.

Evidenz und Inferenz

U. Mortensen

13. 04. 2010

Letzte Korrektur: 05. 05. 2014

FB Psychologie und Sportwissenschaften
Westfälische Wilhelms-Universität Münster

Die Ergebnisse der von der "orthodoxen" Theorie statistischer Tests von Hypothesen (Fisher, Neyman & Pearson) sind nicht notwendig gleich der Evidenz, die in den Daten für oder gegen die betrachteten Hypothesen enthalten ist. Es werden einige Betrachtungen zur üblichen Inferenzstatistik vorgestellt; es wird dann argumentiert, dass die Bayessche Statistik der Forderung nach einer Repräsentation der Evidenz in den Daten am ehesten entspricht.

Inhaltsverzeichnis

Hinweis: außer dem Literaturverzeichnis findet man einen Index am Ende des Textes.

1 Einführung

1.1 Wahrscheinlichkeiten, Evidenz, und Evidenzmaße

Ein wichtiger Aspekt des wissenschaftlichen Handwerks besteht in der Diskussion der Kompatibilität empirischer Daten mit Hypothesen und Theorien. Die Frage ist insbesondere, ob Daten eine Hypothese stützen oder ihr widersprechen, oder ob die Daten gar nichts über die zur Diskussion stehenden Hypothesen aussagen. Es geht um *statistical evidence*, also um die Evidenz[1], die in den Daten in Bezug auf die in Frage stehenden Hypothesen steckt. Der Begriff der Evidenz (evidence) ist komplex (Kelly, 2006) und bezieht sich hier zunächst nur auf die Daten bzw. auf Aspekte der Daten, die für eine Bewertung der Hypothesen relevant sind.

In Wissenschaften wie Psychologie, Biologie, Medizin sowie in den Sozialwissenschaften dominiert eine Art von Hybridansatz, eine Melange aus Fisherschem Signifikanztest und Neyman-Pearsonschem Hypothesentest, deren routinemäßige und oft genug rein schematische Anwendung zwar oft gegeißelt wurde (etwa Gigerenzer (1993), Ostmann & Wutke (1994), Sedlmeier (1996), Moosbrugger & Brandl (2002)), ohne deshalb an Attraktivität eingebüßt zu haben, was einerseits darauf zurückzuführen sein mag, dass viele Anwender der Statistik diese eher als notwendiges Handwerkszeug betrachten, ohne sich mit der damit verbundenen Problematik auseinander zu setzen, und andererseits der Verfügbarkeit dieses Ansatzes in gängigen Programmpaketen wie SPSS und STATISTICA geschuldet sein mag. Den Klagen darüber soll hier keine weitere hinzugefügt werden, vielmehr soll die Frage diskutiert werden, ob Signifikanztests (Fisher) oder Hypothesentests (Neyman & Pearson) die 'Evidenz' in den Daten bezüglich der zur Diskussion stehenden Hypothesen hinreichend ausschöpfen oder sie eventuell eher verschleiern, und ob nicht andere Ansätze, etwa die Likelihood-Inferenz oder Bayessche Verfahren, besser geeignet sind, die relevante Evidenz aus den Daten zu extrahieren.

Der Satz von Bayes: Will man Evidenz quantitativ mit den zur Diskussion stehenden Hypothesen in Beziehung setzen, so liegt es nahe, dies über die bedingten Wahrscheinlichkeiten $P(E|H)$ – die Wahrscheinlichkeit von E, gegeben die Hypothese H – bzw. $P(H|E)$ – die Wahrscheinlichkeit der Hypothese H, gegeben die Evidenz E – zu versuchen. 1764 erschien posthum eine Schrift des Mathematikers und Pfarrers Thomas Bayes, *Essay Towards Solving a Problem in the Doctrine of Chances*. Bayes leitete eine heute unter dem Namen *Satz von Bayes* bekannte Beziehung zwischen Daten bzw. der mit den Daten verbundenen *Evidenz* E, dem Kontext K und einer Hypothese H her:[2]

$$\begin{aligned} P(H|E,K) &= P(E|H,K)\frac{P(H,K)}{P(E)} \\ &= \frac{P(E|H,K)}{P(E|H,K)P(H,K) + P(E|\neg H,K)P(\neg H,K)}, \end{aligned} \tag{1}$$

[1] Der hier gebrauchte Begriff der Evidenz enstpricht dem des englischen 'evidence' mit der Bedeutung von z.B. Beweis, Nachweis, Indiz oder Information. Im Deutschen hat 'Evidenz' auch die Bedeutung von unmittelbarer Einsicht; nach F. Brentano impliziert 'Evidenz' sogar Wahrheit. Diese Bedeutungen werden in diesem Text *nicht* impliziert. Der Grund für diese Verwendung von 'Evidenz' liegt im Gebrauch von 'evidence' in der Literatur und der Umständlichkeit, mit der die Bedeutung des englischen 'evidence' im Deutschen umschrieben werden muß.

[2] Die Formel (1) kommt im Text von Bayes in dieser Form nicht vor (Bayes (1764, 1958); Earman (1992) gibt eine Einführung in die Bayessche Argumentation.

wobei $\neg H$ die Negation von H ("nicht-H") ist, und $P(H, K)$ ist die *a-priori*-Wahrscheinlichkeit der Hypothese H im Kontext K, im Unterschied zur *a-posteriori*-Wahrscheinlichkeit $P(H|E, K)$ von H; weiter gilt $P(\neg H, K) = 1 - P(H, K)$. $P(H|E, K)$ ist die Wahrscheinlichkeit von H, gegeben die Evidenz E (im Kontext K), $P(E|H, K)$ ist die Wahrscheinlichkeit der Evidenz E unter der Annahme von H (im Kontext K). Nach (1) sind die Wahrscheinlichkeiten $P(H|E, K)$ und $P(E|H, K)$ "invers" zueinander, weshalb auch von *inversen Wahrscheinlichkeiten* die Rede ist. Für $P(E|H, K)$ hat sich der von Fisher (1925) eingeführte Ausdruck 'Likelihood' eingebürgert.

Von einem formalen Standpunkt aus gesehen ist (1) eine unmittelbare Implikation der Definition des Begriffs der bedingten Wahrscheinlichkeit,

$$P(A|B) = \frac{P(A \,\&\, B)}{P(B)}, \quad P(B) > 0,$$

wobei A und B zufällige Ereignisse sind. Für $P(A) > 0$ gilt analog dazu $P(B|A) = P(A \,\&\, B)/P(B)$, so dass

$$P(A|B) = P(B|A)\frac{P(A)}{P(B)}$$

folgt. Wenn A und B zufällige Ereignisse sind, ist diese Beziehung unproblematisch. Die Diskussion um den Bayesschen Satz entsteht wegen der Deutung von H als einem zufälligen Ereignis. Bei Jakob Bernoulli, Thomas Bayes und Simon de Laplace scheint die epistemische Deutung des Wahrscheinlichkeitsbegriffs als natürlich empfunden worden zu sein, bis ab Mitte des 19-ten Jahrhunderts die Subjektivität dieser Interpretation kritisiert wurde (s. unten), und spätestens seit Richard von Mises (1921) Charakterisierung der Wahrscheinlichkeit eines zufälligen Ereignisses als Grenzwert der relativen Häufigkeit r_A/n für $n \to \infty$ und damit als einer 'objektiven' Größe kam der epistemische Begriff der Wahrscheinlichkeit in Verruf, obwohl Wahrscheinlichkeit umgangssprachlich durchaus als epistemische Größe verstanden wird. Der subjektive Wahrscheinlichkeitsbegriff wurde wiederbelebt durch Keynes (1921/2008) *Treatise on Probability*, Ramsey (1926), Jeffreys (1939), Bruno de Finetti (1937), Savage (1954/1972), nur um die wichtigsten Autoren zu nennen. Fienberg (2006) gibt einen Überblick über die Entwicklung dieses Begriffs.

