als Verteilung der relativen Häufigkeiten) zugeordnet werden, die nur von der Zahl der Beobachtung (dem 'Stichprobenumfang') und vom zugrunde liegenden Wahrscheinlichkeitsgesetz und damit auch von den zu schätzenden Parametern abhängt. Bei nicht zu kleinen Stichprobenumfängen ('asymptotisch') stimmen (bei gewissen Regularitätsannahmen bezüglich der Likelihood) die Mittelwerte (Erwartungswerte) der Maximum-Likelihood-Schätzwerte mit den zu schätzenden Parametern überein (erwartungstreue Schätzwerte). Die Zuverlässigkeit des Schätzwertes wird mit der Standardabweichung des Schätzwertes, dem 'Standardfehler', oder noch besser mit einem 95%-'Konfidenzintervall' ausgedrückt. Dies ist ein Intervall, das aus den Daten der Stichprobe berechnet wird und in der Folge der unendlichen Wiederholungen in 95% der Fälle den unbekanntem Parameter enthält, in 5% ihn nicht umfaßt. Dieses Konfidenzintervall stimmt zwar bei nicht informativen a priori Verteilungen oft mit dem a posteriori Vertrauensintervall überein, hat aber eine andere Bedeutung. Während das Vertrauensintervall unmittelbar eine Wahrscheinlichkeitsaussage über den Parameter macht (die untere Grenze ist die 2,5%-Perzentile, die obere Grenze die 97,5%-Perzentile der a posteriori Verteilung des Parameters), macht das Konfidenzintervall

nur die qualitative Aussage: 'Der Parameter ist im Intervall enthalten'. Ob er in der Mitte oder mehr an einem der Ränder liegt, darüber wird nichts ausgesagt. Die Konfidenz von 95% bezieht sich nicht auf den Parameter, sondern auf die hypothetische unendliche Folge von Wiederholungen. In dieser Folge wird die Aussage, daß der Parameter im Konfidenzintervall liegt, in 95% der Fälle richtig und in 5% falsch sein. Die 'Konfidenz' von 95% charakterisiert somit primär die Zuverlässigkeit der Methode und nicht die des Schätzwertes. Über die Zuverlässigkeit des Schätzwertes kann nur durch die Breite des Intervalls eine indirekte Aussage abgeleitet werden. Je schmaler das Intervall ist, desto zuverlässiger wird der Schätzwert mit dem Parameter übereinstimmen. Dies kann aber nicht mit einer Wahrscheinlichkeitsangabe präzisiert werden.

2.2.3 Kritik der frequentistischen Statistik

Die Hilfskonstruktion der unendlichen Wiederholung ist nicht unproblematisch und macht die darauf beruhenden Methoden der 'frequentistischen' Statistik (Konfidenzintervalle, Signifikanztests) für den Anwender schwer verständlich. Selbst Egon S. Pearson, der als einer der 'Väter' des Konzepts der Konfidenzintervalle und Signifikanztests gilt, bekennt 1947 etwas resigniert: "The argument that 'if we were to repeatedly do so and so, such and such results would follow in the long run' is at once met by the commonsense answer that we never should carry out a precisely similar trial again... In ... numerous cases there is no repetition of the same type of trial or experiment, but all the same we can and many of us do use the same test rules to guide our decision, following the analysis of an isolated set of numerical data. Why do we do this? What are the springs of decision? Is it because the formulation of the case in terms of hypothetical repetition helps to that clarity of view needed for sound judgement? Or is it because we are content that the application of a rule, now in this investigation, now in that, should result in a long-run frequency of errors in judgement which we control at a low figure? On this I should not care to dogmatize, realizing how difficult it is to analyse the reasons governing even one's own personal decisions" ([12] S. 141/142). Die Methoden der Bayesianischen Statistik sind dem gegenüber unmittelbarer und flexibler einsetzbar. Sie erfordern nicht das Hilfsprinzip der unendlichen Wiederholung, dafür aber die Vorgabe einer a priori Verteilung. Dies ist kein Nachteil, wie man häufig noch glaubt, sondern erweist sich im Gegenteil als Vorteil, da damit Vorwissen eingebracht und individuelle Variationen bei den Wahrscheinlichkeitsgesetzen angemessen berücksichtigt werden können. Die frequentistische Statistik kann als ein singulärer Spezialfall der Bayesianischen Statistik angesehen werden, bei der die singuläre a priori Verteilung angenommen wird, die einem unbekanntem Wert des Parameters die Wahrscheinlichkeit 1 und allen anderen Werten die Wahrscheinlichkeit 0 zuordnet. Mit dieser singulären a priori Verteilung kann aber auch nur eine singuläre a posteriori Verteilung erhalten werden, die nicht weiterhilft. Deshalb mußte in der frequentistischen Statistik die Hilfskonstruktion mit den unendlichen Wiederholungen, auf die sich die Konfidenzwahrscheinlichkeit oder bei statistischen Tests die Signifikanzwahrscheinlichkeit bzw. Irrtumswahrscheinlichkeit bezieht, eingeführt werden. In der Bayesianischen Statistik ist diese Hilfskonstruktion nicht nötig. In den letzten Jahren hat sich vor allem in England eine 'Kuhnsche Revolution', ein Paradigmenwechsel vollzogen (vergl. [9]), bei dem sich die wissenschaftliche Gemeinschaft der Statistiker vermehrt den Bayesianischen Konzepten zuwendet. Wegen der Möglichkeit, die Individualität der Patienten zu berücksichtigen und Vorwissen über die Verfahren einzubringen, dürften diese Methoden nicht zuletzt auch für die alternativen Heilverfahren und für die Auswertung von Beobachtungsstudien interessant sein.

3. Wirksamkeitsnachweis

Ein Patient sieht die Behandlung seiner Krankheit als wirksam an, wenn seine Beschwerden nach der Behandlung verschwunden oder zumindest gemildert sind. Er kann allerdings nicht beurteilen, ob diese Heilung oder Linderung auf die Behandlung zurückzuführen ist oder auch ohne die Behandlung zustande gekommen wäre, vielleicht sogar rascher oder besser als mit der Behandlung. Von der Wirksamkeit einer Behandlungsmethode kann man nur sprechen, wenn mit dieser Methode mehr Patienten eine Heilung ihrer Krankheit oder Linderung ihrer Beschwerden erfahren als ohne die Methode. Der Wirksamkeitsnachweis verlangt somit einen induktiven Wahrscheinlichkeitsvergleich mit den Ergebnissen, die ohne die Methode (oder mit einer anderen Methode) zu erwarten sind. Dieser Vergleich basiert auf Beobachtungen der Therapieergebnisse, die sowohl mit als auch ohne die zu prüfende Behandlungsmethode (Prüftherapie) gewonnen wurden. Die Prinzipien der induktiven Schlußweisen, die bei der Bewertung der Ergebnisse anzuwenden sind, wurden im vorhergehende Abschnitt ausführlich besprochen. Das Besondere des Wirksamkeitsnachweises besteht darin, daß nicht nur ein verallgemeinernder induktiver Schluß, sondern ein Vergleich gefordert wird. Dies setzt aber voraus, daß die Patienten, die mit der Prüftherapie behandelt wurden, in ihren Ausgangsbedingungen und Zusatzbehandlungen mit denen vergleichbar sind, die mit der Vergleichstherapie behandelt wurden (die auch im Verzicht auf eine spezielle Therapie bestehen kann). In Bayesianischer Sicht bedeutet dies, daß die Ausgangslagen der Behandlungen in beiden Gruppen a priori dieselben Wahrscheinlichkeitsverteilungen haben; in frequentistischer Sicht sollen die Patienten beider Gruppen als unabhängige Zufallsstichproben aus derselben Grundgesamtheit angesehen werden können. Dies kann im Prinzip auf zwei verschiedene Arten erreicht werden: durch prospektive kontrollierte klinische Studien mit randomisierter Zuteilung der Behandlungen zu den Patienten und durch Anwendungsbeobachtungen mit methodischem Ausgleich der unterschiedlichen Ausgangsbedingungen.