Trotz der intensiven Kritik der 'Frequentisten' (insbesondere Fisher sowie Neyman & Pearson) ist es intuitiv naheliegend, die epistemische Interpretation zunächst einmal zu akzeptieren, da sie erlaubt, über den Bayesschen Satz genauer zu definieren, was mit Evidenz für eine Hypothese gemeint sein soll: man kann

$$\mathcal{E} = P(H|E) - P(H) \tag{2}$$

als *inkrementelle Evidenz* betrachten, die über E induziert wird. Der Kontext K wurde hier der Übersichtlichkeit wegen weggelassen, d.h. in die Wahrscheinlichkeit P absorbiert. Für $P(H|E) - P(H) > 0$ erhöht E die Wahrscheinlichkeit für H, und für $P(H|E) - P(H) < 0$ ist E gegenindikativ für H. für $P(H|E) - P(H) = 0$ verändert E die Einschätzung von H nicht. $P(H|E) - P(H) > 0$ ist äquivalent zu

$$r(H, E) = \frac{P(H|E)}{P(H)} > 1; \tag{3}$$

Carnap (1950) führte $r(H, E)$ als *Relevanzquotienten* ein. Für $r(H, E) < 1$ spricht die Evidenz gegen H. Je weniger $r(H, E)$ von 1 abweicht, desto geringer ist die Relevanz von E für H. Der Satz von Bayes liefert sofort

$$\frac{P(H|E)}{P(H)} = \frac{P(E|H)}{P(E)} = r(E, H), \tag{4}$$

also

$$r(H, E) = r(E, H). \tag{5}$$

$P(E)$ kann über den Satz der Totalen Wahrscheinlichkeit ausgedrückt werden:

$$P(E) = P(E|H)P(H) + P(E|\neg H)P(\neg H), \;\; P(\neg H) = 1 - P(H).$$

Joyce (2003) gibt allerdings eine interessantere Interpretation von $P(E)$: diese Wahrscheinlichkeit repräsentiere eine "baseline" Vorhersagbarkeit von E aufgrund des Hintergrundwissens, dass im Wahrscheinlichkeitsmaß P kodifiziert wird, und $P(E|H)$ repräsentiert die Vorhersagbarkeit von E, wenn die Hypothese H zu diesem Hintergrund hinzugenommen wird. Das Relevanzmaß $r(E, H)$ reflektiert dann das Ausmaß, in dem E mehr oder weniger vorhersagbar wird relativ zu $P(E)$, wenn H bekannt ist. $r(E, H) = 0$ bedeutet dann, dass H die Daten bzw. die Evidenz $\neg E$ voraussagt. $r(E, H) = 1$ wiederum besagt, dass die Hypothese H die baseline-Vorhersage überhaupt nicht ändert ($P(E|H)/P(E) = 1$ impliziert ja $P(E|H) = P(E)$), und $r(E, H) = 1/P(E)$ schließlich bedeutet, dass H E kategorisch vorhersagt.

Die *Odds*, d.h. die Wettchancen, sind durch den Quotienten

$$\omega(H) = \frac{P(H)}{P(\neg H)} = \frac{P(H)}{1 - P(H)} \tag{6}$$

definiert. Ist etwa $P(H) = 3/5$, $P(\neg H) = 2/5$, so ist $\omega(H) = 3/2$ d.h. die Chance steht 3 zu 2, dass H korrekt ist. Ist P eine irrationale Zahl, lassen sich die bestapproximierenden natürlichen Zahlen finden, um eine derartige anschauliche Beschreibung zu finden. Auch die Odds können als Relevanzmaß betrachtet werden.

Insbesondere ist der *Odds Ratio* als ein weiteres Relevanz- oder Evidenzmaß zu nennen. Dies ist das Verhältnis

$$\rho(H) = \frac{\omega_E(H)}{\omega(H)} = \frac{P(E|H)}{P(E|\neg H)}, \tag{7}$$

mit $\omega_E(H) = P(H|E)/P(\neg H|E))$.

Die bisher eingeführten Maße beziehen sich jeweils auf eine Hypothese und sind also *nicht-relational*. Viele Autoren sind der Ansicht, dass Hypothesen nur relativ zu anderen Hypothesen diskutiert werden könnten, etwa Neyman & Pearson (1928, 1933). Dementsprechend sind *relationale Maße* gewünscht.

Ein relationales Relevanzmaß ist der *Likelihood-Quotient* $\lambda(H, E)$ ein Evidenz- bzw. Relevanzmaß; Dieser ist durch

$$\lambda(H, E) = \frac{P(E|H)}{P(E|\neg H)} \tag{8}$$

definiert. Offenbar ist dann

$$\frac{P(H|E)}{P(\neg H|E)} = \lambda(H, E)\omega(H). \tag{9}$$

Die Interpretation von Wahrscheinlichkeiten: Die Relevanzmaße sind intuitiv plausibel, insbesondere wenn Wahrscheinlichkeiten als epistemische Größen akzeptiert werden. Die Frage ist, in welchem Sinne H und E zufällige Ereignisse sind. Für E ergibt sich die Antwort aus der Tatsache, dass Daten sich als Stichproben und damit als statistische Größen ergeben, die aus H abgeleitet werden können. Andererseits erscheint die Interpretation einer Hypothesen als einem zufälligen Ereignis als gegenintuitiv und erfordert eine geeignete Charakterisierung des Wahrscheinlichkeitsbegriffs. Insbesondere wird die Frage, wie a-priori-Wahrscheinlichkeiten $P(H, K)$ zu definieren sind, kontrovers diskutiert. In Abschnitt 6.2 wird eine Definition von Wahrscheinlichkeiten für Aussagen sowie ein – sehr – kurzer Abriss der existierenden Interpretationen von Wahrscheinlichkeiten gegeben.

Induktion: Darüber hinaus ergibt sich die Frage nach der Möglichkeit der Induktion: findet man etwa $P(H|E, K) > P(H, K)$, so könnte man sagen, dass E eine bestätigende oder gar verifizierende Wirkung in Bezug auf H hat, die als Resultat induktiven Schließens gedeutet werden kann. Aber nach David Hume (1731) ist Induktion nicht möglich, wie Howson (2000) noch einmal versichert. Diese sicherlich etwas saloppe Formulierung des Humeschen Diktums genügt vorerst als Hintergrund für die Feststellung, dass Poppers außerordentlich einflußreicher falsifikationistischer Ansatz (in *Logik der Forschung*, (1934)) auf eben diesem Befund Humes beruht; dort liefert er in einem 1981 hinzugefügten Anhang XVII einen Beweis, dass eine auf (1) beruhende Bayessche Statistik nicht möglich sei, und Popper & Miller (1983) und (1987) liefern weitere Argumente, weshalb Induktion und deshalb auch eine probabilistische Stützung von Hypothesen nicht möglich seien, – es gebe, so Popper & Miller, nur eine deduktive, nicht aber eine induktive Logik. In Abschnitt 6.3.2 (p. 94) werden die Popperschen Argumente detaillierter dargestellt.

Natürlich sind die Popper & Millerschen Beweise nicht unwidersprochen geblieben; in Earman (1992) und Howson (2000) findet man die Gegenargumente und zumindest einen Teil der Literatur zu diesen Behauptungen. Gleichwohl herrscht Skepsis: Moosbrugger & Brandl (2002) ist der Bayes-Ansatz keinen Absatz, sondern nur einen Nebensatz über 'Bayessches Wunschdenken' wert, und das vermeintlich Subjektive des Bayes-Ansatzes führt bei einigen Philosophen der zweiten Hälfte des zwanzigsten Jahrhunderts zu einem nachgerade emotionalen Ausbrüchen:

> Der subjektive Ansatz ... " verliert sich im statistischen Fall in einem dämmrigen Licht von falschen Behauptungen, zweifelhaften Analogien und wirklichkeitsfremden Idealisierungen" (Stegmüller (1983; 244)).