3.1 Klinische Studien

Die bevorzugte Methode des Wirksamkeitsnachweises ist die der **kontrollierten klinischen Studie**. Dabei wird aus den für die Untersuchung in Frage kommenden Patienten nach vorgegebenen Ein- und Ausschlußkriterien eine Zufallsstichprobe ausgewählt und die zu vergleichenden Behandlungsmethoden diesen Patienten zufällig zugeteilt, so daß für jeden Patienten dieselbe Chance besteht, die Prüftherapie oder Vergleichstherapie zu erhalten. Um auch bei beiden Gruppen eine vergleichbare Erwartungshaltung bezüglich des Therapieausgangs zu haben, sollte die Vergleichsgruppe nicht ohne spezielle Therapie bleiben, sondern ein sogenanntes Placebo (d.h. eine Behandlung mit einem Mittel ohne bekannten Wirkstoff) erhalten. Wenn dies sachlich und ethisch vertretbar ist, sollten die beiden Therapieformen in der Handhabung, dem Aussehen und Geschmack gleich sein und weder dem Patienten noch dem Arzt bekannt sein, welches Mittel im Einzelfall angewendet wird. Solche Studien nennt man **Doppelblindstudien**. Die Studiendurchführung muß den ethischen Grundsätzen der 'Deklaration von Helsinki' genügen. Das bedeutet, daß die an der Studie beteiligten Ärzte die Studie für ethisch vertretbar halten (d.h. zu Beginn der Studie davon ausgehen können, daß beide Behandlungsmethoden ohne spezielle Nachteile für die Patienten angewendet werden können), die eingeschlossenen Patienten (nach Möglichkeit) vorher ausführlich aufgeklärt werden, freiwillig ihre Zustimmung zur Teilnahme geben und jederzeit ohne Nachteile von der Teilnahme zu-

rücktreten können. Vor Beginn der Studie ist ein Prüfplan zu erstellen, der der zuständigen Ethikkommission vorzulegen und von dieser zu befürworten ist. Richtlinien zur Erstellung von Prüfplänen, Durchführung und Auswertung der Studien sowie zur Erstellung von Berichten über die Ergebnisse wurden (und werden weiterhin) von der 'European Agency for the Evaluation of Medicinal Products (EMA)' erstellt und sind unter http://www.eudra.org/en_home.htm im Internet abzurufen.

Anwendungsbeobachtungen sind dadurch charakterisiert, daß das individuelle Arzt-Patienten-Verhältnis in Bezug auf Indikationsstellung sowie Wahl und Durchführung der Therapie nicht beeinflußt wird. Insbesondere erfolgt die Zuteilung der Behandlungen nicht randomisiert. Die Studie kann retrospektiv (Auswertung von Krankenakten) oder prospektiv (Dokumentation der Ausgangsdaten, Behandlungen und des Krankheitsverlaufs bei neu zur Behandlung kommenden Patienten) angelegt sein. Nach den 'Empfehlungen zur Planung und Durchführung von Anwendungsbeobachtungen' des BfArM (Bundesinstitut für Arzneimittel und Medizinprodukte) ist "ein Nachweis der Wirksamkeit allein durch AWB (Anwendungsbeobachtungen) ... bis auf besonders begründete Ausnahmefälle nicht möglich". Immerhin wird auch in diesen Empfehlungen zugestanden, daß mit Anwendungsbeobachtungen die "Erkenntnisse zur Wirksamkeit" erweitert werden können. Voraussetzung ist allerdings, daß vor der Durchführung der Beobachtung ein Plan erstellt wurde, in dem die wissenschaftliche Zielsetzung als präzise Fragestellung formuliert ist sowie das gewählte Design (Basis eines Vergleichs, zeitlicher Umfang und Untersuchungsumfang beim einzelnen Patienten, Patientenzahl) und die geplanten Methoden der Datenerhebung und Auswertung festgelegt sind. Um mit Anwendungsbeobachtungen einen Therapievergleich durchführen zu können, müssen Daten von Patienten zur Verfügung stehen, die mit Vergleichstherapien behandelt wurden. Am geeignetsten dafür sind **Kohortenstudien**. Bei diesen Studien wird (möglichst in verschiedenen Einrichtungen) eine 'Kohorte' von Patienten mit der interessierenden Erkrankung ausgewählt und es werden der Ausgangszustand der Patienten, die angewendeten Behandlungen und die damit erzielten Ergebnisse beobachtet und dokumentiert. Werden nur Patienten beobachtet, die mit der Prüftherapie behandelt wurden, dann muß außerhalb der Studie eine Vergleichsbasis geschaffen werden. Diese kann in einem 'historischen Vergleich' mit Patienten bestehen, die früher mit anderen Therapien behandelt wurden und deren Therapieergebnisse dokumentiert oder publiziert sind. Ein solcher Vergleich ist natürlich sehr problematisch, da früher die gesamten Behandlungsbedingungen und auch der Krankheitsverlauf anders gewesen sein können. Aber auch bei Kohortenstudien muß man davon ausgehen, daß sich die mit den Vergleichstherapien behandelten Patienten in ihrer Ausgangssituation deutlich von den mit der Prüftherapie behandelten unterscheiden. Es muß daher auf jeden Fall versucht werden, den Einfluß dieser unterschiedlichen Ausgangslagen auf das Behandlungsergebnis auszugleichen. Dies kann dadurch geschehen, daß man 'Subgruppen' von Patienten bildet, die sich in ihrer Ausgangslage (z.B. Alter, Vorerkrankungen, Begleiterkrankungen, Ausgangsbefund) möglichst ähnlich sind, und die verschiedenen Therapieformen jeweils innerhalb einer Subgruppe vergleicht. Wenn zwar die Therapieergebnisse, aber die Unterschiede in den Therapieergebnissen zwischen den zu vergleichenden Therapien bei verschiedenen Subgruppen nicht wesentlich verschieden sind (keine Wechselwirkung zwischen Therapie und Subgruppe besteht), dann können die Therapieeffekte der Subgruppen zu einem gemeinsamen Schätzwert zusammengefaßt bzw. die a posteriori Verteilung des gemeinsamen Therapieeffekts ermittelt werden. Wenn allerdings bedeutsame Unterschiede bestehen, dann bleibt nur übrig, die einzelnen Subgruppen getrennt zu be-

werten. Man erhält so Informationen, für welche Subgruppen mit der Prüftherapie günstige und für welche ungünstige Ergebnisse zu erwarten sind. Eine extreme Form der Subgruppen-Analyse liefert die 'matched-pairs' Technik. Dabei wird zu jedem mit der Prüftherapie behandelten Patienten ein mit der Vergleichstherapie behandelter ausgesucht, der diesem in der Ausgangslage möglichst gleicht. Der Wirksamkeitsvergleich wird jeweils innerhalb eines solchen Paares durchgeführt und die Vergleiche innerhalb der Paare werden (bei fehlender Wechselwirkung) über die Paare zu einem Gesamtergebnis zusammengefaßt.

Eine Modifikation der kontrollierten Studien bilden die **cross-over-Studien**, bei denen jeder Patient sowohl die Prüf- als auch die Vergleichstherapie bei unterschiedlichen Therapiephasen in randomisierter Reihenfolge erhält. Diese Studienart ist nur dann sinnvoll anzuwenden, wenn sich nach Abschluß der ersten Therapiephase und einer evtl. anschließenden 'wash-out-Phase' ohne Therapie für die zweite Therapiephase dieselbe Verteilung des Ausgangszustandes einstellt wie vor der ersten Therapiephase. Dies schränkt die Anwendbarkeit von cross-over-Studien für den Wirksamkeitsnachweis erheblich ein. Bei akuten Erkrankungen oder bei Erkrankungen, bei denen der Erkrankungszustand durch die Therapie dauerhaft gebessert werden kann, sind diese Studien nicht anzuwenden.