Stegmüller sieht die Gefahr einer radikalen Subjektivierung der Naturwissenschaft, würde sich die subjektive Statistik durchsetzen. Auf die Frage, ob diese Aussage in selbtsreferentieller Weise ihre eigene Charakteristik reflektiert, ist noch einzugehen.

Die Frage nach der Möglichkeit der Induktion erweist sich als quälend, zumal sich sich auch bei nicht-Bayesschen Ansätzen zum Test von Hypothesen stellt; sie werden in Abschnitt 6.3 vorgestellt.

1.2 Hypothesenevaluation: Der orthodoxe Ansatz

Während der Bayessche Ansatz von Laplace (1781) akzeptiert wurde und Eingang u. A. in Laplacesche Arbeiten zur Physik fand, attackierten der Mathematiker George Boole (1854) und der Philosoph John Venn (1866) die Arbeit Bayes insbesondere wegen der Notwendigkeit, eine a-priori-Wahrscheinlichkeit für H definieren zu müssen, und fanden mit ihrer Kritik insbesondere bei Biologen Gehör, die sich mit Darwins Theorie beschäftigten (vergl. Jaynes (2003; 316)). Zu Beginn des zwanzigsten Jahrhunderts waren es wiederum insbesondere Biologen, die sich der Kritik an Bayes und Laplace anschlossen[3] und – wohl unter dem Einfluß Fishers – ihre Aufmerksamkeit auf die Likelihood $P(E|H)$ fokussierten. Sie bildeten eine, wie Jaynes es formuliert, 'extremely aggressive school', die die Entwicklung der Statistik dominierte und den theoretischen Rahmen schuf, der heute als 'orthodoxe Statistik' bezeichnet wird. Jaynes zielt auf Ronald Aylmer Fisher, Jerzy Neyman und Egon Pearson, die sich zwar mit biologischen Fragen beschäftigten, aber alle aus der Mathematik kamen.

In der Tat ist es der zwar nicht von R. A. Fisher erfundene, aber doch von ihm favorisierte, logisch begründete und in vielen konkreten Anwendungen eingesetzte Signifikanztest, der wohl am häufigsten bei der statistischen Überprüfung von Hypothesen angewendet wird; Fishers *Statistical Methods for Research Workers* (1925) – mit 13 Auflagen bis 1960, eine deutsche Ausgabe 1956 – wurde das dominierende Lehrbuch für Anwender der Statistik, geschrieben als

> "... cookbook, which, in effect, told one to 'do this ... then do that ... and don't ask why.' "(Jaynes (2003), p. 492).

Der konzeptionelle Hintergrund für den Signifikanztest ist jedenfalls eine Art Axiom oder Prinzip, das von Royall (1997; 65) als *Law of Improbability* eingeführt wurde: wird ein Ereignis beobachtet, das bei Geltung der Hypothese H eine kleine Wahrscheinlichkeit p hat, so ist dieses Ereignis Evidenz gegen H; man kann den Fisherschen Signifikanztest als stochastische Variante des Popperschen Falsifikationsprinzips sehen: ist p hinreichend klein, etwa $p < .05$, so wird, dem Fisherschen Reglement entsprechend, H "verworfen". Für hinreichend große p-Werte wird H zunächst einmal beibehalten, gilt aber keineswegs als verifiziert, die Möglichkeit, anhand anderer Daten verworfen zu werden, besteht für H auch weiterhin.

Auf den ersten Blick erscheint die Annahme des *Law of Improbability* – der Ausdruck 'Law' ist sicherlich ein wenig überzogen – plausibel, aber ein genauerer Blick darauf wird zeigen, dass man ihm kaum die Grundsätzlichkeit zubilligen kann, die eine intuitive Betrachtung nahelegt. p wurde von Fisher als Maß für die Evidenz gegen die getestete Hypothese $H = H_0$ in einem einzelnen Experiment betrachtet, gewissermaßen als Maß für die Möglichkeit, an H_0 zu glauben bzw. nicht zu glauben, wobei p nicht das einzige Maß sein sollte, um über H_0 zu entscheiden. Für Fisher reflektiert die Wahrscheinlichkeit eines Ereignisses E den Grad an Ungewißheit für eine einmalige Beobachtung von E. Diese Auffassung entspricht derjenigen, die Theorien induktiven Schließens unterliegt. Andererseits ist $p(E)$ durch die relative Häufigkeit von E in der *Referenzmenge* (reference set) definiert; dies ist die Menge der möglichen Stichproben,die bezüglich H_0 gebildet werden können. Man kann dies als Versuch sehen, die Möglichkeit der Induktion mit der Idee objektiver Wahrscheinlichkeiten zu verknüpfen; Fisher spricht explizit von *in-*

[3] Weil sie, so Jaynes, die Mathematik Laplaces nicht verstanden.

ductive inference (Fisher (1959), Johnstone (1987)); das Humesche Diktum über die Unmöglichkeit der Induktion wird mittels intuitiv ansprechender Metaphorik überspielt.

Ein Signifikanztest der Fisherschen Art fokussiert auf eine Hypothese, etwa H_0. Wenn H_0 verworfen wird, weil p "hinreichend" klein ist, so wird die Negation von H_0, $\neg H_0$, akzeptiert. Das Problem mit diesem Vorgehen ist allerdings, dass $\neg H_0$ nicht immer spezifiziert ist. So kann H_0 die Hypothese sein, dass bestimmte Messungen gemäß $N(\mu, \sigma^2)$ verteilt sind, also normalverteilt mit Erwartungswert $\mathbb{E}(X) = \mu$ und Varianz $\mathbb{V}(X) = \sigma^2$ sind. $\neg H_0$ kann bedeuten, dass die X-Werte eben nicht normalverteilt sind, oder normalverteilt sind mit dem Erwartungswert $\mu' \neq \mu$, oder normalverteilt mit der Varianz $(\sigma')^2 \neq \sigma^2$ sind, etc. Wenn darüber hinaus das Prinzip der Kleinen Wahrscheinlichkeit keine universelle Gültigkeit hat, könnte der Vergleich von Hypothesen helfen; man käme zu einem *relationalen Maß* für die Bewertung von Hypothesen.

Die Bewertung von Hypothesen relativ zueinander liegt im Kern des Ansatzes von Neyman & Pearson (1928, 1933). Die ursprüngliche Absicht dieser Autoren war, einige Unklarheiten in Fishers Ansatz klarzustellen, aber ihre Arbeit lieferte dann eine von Fishers Ansatz verschiedene Alternative, die zu einem jahrzehntelangen Streit zwischen Fisher auf der einen Seite und insbesondere Neyman auf der anderen Seite führte. Neyman & Pearson stellen zunächst fest, dass es zwei Arten von Entscheidungsfehlern gibt: beim Fehler der ersten Art wird die Hypothese H_1 verworfen, obwohl H_1 wahr ist, und H_2 akzeptiert, und beim Fehler zweiter Art wird H_2 verworfen, obwohl H_2 wahr ist, und H_1 akzeptiert. Die Entscheidung für eine Handlung impliziert eine Entscheidung für eine korrespondierende Handlung, a_1 oder a_2 (Wahl einer Therapie, Entscheidung für oder gegen eine Investition, etc). Weiter sei α die Wahrscheinlichkeit eines Fehlers erster Art, und β sei die Wahrscheinlichkeit eines Fehlers zweiter Art. Man kann nun, wie Neyman & Pearson zeigen, die Wahrscheinlichkeit α des Fehlers der ersten Art festsetzen, etwa $\alpha = .05$ oder $\alpha = .01$, und dann einen Bereich R definieren derart, dass man sich für H_2 entscheidet, wenn $T(D) \in R$, wobei T eine geeignet auf den Daten D definierte Statistik ist; insbesondere soll gelten (i) $P(T(\mathbf{x}) \in R|H_1) = \alpha$ und (ii) $P(T(\mathbf{x}) \notin R|H_2) = \beta$ minimal ist; die *Macht* (power) des Tests ist dann durch $1 - \beta$ definiert. Für vorgegebenen Wert von α wird also die Macht maximiert. Dies ist der Neyman-Pearson-(NP-)Ansatz. In Abschnitt 7.1, Seite 128, wird eine vollständigere Charakterisierung Der Neyman-Pearson-Theorie gegeben.