Eine spezielle Studienart bilden die **Einstichproben-Studien** (one-sample-design), bei denen die Patienten nur mit der Prüftherapie behandelt werden. Bei diesen Studien kann eine induktive Aussage über die Wirksamkeit gemacht werden, wenn für den relevanten Parameter der Zielgröße (d.h. der Größe, mit der der Therapieerfolg bewertet wird) ein **Sollwert** vorgegeben wird. Die Behandlung wird als wirksam angesehen, wenn der Parameter den Sollwert übertrifft, sonst als unwirksam. Ist z.B. Zielgröße die Heilung der Krankheit, deren Verteilung durch die Heilungswahrscheinlichkeit π vollständig bestimmt ist, dann ist für diese Wahrscheinlichkeit ein Sollwert π_0 vorzugeben. Kann mit vorgegebener Zuverlässigkeit aus den Studienergebnissen induktiv geschlossen werden, daß die Heilungswahrscheinlichkeit π größer als π_0 ist, dann wird die Wirksamkeit angenommen, sonst abgelehnt. Problematisch ist die Festlegung eines anerkannten Sollwertes. Der Wert müßte jedenfalls größer sein, als ohne die Behandlung zu erwarten ist (z.B. größer als die zu erwartende Rate von Spontanheilungen). Woher kann man Informationen über einen solchen Sollwert erhalten? Dies ist letztlich doch nur aus 'historischen' Daten möglich und wir sind wieder mit der Problematik des historischen Vergleichs konfrontiert. Deshalb werden bei der Zulassung von Arzneimitteln die Ergebnisse von Einstichproben-Studien im allgemeinen nicht anerkannt. Diese Studien spielen aber bei der Entwicklung und Testung von Arzneimitteln in der sogenannten Phase II eine große Rolle, in der erstmals an Kranken überprüft wird, ob in der vorgesehenen Dosierung überhaupt ein nennenswerter Therapieerfolg zu erwarten ist. Soll z.B. eine neues Arzneimittel bei einer Krebserkrankung eingesetzt werden, bei der bisher höchstens 15% der Patienten auf eine Therapie angesprochen haben, dann kann eine Ansprechrage von 20% als Erfolg gewertet und als Sollwert vorgegeben werden. Zeigen die Studienergebnisse, daß mit hinreichender Zuverlässigkeit die Wahrscheinlichkeit des Ansprechens 20% oder größer ist, dann wird das Mittel weiter entwickelt, andernfalls wird die Entwicklung gestoppt.

3.2 Wirksamkeitsvergleich

Das Ziel der Studienauswertung besteht darin, aufgrund der beobachteten Ergebnisse induktiv die bei zukünftigen Anwendungen der Prüftherapie zu erwartenden Wir-

kungen mit denen der Vergleichstherapie zu vergleichen. Voraussetzung für eine 'konfirmatorische' Vergleichsaussage ist, daß sowohl bei kontrollierten Studien als auch bei Anwendungsbeobachtungen im Prüfplan eindeutige therapeutische Zielgrößen festgelegt sind und im Auswertungsplan angegeben ist, welche Parameter nach welchen statistischen Verfahren zur Bewertung der therapeutischen Wirksamkeit beurteilt werden. Die Wirksamkeit gilt als nachgewiesen, wenn mit den entsprechenden induktiven (statistischen) Verfahren mit der vorgegebenen Zuverlässigkeit aufgezeigt wird, daß aufgrund der beobachteten Ergebnisse die Verteilung der Zielgrößen sich unter der Prüftherapie von der Verteilung unter der Vergleichstherapie (Placebo) so unterscheidet, daß für die Mehrzahl der künftig zu therapierenden Patienten mit der Prüftherapie ein größerer und klinisch valider Vorteil zu erwarten ist.

3.2.1 Frequentistischer Signifikanztest mit fixem Stichprobenumfang

Im Rahmen der frequentistischen Statistik wird angenommen, daß der relevante Parameter der Zielgrößen unter der Prüf- und Vergleichstherapie jeweils einen festen aber unbekanntem Wert hat. Kriterium für die Wirksamkeit ist der Unterschied zwischen diesen beiden Werten, z.B. die Differenz oder eine sonstige angemessene Funktion. Bei der Zielgröße 'Heilung' ist der relevante Parameter die Heilungswahrscheinlichkeit π und der Unterschied zwischen beiden Behandlungen entspricht der Differenz $\delta = \pi_{\text{Prüf}} - \pi_{\text{Vergl}}$. Bei quantitativen Zielgrößen ist ein relevanter Parameter der Mittelwert (Erwartungswert) μ der Verteilung und der Unterschied zwischen den Therapien wird durch die Differenz $\mu_{\text{Prüf}} - \mu_{\text{Vergl}}$ ausgedrückt. Bei den sogenannten 'parameterfreien' Methoden werden für die Verteilungen der Zielgrößen unter beiden Therapien keine parametrischen Modelle angenommen; der Unterschied wird durch den 'Mann-Whitney-Parameter' $P(X_{\text{Prüf}} > X_{\text{Vergl}})$ ausgedrückt, d.h. durch die Wahrscheinlichkeit, bei einem beliebig herausgegriffenen Patienten, der mit der Prüftherapie behandelt wurde, einen größeren Wert der Zielgröße zu erhalten als bei einem mit der Vergleichstherapie behandelten Patienten.

Die möglichen Werte des Unterschieds in den Zielparametern werden in zwei Bereiche unterteilt: den Bereich der **Nullhypothese**, für den die therapeutische Wirksamkeit nicht gegeben ist (z.B. für Unterschiede kleiner oder gleich 0) und den Bereich der **Alternativhypothese**, für den eine therapeutische Wirksamkeit angenommen wird (z.B. für Unterschiede > 0). Aus den für beide Therapien beobachteten Daten x der Zielgröße wird eine Teststatistik $t(x)$ berechnet, deren Wert um so größer zu erwarten ist, je größer der Parameterunterschied zwischen den beiden Therapien ist. Um mit dieser Teststatistik zu einer Entscheidung über Annahme oder Ablehnung der Nullhypothese zu kommen, wird auf die bereits erwähnte Hilfskonstruktion der unendlichen Folge von hypothetischen Wiederholungen der Studie zurückgegriffen. In dieser Folge haben die möglichen Werte der Teststatistik eine Wahrscheinlichkeitsverteilung, die von den Verteilungen in beiden Gruppen und damit auch vom Unterschied in den Parametern sowie von der Zahl der Beobachtungen (den Stichprobenumfängen beider Gruppen) abhängt. Man legt willkürlich fest, daß die Nullhypothese nur dann verworfen und die Alternativhypothese (und damit die Wirksamkeit) angenommen wird, wenn in dieser Folge bei Gültigkeit der Nullhypothese die Wahrscheinlichkeit für den beobachteten Wert $t(x)$ der Teststatistik oder einen noch größeren höchstens gleich einem vorgegebenen kleinen Wert α (für den 'aus Aberglauben und Gewohnheit' der Wert 0.05 vorgegeben wird) ist. Das 'Niveau' α ist damit die maximale 'Irrtumswahrscheinlichkeit 1. Art', die Nullhypothese zu verwerfen, obwohl sie gilt. Der Signifikanztest läuft also darauf hinaus, daß zu gegebenem α (0.05) für die Teststatistik $t(x)$ die Signifikanzschwelle t_α berechnet wird, so daß in der Folge der Wiederholungen Werte der Teststatistik $\geq t_\alpha$ höchstens mit Wahrscheinlichkeit α

zu erwarten sind. Für $t(x) \geq t_\alpha$ wird die Nullhypothese verworfen, sonst angenommen. Werte der Signifikanzschwellen sind für verschiedene Teststatistiken, Stichprobenumfänge und α -Werte in statistischen Tafeln tabelliert oder können mit entsprechenden Programmen berechnet werden.