Der Neyman-Pearson-Test liefert kein kontinuierliches Maß für die Evidenz für oder gegen eine Hypothese wie $P(H|D)$ oder die Likelihood $P(D|H)$; die Bewertung der Hypothesen auf der Basis der Daten wird dichotomisiert. Wahrscheinlichkeiten werden von Neyman & Pearson streng frequentistisch interpretiert und werden dementsprechend von den Autoren nicht, wie von Fisher, als Ungewißheitsmaße interpretiert. Da darüber hinaus ein NP-Test auf eine Handlungsanweisung hinausläuft, wird auch keine Beziehung zu induktivem *Schließen* hergestellt, die Rede ist vielmehr von induktivem *Verhalten*:

> If a rule R unambiguously prescribes the selection of action for each possible outcome ..., then it is a rule for inductive behavior. (Neyman (1950), p. 10)

Auch hier gilt wegen der Vernachlässigung Humescher Skepsis gegenüber der Induktion die oben in Bezug auf die Fishersche *inductive inference* angefügte Bemer-

kung über eine intuitiv ansprechende Metaphorik.

Die Dichotomisierung hinsichtlich zu wählender Aktionen suggeriert vielleicht, impliziert aber auf jeden Fall nicht, dass eine Entscheidung etwa für H_j bedeutet, dass H_j für wahr und $H_{j'}$ für falsch gehalten wird. In der Tat warnt Neyman (1950) von einer derartigen Interpretation: die Rede von der Akzeptanz und Zurückweisung von Hypothesen sei zwar sehr bequem und üblich, aber sie beziehe sich eben nur auf die Wahl einer Handlung und bedeute nicht, dass eine Hypothese für wahr und die andere für falsch zu halten sei (p. 259). Der Neyman-Pearson-Ansatz diene im Wesentlichen dazu, die Güte von Tests zu erklären: wird für zwei Tests der gleiche α-Wert gewählt, habe aber einer einen kleineren β-Wert als der andere, so sei der erstere der bessere Test.

In diesem Sinne ist die NP-Theorie statistischer Tests eine Theorie optimaler Entscheidungen. Wald (1939, 1950) verallgemeinerte den NP-Ansatzes; manche Autoren sprechen deshalb auch vom Neyman-Pearson-Wald-Ansatz (Basu, 1975). Der Gegenstand der Statistik seien Fragen nach der Art der Entscheidungen unter Ungewißheit.

Beim Signifikanztest Fishers wird aber auf eine Hypothese $H = H_0$ fokussiert; H_0 steht etwa für die Hypothese, dass experimentelle Variablen keinen Einfluß auf eine bestimmte Meßgröße haben, oder dass kein Zusammenhang zwischen Variablen besteht, was den Ausdruck 'Nullhypothese' motiviert. Im Unterschied dazu gehen Neyman & Pearson davon aus, dass ein Test einer Hypothese nur dann sinnvoll ist, wenn gegen eine explizit formulierte Alternativhypothese getestet wird, und diese Alternativhypothese sollte nicht einfach nur, wie bei Fisher, die Negation von H_0 sein. Die Anwendung eines Tests nach Neyman & Pearson verlangt also die Spezifikation einer Alternativhypothese, was durch Annahme einer Effektgröße, die die Alternativhypothese spezifiziert, geschehen kann (Cohen (1969)). Beiden Ansätzen gemein ist die Ablehnung des Prinzips der Inversen Wahrscheinlichkeit (1), d.h.[4]

$$P(H|E) \propto P(E|H)P(H), \tag{10}$$

da die Notwendigkeit, die a-priori-Wahrscheinlichkeit $P(H)$ zu spezifizieren, eine subjektive Komponente in die Entscheidungen bringe. Die Theorie Neyman & Pearsons wird in Lehrbüchern zur Mathematischen Statistik als der kanonische Ansatz schlechthin dargestellt (Lehmann (1994), Kendall & Stuart (1973), Pruscha (2000)), während in Lehrbüchern für Anwender die Fishersche Theorie dominiert, oft mit Komponenten der Neyman-Pearsonschen Theorie versetzt, ohne dass dies deutlich gemacht wird (Huberty (1993)).

Bei wissenschaftlichen Hypothesen ist aber im Lichte der Diskussion über Verifikation und Falsifikation von Theorien und Hypothesen der Ausdruck 'Entscheidung' nicht allzu wörtlich zu nehmen. Daten, die bei Geltung einer bestimmten Hypothese H sehr unwahrscheinlich sind, implizieren ja nicht notwendig, dass H falsch ist, denn auch unwahrscheinliche Ereignisse (hier: Stichproben) treten gelegentlich auf. Daten, die bei Annahme von H unwahrscheinlich sind, können demnach allenfalls Skepsis bezüglich H erzeugen. In der wissenschaftlichen Praxis werden statistische Tests oft nur herangezogen, um das unklare Bild kontrastreicher zu machen; es geht es mehr um mögliche Schlußfolgerungen, die aus den Daten gezogen werden können. Man kann mit Cox (1958, p. 357) sagen,

[4] $\propto$ steht für "ist proportional zu". In (10) wird der Faktor $1/P(E)$ fortgelassen, um auf die inverse Relation von $P(H|E)$ und $P(E|H)$ zu fokussieren.

> "A statistical inference [... is] a statement about statistical populations made from given observations with measured uncertainty."
>
> ...
>
> "A scientific inference in the broader sense is usually concerned with arguing from descriptive facts about populations to some deeper understanding of the system under investigation."

Tukey (1960) hat die Frage nach der Evidenz ebenfalls in grundsätzlicher Form angesprochen: die Frage ist ja, ob Evidenz prinzipiell über Entscheidungen, oder nicht auch einfach über Schlußfolgerungen erklärt werden soll. Rozeboom (1960, p. 420) hat ebenfalls die Frage nach der Angemessenheit von Entscheidungen zwischen Hypothesen diskutiert:

> But the primary aim of a scientific experiment is not to precipitate decisions, but to make an appropriate adjustment in the degree to which one accepts, or believes, the hypothesis or hypotheses being tested."
>
> "... a hypothesis is not something, like a piece of pie offered for dessert, which can be accepted or rejected by a voluntary physical action. Acceptance or rejection of a hypothesis is a cognitive process, a *degree* of believing or disbelieving which, if rational, is no matter of choice but determined solely by how likely it is, given the evidence, that the hypothesis is true." (p. 423)

Die Frage ist also, wie sich die *evidential meaning* von Daten, also die Bedeutung von Daten für die Evaluation der Geltung von Hypothesen, charakterisieren läßt.