Wird mit den Daten einer Studie die Nullhypothese verworfen und somit die Wirksamkeit angenommen, dann genügt ein solcher 'reiner' Signifikanztest. Kann die Nullhypothese nicht verworfen werden, dann ist das Vorgehen unbefriedigend. Man kann dann nämlich nicht ohne weiteres annehmen, daß die Prüftherapie nicht besser als die Vergleichstherapie ist; es kann ein Fehler 2. Art vorliegen, die Nullhypothese anzunehmen, obwohl sie falsch ist. Die Wahrscheinlichkeit für diesen Fehler, die Irrtumswahrscheinlichkeit 2. Art, hängt (bei gegebenem α) vom tatsächlichen Unterschied δ und dem Umfang beider Stichproben ab. Für einen vorgegebenen Unterschied δ_1 kann dieser Fehler (bei vorgegebenem Niveau α) durch die Wahl des Stichprobenumfangs kontrolliert werden, indem für die Irrtumswahrscheinlichkeit 2. Art eine obere Schranke β (meist 0.2) vorgegeben wird. Die Stichprobenumfänge sind so zu wählen, daß diese Schranke für $\delta=\delta_1$ eingehalten wird. Statt der Irrtumswahrscheinlichkeit 2. Art wird häufig ihr Komplement, die Power (Teststärke, Testschärfe) $1-\beta$ vorgegeben; d.i. die Wahrscheinlichkeit für eine 'signifikantes' Ergebnis bei einem Unterschied δ_1 .

Das Verfahren des Signifikanztests soll an einem einfachen Beispiel demonstriert werden. Es wird eine quantitative Zielgröße x (z.B. die Blutdrucksenkung bei Hypertonikern oder Änderung des Hamilton-Scores bei Depression) und als relevanter Parameter der Mittelwert angenommen, der für die Prüftherapie mit μ_1 und für die Vergleichstherapie mit μ_2 bezeichnet wird. Um das Beispiel nicht unnötig zu komplizieren, wird angenommen, daß die Werte der Zielgröße x normal verteilt sind mit den Mittelwerten μ_1 bzw. μ_2 und der Varianz $\frac{1}{2}$ in beiden Gruppen. Als Parameter für den Unterschied zwischen beiden Therapien wird die Differenz $\delta=\mu_1-\mu_2$ genommen. In einer klinischen Studie werden in randomisierter Zuteilung jeweils n Patienten mit der Prüf- und Vergleichstherapie behandelt und die Werte x_{1i} und x_{2i} beobachtet. Die Nullhypothese umfaßt die Werte $\delta \leq 0$, die Alternativhypothese die Werte $\delta > 0$. Teststatistik ist die Differenz der beiden Mittelwerte $t = \bar{x}_1 - \bar{x}_2$, die in der Folge der unendlichen Wiederholungen normal verteilt ist mit dem Mittelwert δ und der Varianz $1/\sqrt{n}$. Die Studie kann auch so durchgeführt werden, daß n Paare von Patienten gebildet werden und bei jedem Paar einem Patienten die Prüftherapie, dem anderen die Vergleichstherapie randomisiert zugeteilt wird (Blockrandomisation mit Blocklänge 2). Tabelle 1 zeigt zu verschiedenen n die Signifikanzschwelle t_α und die Power für $\delta_1=1$.

Umfang pro Gruppe	Signifikanzschwelle t_α	Power
1	1.645	0.260
2	1.163	0.409
3	0.950	0.535
4	0.822	0.639
5	0.736	0.723
6	0.672	0.789
7	0.622	0.842
8	0.582	0.882
9	0.548	0.912
10	0.520	0.935

Tabelle 1: Signifikanzschwelle und Power bei verschiedenen Stichprobenumfängen

Bei einem Stichprobenumfang von 7 Patienten pro Gruppe (7 Paaren) ist die Power 0.842 und damit erstmals größer als 0.8. Der Mittelwert in der Prüfgruppe muß mindestens um den Betrag 0.622 größer sein als der in der Vergleichsgruppe, um die Nullhypothese abzulehnen.

3.2.2 Frequentistische Sequentialtests

Dieses übliche Verfahren, den Stichprobenumfang durch die Vorgabe einer relevanten Differenz δ_1 und der entsprechenden Power festzulegen und erst nach Abschluß der Studie über die Wirksamkeit der Prüftherapie zu entscheiden, ist besonders bei Langzeitstudien ethisch bedenklich. Es könnte sein, daß die Prüftherapie deutlich besser ist, als ursprünglich angenommen wurde. Man könnte dann bereits bei weniger Patienten als vorgesehen diesen Vorteil erkennen und würde nicht unnötig den Patienten der Vergleichsgruppe die schlechtere Therapie zumuten. Nach der Deklaration von Helsinki muß sich der Prüfarzt auch im Verlauf der Studie vergewissern, ob die Prüf- und Vergleichstherapie noch als gleich indiziert anzusehen sind. Dies bedingt, in hinreichend engen Intervallen Zwischenauswertungen vorzunehmen. Man nennt solche Auswertungsverfahren **Sequentialverfahren**.

Dabei ergibt sich aber bei dem frequentistischen Konzept das Problem der multiplen Testung. Wird bei einem geringen Stichprobenumfang eine Auswertung vorgenommen und zum Niveau α über die Ablehnung der Nullhypothese entschieden, dann kann bei einer Fortsetzung der Studie niemals wieder auf demselben Niveau α die Nullhypothese abgelehnt werden. Es bestand ja schon bei der ersten Auswertung die Wahrscheinlichkeit α , eine gültige Nullhypothese irrtümlich abzulehnen. Legt man auch bei der zweiten Auswertung dieses Niveau zugrunde, dann ist die Wahrscheinlichkeit, bei einer dieser beiden Auswertungen die Nullhypothese irrtümlich abzulehnen, größer als α ; sie kann maximal 2α betragen (Bonferroni-Ungleichung). Um also insgesamt bei den verschiedenen Zwischenauswertungen das Niveau α einzuhalten, muß bei den einzelnen Auswertungen zu einem geringeren Niveau entschieden werden; das vorgegebene Niveau α muß auf die einzelnen Auswertung so 'verteilt' werden, daß die multiple Wahrscheinlichkeit, bei irgendeiner Auswertung die Nullhypothese irrtümlich abzulehnen, auf α beschränkt bleibt (α -spending function). Entsprechendes gilt für die Irrtumswahrscheinlichkeit β , die Nullhypothese anzunehmen, obwohl der Unterschied δ_1 beträgt. Die Verteilung der α - und β -Werte auf die einzelnen Auswertungsschritte kann nach verschiedenen Funktionen geschehen. Pocock [13] hat für alle Auswertungsschritte dasselbe nominelle Niveau α^* (das natürlich kleiner als α sein muß) vorgeschlagen; O'Brien und Fleming [11] haben für die anfänglichen Zwischenauswertungen sehr kleine α^* -Werte, für die späteren größere vorgeschlagen, um eine zu frühe Beendigung der Studie zu vermeiden. Die optimalen Prozeduren (mit geringster zu erwartender Patientenzahl) liegen zwischen diesen beiden Ansätzen.

Ein weiterer Unterschied besteht zwischen **Gruppen-sequentiellen** Verfahren und **kontinuierlichen** Sequentialverfahren. Im ersten Fall werden die Anzahl der möglichen Zwischenauswertungen und die Stichprobenumfänge der Auswertungsgruppen vor Beginn der Studie festgelegt. Es werden dann für diese Stufen die nominellen α^* - und β^* -Werte bzw. die entsprechenden Schwellenwerte der Teststatistik (die mit allen bei der jeweiligen Stufe vorliegenden Daten gebildet wird) berechnet, die erreicht werden müssen, um die Studie abzuschließen und entweder die Nullhypothese oder

die Alternativhypothese anzunehmen. Erreicht bei einer Stufe der Pfad der Teststatistik (d.h. der Wert der Teststatistik als Funktion der bis dahin erfaßten Patientenzahl) keine dieser beiden Grenzen, dann wird die Studie mit der nächsten Stufe (Gruppe) fortgesetzt. Für die letzte Stufe ist stets eine Entscheidung vorgegeben. Bei den kontinuierlichen Verfahren werden Zeitpunkt und Stichprobenumfang für die einzelnen Zwischenauswertungen nicht vorgegeben. Dies macht es erforderlich, für jeden möglichen Pfad der Teststatistik Annahmegrenzen für die Null- und Alternativhypothese vorzugeben (deshalb die Bezeichnung 'kontinuierlich'). Verläuft der Pfad der Teststatistik innerhalb dieser Grenzen (im Fortsetzungsbereich), wird die Studie fortgesetzt; erreicht oder überschreitet er eine der Grenzen, wird die Studie abgeschlossen und die entsprechende Entscheidung getroffen. Dies bedeutet aber nicht (entgegen der irrigen Auffassung mancher Statistiker), daß auch für jeden Stichprobenumfang (d.h. immer wenn die Daten eines neuen Patienten vorliegen) ausgewertet werden muß. Vielmehr kann auch bei den kontinuierlichen Sequentialverfahren die Auswertung in Gruppen erfolgen, wobei aber Zeitpunkt und Gruppengröße der Zwischenauswertungen nicht vorher festgelegt werden müssen, sondern im Verlauf der Studie ad hoc bestimmt werden. Dies ist ein wesentlicher Vorteil der kontinuierlichen gegenüber den Gruppen-sequentiellen Verfahren. Die Zuverlässigkeit der Testentscheidung (d.h. die vorgegebenen Irrtumswahrscheinlichkeiten) wird durch die ad hoc Wahl der Zwischenauswertungen nicht beeinflusst. Es kann höchstens vorkommen, daß man bei einer Zwischenauswertung nach Vorliegen der Daten einer größeren Gruppe feststellt, daß die Studie bereits eher hätte beendet werden können, da der Pfad der Teststatistik bereits im Verlauf der Gruppe eine der Entscheidungsgrenzen erreicht hat. Dieses 'overshooting' kann vermieden werden, wenn zu Beginn der Studie die Zwischenauswertungen mit größeren Gruppengrößen, bei Annäherung an eine der Entscheidungsgrenzen mit kleinen Gruppengrößen durchgeführt werden.

Das erste kontinuierliche Sequentialverfahren wurde am Ende des zweiten Weltkrieges im Rahmen der RAND-Corporation von Abraham Wald entwickelt und 1947 veröffentlicht [18]. Als Teststatistik verwendet Wald die Likelihood-Ratio-Statistik; d.h. der Quotient aus der Likelihood für die Alternativhypothese zu der für die Nullhypothese. Das Verfahren heißt deshalb auch 'Sequential Probability Ratio Test'. Der Fortsetzungsbereich wird durch zwei Geraden parallel zur n-Achse begrenzt. Das Verfahren ist bei Gültigkeit der Null- oder Alternativhypothese optimal; d.h. es kommt bei kleinster mittlerer Stichprobengröße zu einer Entscheidung (minimale 'Average Sample Number' ASN). Es hat aber den Nachteil, daß sich eine maximale Stichprobengröße, bei der sicher eine Entscheidung getroffen wird, nicht angeben läßt. Man nennt solche Verfahren 'offene' Sequentialverfahren. Ein kontinuierliches 'geschlossenes' Verfahren wurde 1983 von Whitehead veröffentlicht [19]. Im oben gebrachten Beispiel eines Signifikanztests für den mittleren Unterschied zwischen Meßwertpaaren, deren Differenz normal verteilt ist mit dem Mittelwert δ und der Varianz 1, ist sequentielle Teststatistik die Summe der Differenzen der bis zur Zwischenauswertung beobachteten Meßwertpaare, die als Sequenzpfad über die Zahl der Paare aufgetragen wird. Ein simulierter Sequenzpfad für paarweise Differenzen, die um den Mittelwert 1 mit der Varianz 1 normal verteilt sind, ist zusammen mit dem Fortsetzungsbereich in Abbildung 2 zu $\alpha=0.05$ und $\beta=0.2$ bei $\delta_1=1$ angegeben. Der Fortsetzungsbereich wird durch zwei Geraden begrenzt, die ein Dreieck bilden. Solange sich der Pfad in diesem Dreieck bewegt, wird die Studie fortgesetzt. Sobald eine der Grenzen erreicht oder überschritten wird, wird die Studie beendet. Es wird die Alternativhypothese H_1 angenommen, wenn die obere Grenze erreicht wird, und die Nullhypothese H_0 beim Erreichen der unteren Grenze.

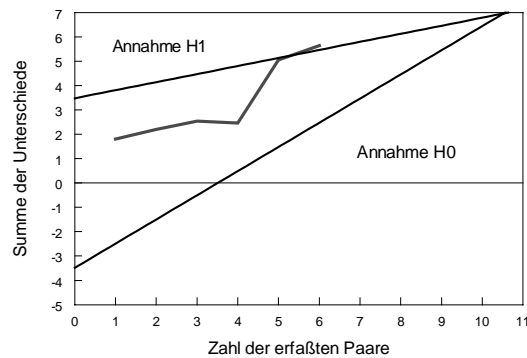


Abbildung 2: Sequentieller Dreieckstest

Im Beispiel der Abb. 2 hat der Sequenzpfad bei $n=6$ Paaren die obere Grenze erstmals überschritten. Die Studie wird somit bei insgesamt 12 Patienten gestoppt und die Alternativhypothese H_1 , daß die Prüftherapie der Vergleichstherapie überlegen ist, angenommen. Wie aus Tabelle 1 hervorgeht, hätte man in derselben Situation mit einem fixen Test 7 Paare (also insgesamt 14 Patienten) benötigt. Allerdings kann beim sequentiellen Dreiecksverfahren die maximale Zahl der Paare bis zu einer Entscheidung 11 (22 Patienten) betragen. Es ist aber äußerst unwahrscheinlich, daß diese maximale Zahl tatsächlich erreicht wird. Im Mittel kann bei diesem Beispiel mit dem Dreiecksverfahren nach etwas weniger als 5 Paaren die Studie beendet werden. Die Wahrscheinlichkeit, daß 7 oder mehr Paare benötigt werden, ist nur etwa 10%. Die Sequentialverfahren bringen also gegenüber dem fixen Testverfahren erhebliche Vorteile.

3.2.3 Kritik des frequentistischen Konzepts

Aus Tabelle 1 geht hervor, daß nach dem frequentistischen Konzept bei Vorgabe eines Stichprobenumfangs von 1 Paar die Nullhypothese $\delta \leq 0$ zu Gunsten der Alternativhypothese $\delta > 0$ auf dem Niveau $\alpha=0.05$ abgelehnt wird, wenn die beobachtete Differenz größer als 1.645 ist. Wäre die Studie sequentiell nach dem Dreiecksverfahren geplant, dann müßte beim ersten Paar die Differenz größer als 3.811 sein, um zur Beendigung und Ablehnung der Nullhypothese zu führen. Der Grund liegt darin, daß beim sequentiellen Verfahren die Studie fortgesetzt und damit die Entscheidung wiederholt werden kann. Deshalb müssen die vorgegebenen Irrtumswahrscheinlichkeiten über den gesamten möglichen Verlauf verteilt werden. Die Entscheidung über Annahme oder Ablehnung der Nullhypothese hängt also beim frequentistischen Konzept nicht nur von den beobachteten Daten, sondern auch noch von der vorgesehenen Auswertungsmethode, d.h. der vorgesehenen 'stopping rule', ab. Dies ist eine Merkwürdigkeit des frequentistischen Konzepts, die von verschiedenen Autoren heftig kritisiert wurde. Als Beispiel sei Jerome Cornfield vom National Institute of Health der USA zitiert, der bemerkte: "To most scientists without previous exposure to statistics, as well as to most intelligent laymen, any dependence of conclusions on stopping rules ... seems like a violation of common sense. Those biostatisticians who defend sequential analysis on the other hand would argue that dependence of conclusions on stopping rules is required to preserve the critical level, i.e., the lowest significance level at which the hypothesis can be rejected for given data. If one accepts the importance of preserving the critical level, then clearly conclusions must depend on the stopping rule. But what is not immediately obvious is that the critical level pro-

vides an appropriate measure of the amount of evidence in the data for or against the hypothesis" ([3] S. 18]. Diese 'Evidenz der Daten für oder gegen Hypothesen' kann mit dem Bayesianischen Konzept viel besser erbracht werden als mit dem frequentistischen Konzept des Signifikanztests. Im folgenden soll dieses Konzept am oben diskutierten Beispiel des Wirksamkeitsvergleichs zweier Behandlungen erläutert werden.