In einem dritten Ansatz wird der Versuch unternommen, den Unzulänglichkeiten des Fisherschen Signifikanztests einerseits und des Neyman-Pearsonschen Hypothesentests andererseits zu entkommen, aber das Prinzip, dass eine Hypothese nur gegen eine explizit formulierte Alternativhypothese getestet werden kann, beizubehalten. Es beruht auf dem *Likelihood-Prinzip*, demzufolge für die Bewertung einer Hypothese H nur die Likelihood $P(E|H)$ relevant ist (Savage (1961), zitiert nach Hacking (1965), p. 65). Hacking (1965), führt noch das schärfer formulierte *Law of Likelihood* ein:

> "If h and i are joint propositions and e is a joint proposition, and e includes both h and i, then e supports h better than i if the likelihood of h exceeds the of i. " (p. 59)
>
> Und weiter: "If h and i are simple joint propositions included in the joint proposition e, then e supports h better than i if the likelihood ratio of h to i exceeds i" (p. 70)

Insbesondere der Likelihood-Quotient $\lambda = P(E|H_1)/P(E|H_2)$ wird zur Basis den Evaluation von Hypothesen (Hacking (1965), Barnard (1967), Edwards (1982), Royall (1997)). So plausibel der Ansatz erscheint, – auch hier ergeben sich grundsätzliche Zweifel. Fasst man die gesammelten Zweifel an Fishers Signifikanztests, Neyman-Pearsonschen Hypothesentests und Hacking-Royallschen Likelihood-Tests zusammen, findet man sich der in der Position Earmans (1992):

> "I confess that I am an Baysian – at least I am on Mondays, Wednesdays, and Fridays. And when I am a Bayesian I am an imperialistic

> apostle in insisting that every sin and virtue in confirmation theory should be explained in Baysian terms. ... On Tuesdays, Thursdays, and Saturdays, however, I have my doubts not only about the imperialistic ambitions of Bayesianism but also about its viability as a basis of scientific inference. (On sundays I try not to think about the matter)." Earman (1992), p. 1 – 2.

Überblick: Es wird zunächst der Fishersche Signifikanztest formal eingeführt. Dann werden die erwähnten Implikationen dieses Tests betrachtet. Bei einigen Argumenten wird dabei vom Bayes-Ansatz Gebrauch gemacht; für Anti-Bayesianer werden diese Argumente vermutlich ein geringes Gewicht haben, aber sie sind dann eingeladen, die pro-Bayesschen Argumente zu widerlegen, der bloße Hinweis auf die Subjektivität der a-priori-Verteilungen wird als Argument nicht genügen.

Anschließend wird der Neyman-Pearson-Ansatz einschließlich der Theorie der Konfidenzintervalle vorgestellt; die Notwendigkeit Power-Betrachtungen im Sinne Cohens (1993) anstellen zu müssen, wird ebenfalls im Zusammenhang mit der Bayesschen Statistik betrachtet werden. Eine Darstellung der Likelihood-Inferenz im dritten Teil schließt sich an.

Im vierten Teil werden einige grundsätzliche Betrachtungen zur Induktion und zum Bayesschen Ansatz vorgestellt. Es wird nicht argumeniert werden, dass dieser Ansatz – und seine diversen Unteransätze – der einzig mögliche oder sinnvolle ist. Aber die Wahrscheinlichkeit, sich nicht in der oben genannten Earmanschen Zuständlichkeit zu befinden, ist kleiner als α.

2 Einige Prinzipien

2.1 'Evidential Meaning' und Birnbaums These

Es geht um *informative Schlußfolgerungen* (*informative inferences*, die aus Daten in Bezug auf eine Hypothese H gezogen werden können. Für diese Schlußfolgerungen hat Birnbaum (1962) das Symbol $(E, \mathbf{x})$ eingeführt: E ist die Untersuchung (das Experiment), $\mathbf{x}$ sind die Daten. $(E, \mathbf{x})$ ist ein spezifisches Ergebnis des Experiments E. θ ist ein Parameter oder auch Vektor von Parametern aus einem Parameterraum Θ. Für jedes $\theta \in \Theta$ existiert eine Wahrscheinlichkeitsverteilung oder - dichte $f(\mathbf{x}|\theta)$.

Birnbaum betrachtet zwei Aspekte der informativen Schlußfolgerung: (i) die mathematische Charakterisierung der statistischen Information (statistical evidence), und (ii) die Interpretation der 'evidence' in den Daten, wozu die Charakterisierung der Begriffe zählt, die herangezogen werden sollen, um die Bedeutung der Daten für die Hypothese zu erfassen.

Mit $Ev(E, \mathbf{x})$ bezeichnet Birnbaum die *evidential meaning*, also die Bedeutung der Information in den Daten, die sich aus einem Experiment ergeben. Liefern zwei Untersuchungen E und E' die gleiche Information bezüglich H, so kann man von *informativer Äquivalenz* (*evidential equivalence*) sprechen:

$$Ev(E, \mathbf{x}) \equiv Ev(E', \mathbf{y}). \tag{11}$$

2.1.1 Das Suffizienz-Prinzip

Vorausgesetzt wird der von Fisher (1922) eingeführte Begriff der suffizienten Statistik; er wird im Anhang definiert. Es sei $x_1, \ldots, x_n$ eine Stichprobe und $g(x_1, \ldots, x_n|\theta)$ die vom Parameter θ abhängende Stichprobenverteilung. Läßt sich g in der Form

$$g(x_1, \ldots, x_n|\theta) = p(r|\theta)b(x_1, \ldots, x_n) \tag{12}$$

schreiben, $p(r|\theta)$ die Randverteilung für $r(x_1, \ldots, x_n)$. r ist dann eine suffiziente Statistik; sie enthält alle Information in den Daten über θ. So kann u.U. das arithmetische Mittel $r(x_1, \ldots, x_n) = \bar{x}$ alle Information in den Daten $\mathbf{x}$ über den Erwartungswert μ einer zufälligen Veränderlichen enthalten.

Definition 2.1 *Es sei E eine Untersuchung mit den resultierenden Daten $\boldsymbol{x}$. $T(\boldsymbol{x})$ sei eine suffiziente Statistik. E' sei ein von E abgeleitetes Experiment, bei dem jede Stichprobe $\boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ nur durch die dazu korrespondierende Statistik $T(\boldsymbol{x})$ repräsentiert wird. Dann gilt, für jede Stichprobe $\boldsymbol{x}$,*

$$Ev(E, \boldsymbol{x}) \equiv Ev(E', T(\boldsymbol{x})). \tag{13}$$

Anmerkung: Das Suffizienz-Prinzip bedeutet, dass Daten, die von der suffizienten Statistik $T(\mathbf{x})$ unabhängig sind, irrelevant sind. □

2.1.2 Das Konditionalitäts-Prinzip

Zur Vorbereitung wird der Begriff der *Ancillary Statistic*[5] eingeführt. Dieser Begriff wurde von Fisher (1934) als *ad hockery* eingeführt, um mit einer "Merkwürdigkeit orthodoxer Verfahren" zurechtzukommen (Jaynes (2003, p. 253). Es wird der Fall betrachtet, dass ein Lokationsparameter θ für eine Stichprobenverteilung $p(\mathbf{x}|\theta I) = f(\mathbf{x} - \theta|I)$ geschätzt werden soll. Nun können zwei Datenmengen (Stichproben) die gleiche Schätzung für θ liefern, obwohl andere Statistiken wie der Range und höhere Momente verschieden sind, so dass die Schlußfolgerungen bezüglich der Genauigkeit der Schätzungen verschieden sind. Fishers Idee war nun, eine Statistik $z(\mathbf{x})$ einzuführen, die nicht vom Parameter θ abhängt, derart, dass sie Information über die Daten liefert, die *nicht* von θ abhängt; $z(\mathbf{x})$ ist die 'ancillary statistic'. Jaynes (2003) liefert eine Interpretation der Hilfsstatistik, indem er sie in einem Bayesschen Rahmen interpretiert; darauf soll an dieser Stelle nicht eingegangen werden.

Weiter wird der Begriff der *Mischung von Experimenten* benötigt. Ein Experiment E ist eine Mischung von Experimenten (oder einfach eine Mischung) mit Komponenten E_h, wenn es mathematisch äquivalent zu einem Zweistufenexperiment der Form

(a) Gegeben eine zufällige Veränderliche H mit der Verteilungsfunktion G; ein Wert h dieser Variablen wird bestimmt,

(b) das Experiment E_h wird durchgeführt, mit dem Resultat x_h.

Jedes Resultat E ist damit äquivalent einem Paar (E_h, x_h), mit dem gleichen Parameterraum.

[5] Hilfsstatistik. Der Ausdruck 'Anzillarstatistik' scheint im Deutschen nicht üblich zu sein.