3.2.4 Bayesianisches Konzept des Wirksamkeitsvergleichs

Wie bereits im Abschnitt 2.2.1 dargelegt wurde, beruht das Bayesianische Konzept der induktiven Aussagen auf der Annahme, daß alle unbekanntes Größen (Parameter der Wahrscheinlichkeitsverteilung oder sonstige unbekanntes Größen) Zufallsgrößen sind, denen a priori (d.h. vor den Beobachtungen) eine Verteilung zugeordnet wird. Die beobachteten Daten liefern zusammen mit der a priori Verteilung die a posteriori Verteilung der unbekanntes Größen, wobei als Bindeglied die Likelihood fungiert, d.h. die Verteilung, die den beobachteten Daten zukommt, wenn für die unbekanntes Größen bestimmte Werte angenommen werden. Alle induktiven Aussagen über die unbekanntes Größen werden von ihrer a posteriori Verteilung hergeleitet.

Im oben diskutierten Beispiel des Wirksamkeitsvergleichs zweier Therapien, mit denen jeweils ein Paar von Patienten behandelt wird, sind die beobachteten Daten die Differenzen d_i der Zielgröße zwischen dem mit der Prüftherapie und dem mit der Vergleichstherapie behandelten Patienten. Es wird angenommen, daß diese Daten unabhängig normal verteilt sind mit dem Mittelwert δ und der Varianz $\sigma^2=1$. Diese Verteilung ergibt die Likelihood für die beobachteten Daten als Funktion des Parameters δ . Vor der Analyse muß für δ eine a priori Verteilung angenommen werden. Zweckmäßig wählt man hierfür auch eine Normalverteilung mit dem Mittelwert δ_0 und der Varianz σ^2/n_0 (sog. conjugate prior). Der Parameter n_0 charakterisiert die 'Präzision' der Verteilung. Ist über δ a priori nichts bekannt, dann wird die Präzision der a priori Verteilung gleich 0 gesetzt. Das bedeutet, daß jeder mögliche Wert δ mit gleicher Wahrscheinlichkeit (die allerdings 0 ist) erwartet wird. Dies ist eine 'uneigentliche' Verteilung, da ihr Integral 0 und nicht 1 ist. Sie führt aber zu einer regulären a posteriori Verteilung. Diese a posteriori Verteilung ist ebenfalls eine Normalverteilung, deren Mittelwert δ_n das gewichtete Mittel aus dem a priori Mittelwert δ_0 und der beobachteten mittleren Differenz \bar{d} und deren Varianz σ_n^2 umgekehrt proportional zur Summe aus der a priori 'Präzision' n_0 und dem Stichprobenumfang n ist. Mit zunehmendem Stichprobenumfang n überwiegt der Einfluß der Beobachtungen, so daß die a posteriori Verteilung immer weniger von den a priori Annahmen abhängt. Soll im Wirksamkeitsvergleich zwischen der Nullhypothese $H_0: \delta \leq 0$ und der Alternative $H_1: \delta > 0$ entscheiden werden, dann sind für beide Hypothesen die a posteriori Wahrscheinlichkeiten zu bilden. Diese sind (mit $\Phi(\cdot)$ als Standard-Normalverteilung):

$$\text{für } H_0: P(H_0|d_1, \dots, d_n) = \Phi\left(\frac{-\delta_n}{\sigma_n}\right) \text{ und für } H_1: P(H_1|d_1, \dots, d_n) = \Phi\left(\frac{\delta_n}{\sigma_n}\right)$$

In Tabelle 2 sind diese beiden Wahrscheinlichkeiten für die simulierten Daten des Sequenzpfades der Abb. 2 angegeben. Als a priori Verteilung für δ wurde die nicht informative Verteilung mit $\delta_0=0$ und $n_0=0$ angenommen.

Im Beispiel ist die a posteriori Wahrscheinlichkeit für H_0 bereits bei der ersten Differenz d_1 von 1.808 kleiner als 0.05. Im frequentistischen Konzept würde man bei dieser Differenz die Nullhypothese auf dem Niveau 0.05 ablehnen, wenn nur 1 Paar vorgesehen ist.

n	d_i	\bar{d}	$P(H_0 \bar{d})$	$P(H_1 \bar{d})$
1	1.808	1.808	0.03531	0.96469
2	0.386	1.097	0.06038	0.93962
3	0.344	0.846	0.07137	0.92863
4	-0.075	0.615	0.10900	0.89100
5	2.588	1.010	0.01195	0.98805
6	0.598	0.941	0.01055	0.98945

Tabelle 2: a posteriori Wahrscheinlichkeit für H_0 und H_1

Im weiteren Verlauf der Datenerhebung steigt die a posteriori Wahrscheinlichkeit für H_0 etwas an (vor allem nach dem 4. Paar, wo mit der Prüfltherapie ein um 0.075 geringeres Ergebnis als mit der Vergleichstherapie erzielt wurde), fällt aber dann auf einen Wert von etwa 0.01, auf dem sie auch beim nächsten Paar verbleibt. Die Evidenz für die Nullhypothese (ausgedrückt durch die a posteriori Wahrscheinlichkeit) ist demnach so gering und die für die Alternativhypothese so groß, daß induktiv die Nullhypothese abzulehnen und die Alternativhypothese anzunehmen ist.

Generell gilt, daß bei Daten, die aus einer Verteilung mit festem aber unbekanntem Parameter stammen, die a posteriori Verteilung für diesen Parameter sich mit zunehmendem Stichprobenumfang immer mehr auf den Parameter konzentriert; und zwar unabhängig von der gewählten a priori Verteilung. Die Evidenz des 'richtigen' Parameterwerts wird mit zunehmendem Stichprobenumfang unabhängig von der a priori Verteilung immer deutlicher. Sie hängt nicht von irgendwelchen 'stopping rules' ab. Es ist daher im Bayesianischen Konzept nicht erforderlich, bei der induktiven Entscheidung über die Wirksamkeit solche 'stopping rules' zu berücksichtigen; die Entscheidung wird nur auf der Basis der vorliegende Daten getroffen. Man kann jederzeit eine Studie fortsetzen, wenn dies aufgrund der Datenlage gerechtfertigt erscheint, ohne daß dadurch andere Auswertungsverfahren erforderlich sind. Ein prinzipieller Unterschied zwischen der Auswertung bei fixem Stichprobenumfang und bei einem sequentiellen Verfahren existiert im Bayesianischen Konzept nicht.

Allerdings erfordert eine gute induktive Aussage eine hinreichende Stabilität der Datenlage. Diese Stabilität kann mit einer 'Sensitivitätsanalyse' überprüft werden. Dabei werden mit der aus den bisherigen Daten ermittelten a posteriori Verteilung der Parameter zukünftige Daten vorausgesagt (prädiziert) und mit den bisherigen erfaßten Daten verglichen. Diese zukünftigen Daten sind unbekannte Größen, denen analog zu den unbekanntem Parametern eine a posteriori Wahrscheinlichkeitsverteilung zugeordnet ist, die mit den bekannten Daten bestimmt wird. Stimmt die Verteilung der zukünftigen Daten mit der Verteilung der beobachteten Daten gut überein (was analog zum frequentistischen Konzept mit einem Signifikanztest überprüft werden kann), dann können die Ergebnisse der Bayesianischen Analyse akzeptiert werden; stimmen die Verteilungen schlecht überein, dann sind neue Daten zu erheben. Ein schlechte Übereinstimmung kann auch auf ein nicht adäquates Modell (Likelihood oder a prior Verteilung) hinweisen. Man muß also auch die Modellannahmen modifizieren und den Einfluß der Modellmodifikation auf die prädizierten Werte untersuchen. Sensitivitätsanalyse und Modellüberprüfung sind wesentliche Bestandteile der Bayesianischen Analyse.