Definition 2.2 *Ein Experiment E sei mathematisch äquivalent einer Mischung G von Komponenten $\{E_h\}$ mit den möglichen Resultaten (E_h, x_h). Es gelte*

$$Ev(E, (E_h, x_h)) = Ev(E_h, x_h), \tag{14}$$

d.h. die evidential meaning *eines Ergebnisses (E_h, x_h) eines Experiments E mit einer Mischungsstruktur ist die gleiche wie die evidential meaning des Resultats x_h des Experiments E_h, wobei die Gesamtstruktur des ursprünglichen Experiments E vernachlässigt wird. (14) heißt dann das* Konditionalitätsprinzip.

Die Statistik h ist offenbar eine Hilfsstatistik im oben eingeführten Sinne. Das Konditionalitätsprinzip besagt, salopp ausgedrückt, die Irrelevanz von Komponenten des Experiments, die nicht explizit durchgeführt wurden. Eine weitere, ausführliche Diskussion findet man in Cox (1958).

Beispiel 2.1 Der Parameter θ möge in einem von zwei möglichen Experimenten E_1 oder E_2 schätzbar sein. Um zu entscheiden, welches Experiment durchgeführt werden soll, wird eine Münze geworfen. Die Schlußfolgerungen, die über θ gezogen werden, sollten nur von dem tatsächlich durchgeführten Experiment abhängen, nicht von dem nicht durchgeführten Experiment. □

Kritik: die Frage ist, wie ein Experiment unter den möglichen ausgesucht wird. In Beispiel 2.1 wurde gewürfelt, – aber das ist nur eine Möglichkeit, das Experiment auszuwählen. Eine andere Möglichkeit ist, das Experiment nach bestimmten, bekannten Kriterien auszusuchen. So könnten bestimmte Experimente besonders geeignet sein, kleine Werte von θ zu bestimmen, während andere Experimente eher für große Werte von θ geeignet sind (Jaynes 2003; 251).

Birnbaum beweist nun die folgende Aussage:

Satz 2.1 Satz von Birnbaum *Das Likelihood-Prinzip impliziert sowohl das Suffizienz- wie das Konditionalitätsprinzip, und umgekehrt implizieren das Suffizienz- und das Konditionalitätsprinzip das Likelihood-Prinzip.*

Beweis: Birnbaum (1962), p. 284.

Birnbaum führt nun aus, dass das Suffizienz- und das Konditionalitäts-Prinzip – und damit das Likelihood-Prinzip – eine mathematische Charakterisierung der statistischen Evidenz liefern.

Interpretationen von Likelihood-Funktionen: Die statistische Evidenz von $E, \mathbf{x})$ werde in Form einer Likelihood-Funktion repräsentiert, $L(\theta) = cf(\mathbf{x}, \theta)$, c eine positive Konstante, – was sind die qualitativen und quantitativen Eigenschaften der statistischen Evidenz, die durch $L(\theta)$ abgebildet werden? Nach Fisher (1956)[6] sind relative Likelihoods alternativer Werte von Parametern natürliche Präferenzordnungen von Möglichkeiten, die zeigen, welche Werte der Parameter unplausibel sind.

[6] Fisher, R.A.: Statistical methods and scientific inferences. Edinburgh 1956

2.1.3 Das Likelihood-Prinzip

Gesucht wird ein Ausdruck, der in irgendeiner Weise die Bedeutung von Daten $\mathbf{x} = (x_1, \ldots, x_n)$ für die Gültigkeit einer Hypothese – die 'evidential meaning' der Daten – ausdrückt. Für Fisher erschien die Likelihood der Daten, gegeben eine Hypothese, die 'evidence' der Daten in Bezug auf diese Hypothese auszudrücken. Das im Folgenden definierte Likelihood-Prinzip kann als eine Spezifikation dieses Gedankens gesehen werden.

Definition 2.3 *Es bezeichne $\boldsymbol{X} = (X_1, \ldots, X_n)$ eine Stichprobe vom Umfang n, und $\boldsymbol{x} = (x_1, \ldots, x_n)$ werde beobachtet. Weiter sei $f(X|\theta)$ die gemeinsame Dichte der X_j, $j = 1, \ldots, n$, oder die gemeinsame Wahrscheinlichkeitsfunktion, wobei θ ein Parameter der Dichte bzw. Wahrscheinlichkeitsfunktion ist. Dann ist*

$$L(\theta|\boldsymbol{x}) = f(\boldsymbol{x}|\theta) \tag{15}$$

die Likelihood-Funktion *für $\boldsymbol{x}$.*

$X = (X_1, \ldots, X_n)$ sei diskret, so dass $f(\mathbf{x}|\theta)$ eine Wahrscheinlichkeitsfunktion ist. Gilt $L(\theta_1|\mathbf{x}) > L(\theta_s|\mathbf{x})$, d.h. $f(\mathbf{x}|\theta_1) > f(\mathbf{x}|\theta_2)$, so ist offenbar $\mathbf{x}$ wahrscheinlicher, wenn der Parameter den Wert θ_1 hat, als wenn $\theta = \theta_2$. Generell bedeutet dann $L(\theta|\mathbf{x}) = f(\mathbf{x}|\theta)$, dass L bzw. f die Wahrscheinlichkeit von $\mathbf{x}$ als Funktion von θ. Ist f eine Dichtefunktion, so ist $f(\mathbf{x}|\theta)$ bekanntlich keine Wahrscheinlichkeit. Es läßt sich aber eine Beziehung zu Wahrscheinlichkeiten herstellen. Dazu werde angenommen, dass $f(\mathbf{x}|\theta)$ stetig in $\mathbf{x}$ ist. Man kann dann die Wahrscheinlichkeit, dass $\mathbf{X}$ im Intervall $(\mathbf{x} - \varepsilon, \mathbf{x} + \varepsilon)$ liegt betrachten. Es ist

$$P(\mathbf{x} - \varepsilon < \mathbf{X} < \mathbf{x} + \varepsilon|\theta) = F(\mathbf{x} + \varepsilon) - F(\mathbf{x} - \varepsilon) = \int_{\mathbf{X}-\varepsilon}^{\mathbf{X}+\varepsilon} f(\mathbf{X}|\theta) d\mathbf{X}.$$

Nun ist

$$\frac{F(\mathbf{x} + \varepsilon) - F(\mathbf{x} - \varepsilon)}{2\varepsilon} \approx f(\mathbf{x}|\theta)$$

für hinreichend kleines ε, d.h.

$$F(\mathbf{x} + \varepsilon) - F(\mathbf{x} - \varepsilon) \approx 2\varepsilon f(\mathbf{x}|\theta),$$

so dass

$$\frac{P(\mathbf{x} - \varepsilon < \mathbf{X} < \mathbf{x} + \varepsilon|\theta_1)}{P(\mathbf{x} - \varepsilon < \mathbf{X} < \mathbf{x} + \varepsilon|\theta_2)} \approx \frac{2\varepsilon L(\theta_1|\theta_2)}{2\varepsilon L(\theta_2|\mathbf{x})} = \frac{L(\theta_1|\mathbf{x})}{L(\theta_2|\mathbf{x})}. \tag{16}$$

Der Likelihood-Quotient $L(\theta_1|\mathbf{x})/L(\theta_2|\mathbf{x})$ ist also approximativ gleich dem Quotienten der Wahrscheinlichkeiten, dass $\mathbf{X}$ in einem Intervall $\mathbf{x} \pm \varepsilon$ liegt, bedingt auf $\theta = \theta_1$ und $\theta = \theta_2$.

Definition 2.4 *Es seien $\boldsymbol{x}$ und $\boldsymbol{y}$ zwei Stichproben und es gelte*

$$L(\theta|\boldsymbol{x}) = C(\boldsymbol{x}, \boldsymbol{y}) L(\theta|\boldsymbol{y}), \tag{17}$$

wobei $C(\boldsymbol{x}, \boldsymbol{y})$ ein nur von $\boldsymbol{x}$ und $\boldsymbol{y}$, nicht aber von θ abhängender Faktor sei, so dass die Folgerungen aus $\boldsymbol{x}$ und $\boldsymbol{y}$ bezüglich θ dieselben sind. Dann repräsentiert (17) das Likelihood-Prinzip.

Anmerkungen:

1. Das Likelihood-Prinzip besagt, dass die Bedeutung der Daten für eine Hypothese H vollständig durch die Likelihood der Daten, gegeben H, enthalten ist. Die Art der Untersuchung, aus der die Daten hervorgegangen sind, spielt dabei keine Rolle: liefert etwa das Experiment I die Daten $\mathbf{x}$ und das Experiment II die Daten $\mathbf{y}$, und erfüllen die jeweiligen Likelihood-Funktionen die Bedingungen (17), also $C(\mathbf{x}, \mathbf{y})$ unabhängig von θ, so ist alle Information in der Likelihood-Funktion $L(\theta|\mathbf{x})$ bzw. $L(\theta|\mathbf{y})$ enthalten; die Strukturen der beiden Experimente können verschieden sein: so kann in Experiment I die Anzahl n der Versuchsdurchgänge vor Beginn des Experiments festgelegt worden sein und man bestimmt die Anzahl k von interessierenden Ereignissen, während in Experiment II die Untersuchung beendet wird, wenn eine vorgegebene Anzahl k von interessierenden Ereignissen eingetreten ist.

2. Nach der Bayes-Beziehung

$$P(H|D) = P(D|H)P(H)$$

   hängt $P(H|D)$ von den Daten nur über die Likelihood $P(D|H)$ ab; vom Bayesschen Standpunkt aus steckt deshalb alle Information über H in der Likelihood und die Definition eines *Likelihood-Prinzips* ist insofern überflüssig, weil selbstverständlich. Wie Jaynes (2003), p. 250, ausführt, ist diese Selbstverständlichkeit aber für Nicht-Bayesianer, für die eine Wahrscheinlichkeitsfunktion eine Beschreibung physikalischer ("objektiver") Sachverhalte ist, nicht gegeben, so dass das Likelihood-Prinzip Gegenstand einer länglichen Kontroverse ist.

3. Eine Betrachtung von Barnard (1947) liefert eine weitere Motivation für die Einführung des Likelihood-Prinzips. Die allgemeine Idee ist, dass irrelevante Daten nicht in die Schlußfolgerungen über die Hypothesen eingehen sollten. Dazu sei $T = \{K, Z\}$ eine zufällige Veränderliche, die das Resultat eines Münzwurfs anzeige: für $T = K$ liegt der Kopf oben, für $T = Z$ liegt die Zahl oben. Die Stichprobenwahrscheinlichkeit für alle Daten ist dann

$$P(D\,T|\theta) = P(D|\theta)P(T),$$

   da D und T stochastisch unabhängig sind. Dies bedeutet, dass der Münzwurf uns nichts weiter über den Parameter θ vermittelt, als was nicht schon in der Likelihood $P(D|H)$ steckte, d.h. die Schlußfolgerungen aus $P(D\,T|\theta)$ sind die gleichen wie die aus der Likelihood $P(D|H)$; der Wert der zufälligen Veränderlichen ist irrelevant für die Interpretation. Verallgemeinert bedeutet dies eben, dass konstante Faktoren in der Likelihood-Funktion irrelevant für Schlußfolgerungen sind, und Schlußfolgerungen über θ aus den Werten des Likelihood-Quotienten

$$\frac{L_1}{L_2} = \frac{P(D\,T|\theta_1)}{P(D\,T|\theta_2)} = \frac{P(D|\theta_1)}{P(D|\theta_2)}$$

   gezogen werden können.

Beispiel 2.2 Es werden zwei Folgen von Bernoulli-Experimenten durchgeführt, wobei für die erste Folge die Anzahl n der Versuche vor ihrer Durchführung festgelegt wird, während bei der zweiten Folge gefordert wird, dass die Versuche so lange fortgeführt werden, bis genau k "Erfolge" eingetreten sind. Der Parameter θ

ist in beiden Fällen die Wahrscheinlichkeit eines Erfolgs in einem Versuch, und der Wert von θ ist in beiden Versuchsreihen identisch. Es zeigt sich, dass in der ersten Versuchsreihe ebenfalls k Erfolge beobachtet wurden; in der zweiten Versuchsreihe wurden $m > n$ Versuche benötigt, um die k Erfolge zu erzielen.

In der ersten Versuchsreihe mit festgelegter Anzahl n von Versuchen ist die Anzahl der Erfolge binomialverteilt, so dass insbesondere

$$P(X = k|\theta, n) = \binom{n}{k}\theta^k(1-\theta)^{n-k}.$$

In der zweiten Versuchreihe ist die Anzahl der Versuche negativ binomialverteilt: hat man insbesondere insgesamt j Mißerfolge *vor* dem k-ten Erfolg, so verteilen sich die $k-1$ Erfolge gemäß

$$\binom{k+j-1}{k-1}\theta^{k-1}(1-\theta)^j$$

auf die insgesamt $k+j-1$ Versuche; θ^{k-1} ist proportional der Wahrscheinlichkeit von $k-1$ Erfolgen, $(1-\theta)^j$ ist proportional der Wahrscheinlichkeit von j Mißerfolgen. Dass im $(k+j)$-ten Versuch ein Erfolg beobachtet wird, hat die Wahrscheinlichkeit θ und ist unabhängig von den vorangegangenen Versuchen, also ist die Wahrscheinlichkeit, mit dem $m = (k+j)$-ten Versuch gerade k Erfolge zu haben,

$$P(Y = m|\theta, j) = \binom{k+j-1}{k-1}\theta^{k-1}(1-\theta)^j\theta = \binom{m-1}{k-1}\theta^k(1-\theta)^j.$$

Die Likelihood-Funktionen sind

$$L_1(\theta|k,n) = P(X = k|\theta, n), \quad L_2(\theta|k,m) = P(Y = m|\theta, j).$$

Man kann nun nach den Bedingungen fragen, unter denen $L_2 = CL_1$ gilt mit C unabhängig von θ:

$$\binom{n}{k}\theta^k(1-\theta)^{n-k} = C(\mathbf{x},\mathbf{y})\binom{k+j-1}{k-1}\theta^k(1-\theta)^j,$$

und alle Terme, die θ enthalten, kürzen sich heraus, wenn $j = n-k$. Hat man etwa $n = 12$, $k = 3$, so ergibt sich C aus

$$\binom{12}{3} = C\binom{11}{2}.$$

□

Beispiel 2.3 Es seien $\mathbf{x}$ und $\mathbf{y}$ zwei Stichproben aus einer normalverteilten Population. Die Likelihood-Funktion für $\mathbf{x}$ ist, mit $\theta = \mu$,

$$f(\mathbf{x}|\mu) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right].$$

Es ist aber

$$x_i - \mu = (x_i - \bar{x}) + (\bar{x} - \mu),$$

$\bar{x}$ das arithmetische Mittel der x_i, d.h. es ist

$$(x_i - \mu)^2 = (x_i - \bar{x})^2 + (\bar{x} - \mu)^2 + 2(x_i - \bar{x})(\bar{x} - \mu).$$

Summation über die i liefert

$$\sum_{i=1}^{n}(x_i - \mu)^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 + n(\bar{x} - \mu)^2, \tag{18}$$

da ja $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$. Mithin folgen die Beziehungen

$$L(\mu|\mathbf{x}) = f(\mathbf{x}|\mu) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \bar{x})^2 + \sum_{i=1}^{n}(\bar{x} - \mu)^2\right] \tag{19}$$

$$L(\mu|\mathbf{y}) = f(\mathbf{y}|\mu) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \bar{y})^2 + \sum_{i=1}^{n}(\bar{y} - \mu)^2\right]. \tag{20}$$