Bei der Versuchsplanung ist zwar keine Vorgabe des Stichprobenumfangs erforderlich, da mit der Sensitivitätsanalyse jederzeit entschieden werden kann, ob die Studie

fortzusetzen oder abzuschließen ist. Man möchte aber zumindest einen Anhalt für den Umfang der geplanten Studie haben. In der Situation des als Beispiel gebrachten Therapievergleichs kann man, analog zum frequentistischen Signifikanztest, auf der Basis der a priori Verteilung von δ die Ergebnisse der zukünftigen Stichprobe präzisieren, mit denen die a posteriori Wahrscheinlichkeit für H_0 höchstens α wird, und dann den Stichprobenumfang n so festlegen, daß die a posteriori Wahrscheinlichkeit für einen Bereich $\delta > \delta_1$ bei diesen Daten mindestens β ist. Für eine nicht informative a priori Verteilung führt dieses Verfahren genau zu dem Stichprobenumfang, der auch beim Signifikanztest erforderlich ist. Nach einer Modifikation von David J. Spiegelhalter [17] kann man den Stichprobenumfang auch über die 'prädiktive Power' festlegen, d.i. die Wahrscheinlichkeit für zukünftige Ergebnisse, die zu einer a posteriori Wahrscheinlichkeit α für H_0 führen. Die prädiktive Power wird mit der bei der Planung zur Verfügung stehenden Verteilung der Parameter berechnet. Der Stichprobenumfang wird so festgelegt, daß diese Power mindestens den Wert $1-\beta$ hat. Dieses Verfahren ist allerdings nicht brauchbar, wenn zu Beginn der Studie nur eine nicht informative a priori Verteilung zur Verfügung steht, da hierfür die prädiktive Power stets 0.5 ist. Man muß also schon aus beobachteten Daten eine informative Verteilung der Parameter ermittelt haben, die auf die Gültigkeit der Alternative hinweist, um überhaupt für irgendein n eine prädiktive Power >0.5 erwarten zu können. Wurde z.B. in einem Vorversuch bei 10 Paaren eine mittlere Differenz von 0.5 beobachtet, dann wäre eine Stichprobe von weiteren 30 Paaren erforderlich, um eine prädiktive Power von 0.8 zu erreichen. Wurde im Vorversuch die mittlere Differenz von 0.5 nur bei 5 Paaren beobachtet, dann wären für eine prädiktive Power von 0.8 weitere 150 Paare erforderlich. Wurde im Vorversuch bei 5 Paaren eine mittlere Differenz von 0.7 beobachtet, dann würde bereits mit weiteren 15 Paaren eine prädiktive Power von 0.8 erreicht. Diese Beispiele zeigen, wie im Bayesianischen Konzept eine Versuchsplanung betrieben werden kann und wie wichtig dabei die Vorkenntnis über die Verteilung der Parameter ist. Die klinische Prüfung sollte daher stets in Stufen (d.h. sequentiell) durchgeführt werden.

Die Anwendung des Bayesianischen Konzepts ist nicht auf die Annahme beschränkt, daß die Ergebnisse bei allen Patienten nach Anwendung der Prüf- bzw. Vergleichstherapie jeweils derselben Verteilung gehorchen. Diese Situation ist unrealistisch, da jedem Arzt bekannt ist, daß es für jede Therapie 'Responder' und 'Non-Responder' gibt. Dies wird im Bayesianischen Konzept dadurch berücksichtigt, daß die Parameter der Response-Verteilung zufällig von Patient zu Patient variieren können. Die Verteilung der beobachteten Ergebnisse ist also stets eine Mischverteilung mit unterschiedlichen Parameterwerten. Nach dem oben angegebenen Grenzwertsatz wird mit zunehmendem Stichprobenumfang die a posteriori Verteilung zur 'wahren' Verteilung der Parameterwerte konvergieren. Es macht dabei wenig Sinn, für die Parameterwerte Null- und Alternativhypothesen vorzugeben. Sinnvoller ist es, zur Bewertung der Wirksamkeit die a posteriori Wahrscheinlichkeit zu ermitteln, mit der die Parameter der Prüftherapie besser sind als die der Vergleichstherapie (vergl. Schneider [16]). Durch die Einführung von 'Verlustfunktionen', die den Verlust bei falscher Entscheidung quantifizieren, kann der Wirksamkeitsvergleich als quantifizierbares Entscheidungsproblem behandelt werden.

Eine weitere Verfeinerung der Auswertung wird durch die Verwendung 'hierarchischer Modelle' erreicht, bei denen die Verteilung der Ergebnisse in hierarchisch gegliederten Stufen modellmäßig aufgebaut werden; z.B. wird auf der 1. Stufe das Ergebnis bei einem Patienten als normal verteilt angenommen, deren Mittelwert auf der

2. Stufe als Funktion der Behandlung und von Charakteristika des Patienten (Alter, Vorerkrankung u.ä.) modelliert wird, deren Funktionsparameter auf der 3. Stufe als Zufallsgrößen mit sogenannten 'Hyperparametern' angesetzt werden. Die Verteilungen der unbekanntenen Größen einer Stufe hängen nur von den Parametern der unmittelbar übergeordneten Stufe ab. Mit den beobachteten Daten werden für die einzelnen Parameter oder sonstigen unbekanntenen Größen (z.B. auch für fehlende Werte) die a posteriori Verteilungen berechnet. Hierfür haben sich komplexe Simulationsverfahren nach dem Prinzip der Markov-Ketten (z.B. Gibbs-Sampling) bewährt. Diese Modelle sind geeignet, auch bei Studien mit einem komplexen Design (wie z.B. Anwendungsbeobachtungen) die Zusammenhänge problemadäquat zu beschreiben. In einer für einen individuellen Wirksamkeitsnachweis interessanten Modifikation kann bei jedem Patienten sowohl für die Prüftherapie als auch für die Vergleichstherapie ein Ergebnis angenommen werden, von denen aber nur das mit der tatsächlich angewendeten Therapie erzielte beobachtet werden kann. Das Ergebnis für die andere Therapie ist eine unbekanntene Zufallsgröße, deren a posteriori Verteilung mit den beobachteten Daten bestimmt wird. So läßt sich für jeden Patienten individuell zu seinen Charakteristika ein Therapievergleich durchführen.

Zahlreiche hierarchische Bayesianische Modelle sind in dem von Spiegelhalter und Mitarbeitern entwickelten Programmsystem BUGS (Bayesian inference Using Gibbs Sampling) als Beispiele enthalten, das zur Bayesianischen Analyse über das Internet abgerufen werden kann (<http://www.mrc-bsu.cam.ac.uk/bugs.html>). Es ist zu hoffen, daß damit die Bayesianischen Verfahren, die den heute noch weitgehend verwendeten frequentistischen Verfahren überlegen sind und mit denen insbesondere auch die Individualität der Patienten berücksichtigt werden kann, größere Verbreitung erfahren. In einer EMEA Richtlinie wird ausdrücklich festgestellt, daß diese Auswertungsverfahren, die 'not p-value based' sind, von den Zulassungsbehörden akzeptiert werden.

Schlußbemerkungen

Es wurde das Problem der Induktion, d.h. von Aussagen, die aus der Erfahrung über allgemeine Zusammenhänge gemacht werden, behandelt und gezeigt, wie mit induktiven Schlußweisen die Wirksamkeit von Arzneimitteln nachgewiesen werden kann. Die Wirksamkeit ist zu verstehen als die Fähigkeit des Arzneimittels, bei der vorgesehenen Indikation mehr Patienten Heilung oder Linderung ihrer Beschwerden zu bringen, als ohne das Arzneimittel zu erwarten wäre. Zum Wirksamkeitsnachweis genügt es nicht, auf die Erfahrung hinzuweisen, die mit dem Arzneimittel bei einzelnen Patienten gemacht wurde. Diese Erfahrung muß vielmehr induktiv verallgemeinert und mit der Erfahrung verglichen werden, die ohne das Arzneimittel gemacht wurde. Dies setzt eine hohe Qualität der Erfahrung, wie sie am ehesten in Studien (kontrollierte Studien, Anwendungsbeobachtungen als Kohortenstudien) erreicht wird, und die Anwendung adäquater statistischer Auswertungsmethoden voraus, von denen insbesondere die Bayesianischen Verfahren empfohlen werden können. Diese Grundsätze gelten gleichermaßen sowohl für Arzneimittel und Behandlungsmethoden der 'Schulmedizin' als auch für die der Naturheilverfahren oder alternativer Therapierichtungen. Dabei können aber durchaus die Studien und induktiven Auswertungsverfahren den jeweiligen Therapiekonzepten angepaßt werden.

Neben der Erfahrung und den daraus gewonnenen induktiven Schlüssen werden oft auch theoretische Vorstellungen in Form von 'Kausalmodellen' oder 'Kausalgesetzen'

für die Behauptung der Wirksamkeit angeführt. Solche Vorstellungen sind sehr nützlich zur **Erklärung** einer möglichen Wirksamkeit. Für den Wirksamkeitsnachweis sind sie aber weder notwendig noch hinreichend. Dies können sie schon deshalb nicht sein, da sich 'Kausalgesetze' nie vollständig verifizieren lassen. Auch die zur Zeit anerkannten 'Naturgesetze' sind keine allgemeingültigen Regeln, sondern nur 'vorläufige Hypothesen', mit denen wir die Erscheinungen der Natur in Zusammenhang zu bringen versuchen. Darauf hat bereits Ernst Mach hingewiesen: "Man spricht oft von Naturgesetzen. Was bedeutet dieser Ausdruck? Gewöhnlich wird man der Meinung begegnen, die Naturgesetze seien Regeln, nach welchen die Vorgänge in der Natur sich richten *müssen*, ähnlich den bürgerlichen Gesetzen, nach welchen die Handlungen der Bürger sich richten *sollen*. Einen Unterschied pflegt man darin zu sehen, daß die letzteren Gesetze auch übertreten werden können, während man Abweichungen der Naturvorgänge von ersteren für unmöglich hält. Diese Auffassung der Naturgesetze wird aber erschüttert durch die Überlegung, daß wir ja nur aus den Naturvorgängen selbst die Naturgesetze ablesen, abstrahieren, und daß wir hierbei vor Irrtümer durchaus nicht gesichert sind. ... Angesichts dieser Überlegungen wird vielleicht folgende naheliegende Fassung Zustimmung finden: Ihrem Ursprunge nach sind die 'Naturgesetze' Einschränkungen, die wir unter Leitung der Erfahrung unserer Erwartung vorschreiben" [10].

Ein nach den hier behandelten Grundsätzen erbrachter Wirksamkeitsnachweis kann nicht deshalb verworfen werden, weil für die Wirksamkeit keine 'Kausalhypothese' existiert oder die Wirksamkeit angeblichen 'Naturgesetzen' widerspricht. Umgekehrt können 'Kausalhypothesen' oder andere hypothetische 'Prinzipien' den induktiven Wirksamkeitsnachweis mit Studien nicht ersetzen. Dies gilt auch für die Prinzipien alternativer Therapierichtungen, wie der Homöopathie oder Anthroposophie. Unbestreitbar ist allerdings, daß solche Prinzipien oder auch nur 'Vorstellungen' die Akzeptanz bei den Patienten erheblich verbessern können, oft weit mehr als ein korrekt durchgeführter Wirksamkeitsnachweis. Dies ist aber eine Frage der Psychologie, auf die nicht weiter eingegangen werden soll.

Werden hypothetische Vorstellungen zur Begründung der Wirksamkeit gebracht, so müssen diese Vorstellungen dem Abgrenzungskriterium Poppers (vgl. Abschnitt 1.3) entsprechen, wenn sie von der etablierten Wissenschaft akzeptiert werden sollen. Das bedeutet, daß die Theorien in sich geschlossen und logisch konsistent sein müssen. Aus ihnen müssen logisch deduktiv 'Basissätze' ableitbar sein, die empirisch nachprüfbar und damit auch falsifizierbar sind. Der bloße Hinweis auf 'geheimnisvolle Kräfte' (z.B. bei der Bachblütentherapie) oder auf mehr oder minder apokryphe Texte von 'Weisen' oder 'großen Heilern' wird ebensowenig als wissenschaftliche Erklärung akzeptiert werden, wie der zusammenhanglose Hinweis auf naturwissenschaftliche Theorien (wie z.B. der Quantentheorie oder Chaostheorie).

Literaturverzeichnis

1. Bayes T.: An Essay towards solving a Problem in the Doctrine of Chances. Philosophical Transactions of the Royal Society of London. 58, 370-418, 1763
2. Cassirer E.: Das Erkenntnisproblem in der Philosophie und Wissenschaft der neueren Zeit. Sonderausgabe, Reprint, Wiss. Buchges. Darmstadt, 1994
3. Cornfield J.: Sequential Trials, Sequential Analysis and the Likelihood Principle. The American Statistician, 20, 18-23, 1966
4. Flach W.: Grundzüge der Erkenntnislehre: Erkenntniskritik, Logik, Methodologie. Königshausen und Neumann Würzburg 1994
5. Hume D.: Eine Untersuchung über den menschlichen Verstand. Philosophische Bibliothek Band 35, Felix Meiner Verlag Hamburg 1973
6. Kant I.: Kritik der reinen Vernunft. Philosophische Bibliothek Band 37a, Felix Meiner Verlag Hamburg 1976
7. Kuhn T.S.: Die Struktur wissenschaftlicher Revolutionen. Suhrkamp Verlag Frankfurt 1967
8. Kuhn T.S.: Die Entstehung des Neuen. Suhrkamp Taschenbuch Wissenschaft 236, Frankfurt 1977
9. Lauritzen S.L. and Spiegelhalter D.J.: Local computations with probabilities on graphical structures and their application to expert systems (with discussion) Journal of the Royal Statistical Society Ser. B, 50, 157-224, 1988
10. Mach E.: Erkenntnis und Irrtum. Reprografischer Nachdruck, Wiss. Buchges. Darmstadt 1991
11. O'Brien P.C. and Fleming T.R.: A Multiple Testing Procedure for Clinical Trials. Biometrics, 35, 549-556, 1979
12. Pearson E.S.: The Choice of Statistical Tests Illustrated on the Interpretation of Data Classed in a 2x2 Table. Biometrika, 34, 139-167, 1947
13. Pocock S.J.: Interim Analyses for Randomized Clinical Trials: The Group Sequential Approach. Biometrics, 38, 153-162, 1982
14. Popper K.R.: Logik der Forschung. 7. Auflage, J.C.B. Mohr Tübingen 1982
15. Popper K.R.: Objective Knowledge. An Evolutionary Approach. 7th impression, Clarendon Press Oxford 1992
16. Schneider B.: Bayesian Models for Clinical Studies. Methods of Information in Medicine, 23, 147-153, 1984
17. Spiegelhalter D.J., Freedman L.S. and Parmar M.K.B.: Bayesian Approaches to Randomized Trials. In: Bayesian Biostatistics (edited by Donald A. Berry and Dalene K. Stangl). Marcel Dekker, Inc. New York, Basel, Hong Kong 1996
18. Wald A.: Sequential Analysis. John Wiley & Sons, Inc., New York, London, Sidney 1947
19. Whitehead J.: The Design and Analysis of Sequential Clinical Trials. Ellis Horwood Ltd., Chichester 1983 (2nd edition 1992)