Das Likelihood-Prinzip ist erfüllt, wenn ein $C(\mathbf{x}, \mathbf{y})$ existiert derart, dass (17) erfüllt ist, wobei $C(\mathbf{x}, \mathbf{y})$ nicht von μ abhängen darf. Die Definition (17) bedeutet, dass

$$C(\mathbf{x}, \mathbf{y}) = L(\mu|\mathbf{x})/L(\mu|\mathbf{y}),$$

und (19) und (20) implizieren

$$\begin{aligned} C(\mathbf{x}, \mathbf{y}) &= \frac{L(\mu|\mathbf{x})}{L(\mu|\mathbf{y})} = \\ &= \exp\left[-\frac{1}{\sigma^2}\left(\sum_{i=1}^{n}(x_i - \bar{x})^2 - n(\bar{x} - \mu)^2 - \right.\right. \\ &\qquad \left.\left. - \sum_{j=1}^{n}(y_j - \bar{y})^2 + n(\bar{y} - \mu)^2\right)\right]. \end{aligned} \tag{21}$$

$C(\mathbf{x}, \mathbf{y})$ hängt offenbar genau dann nicht von μ ab, wenn

$$-n(\bar{x} - \mu)^2 + n(\bar{y} - \mu)^2 = 0,$$

d.h. wenn $\bar{x} = \bar{y}$. In diesem Fall ist

$$\begin{aligned} f(\mathbf{x}|\mu) &= \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}(x_i - \bar{x})^2 + n(\bar{x} - \mu)^2\right)\right] \\ &= \frac{C(\mathbf{x}, \mathbf{y})}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}(y_i - \bar{y})^2 + n(\bar{y} - \mu)^2\right)\right] = C(\mathbf{x}, \mathbf{y}) f(\mathbf{y}|\mu), \end{aligned}$$

d.h. für $\bar{x} = \bar{y}$ ist das Likelihood-Prinzip erfüllt, die Folgerungen aus $\mathbf{x}$ und $\mathbf{y}$ bezüglich des unbekannten Parameters μ sind in diesem Fall identisch. □

Natürlich sind die Aussagen Birnbaums nicht ganz unwidersprochen geblieben. Fraser (1963) argumentiert, dass das Likelihood-Prinzip, wie es von Birnbaum formuliert wurde, für bestimmte Beziehungen zwischen der gemessenen Variablen und dem zu schätzenden Parameter zu stark sein kann, und Durbin (1970) zeigte, dass die Birnbaumschen Folgerungen von der Formulierung des Konditionalitätsprinzips abhängen.

Tabelle 1: Die Entwicklung des statistischen Testens vor R. A. Fisher (nach Huberty (1993)

Jahr	Person	Thema
1710	John Arbuthnot (1667 - 1735)	Verhältnis der Anzahl von Jungen- und Mädchengeburten
1767	John Michell (1724 - 1793)	zufällige Verteilung von Sternen und Sternhaufen
1823	Pierre Simon Laplace (1749 - 1827)	Mondphase und barometrische Veränderungen
1900	Karl Pearson (1857 - 1936)	Güte der Anpassung
1908	William Sealy Gosset (1876 - 1937)	Prüfung von Mittelwertsdifferenzen

3 Signifikanztests und Evidenz

3.1 Das Prinzip des Fisherschen Signifikanztests

Das Prinzip des Fisherschen Signifikanztests geht nicht auf Fisher zurück, wohl aber der Ausdruck *test of significance* (Fisher (1925), p. 43). Die Tabelle 1 auf Seite 19 liefert eine kleine Liste von Vorläufern, die das gleiche, zugrundeliegende Prinzip verwendet haben, das von Royall (1997) so genannte:

> *Law of Improbability* If hypothesis A implies that the probability that a random variable X takes on the value x is quite small, say, $p_A(x)$, then the observation $X = x$ is evidence against A, and the smaller $p_A(x)$, the stronger that evidence.

John Arbuthnot gilt als der Erste, der eine dem p-Wert eigene Logik verwendete. Arbuthnot hatte aus Londoner Unterlagen für 82 Jahre die Anzahl der Jungengeburten größer als die Anzahl der Mädchengeburten war. Er nahm an, dass die Wahrscheinlichkeit einer Jungen- bzw. Mädchengeburt gleich groß sei, nämlich $p = 1/2$, weiter an, dass die Geburten alle statistisch unabhängig voneinander sind und errechnete eine Wahrscheinlichkeit von $(1/2)^{82} = 2^{-82}$ für seinen Befund. Er hat gewissermaßen die Likelihood der Daten unter der Bedingung, dass seine Annahmen korrekt sind, berechnet. Die Wahrscheinlichkeit der Daten unter der Hypothese, dass Mädchen und Jungen mit gleicher Häufigkeit geboren werden, erschien ihm so gering, dass er die Hypothese verwarf und auf eine göttliche Vorsehung (Divine Providence) schloß. Die Schlußfolgerung mag man anzweifeln, aber von der Unwahrscheinlichkeit der Daten, gegeben eine bestimmte Hypothese, auf die Unwahrscheinlichkeit der Hypothese zu schließen, ist eine auch heut noch geübte Praxis; man erinnere sich an das in der Einführung erwähnte Royallsche *Law of Improbability*.

John Michell[7] folgte ebenfalls dieser Logik. Er berechnete die Wahrscheinlichkeit für die Möglichkeit, dass bestimmte Gruppen von Sternen wie die Pleiaden lediglich eine zufällige Anhäufung von Sternen seien, zu 1 : 496000. Diese Wahrscheinlichkeit sei zu klein, um an die Zufälligkeit der Anhäufung zu glauben, also

[7] 1724 – 1793; Geologe und Astronom.

könne man davon ausgehen, dass irgendwelchen systematischen Effekte wirksam seien. Hardin (1966, p. 36) liefert eine explizite Darstellung der Michellschen Argumentation.

Das Schema des Signifikanztests: Es sei $\mathbf{X}$ irgendeine Strichprobe und $\mathbf{x}$ die spezielle Stichprobe, die sich bei einer Untersuchung ergeben hat. θ sei ein unbekannter Parameter, $\theta \in \Theta \subseteq \mathbb{R}$, Θ der Parameterraum. $f(\mathbf{X}|\theta)$ sein die Dichtefunktion (oder Wahrscheinlichkeitsfunktion) für $\mathbf{X}$. Es soll die Hypothese $\theta = \theta_0$ getestet werden. Es wird eine Teststatistik $T(\mathbf{X})$ definiert und die Wahrscheinlichkeit

$$p = P(T(\mathbf{X}) > T(\mathbf{x})|H_0) = P_{\theta_0}(T(\mathbf{X}) > T(\mathbf{x})) \tag{22}$$

betrachtet. p heißt das *Signifikanzniveau*. Ist p (der p-Wert) klein, so ist der beobachtete Wert $T(\mathbf{x})$ "groß" relativ zur Gesamtmenge der Werte von T, wenn H_0 gilt. In diesem Sinne kann man sagen, dass die Daten $\mathbf{x}$ unter der Bedingung, dass H_0 gilt, eher unwahrscheinlich sind; je kleiner der Wert von p, desto geringer ist die Wahrscheinlichkeit der Daten unter H_0. Einer oft angewandten Regel entsprechend wird demnach H_0 "verworfen" oder "zurückgewiesen", wenn etwa $p \leq .05$; das Ergebnis der Untersuchung mit dem Resultat $\mathbf{x}$ ist dann *signifikant*.

Kommentar: Ein erster Kommentar ergibt sich aus der Charakterisierung (22) des Signifikanztests: es werden nicht nur die tatsächlich beobachteten Daten $\mathbf{x}$ bzw. die Statistik $T(\mathbf{x})$ zur Bewertung der Hypothese(n) betrachtet, sondern alle Werte von T, die größer sind als $T(\mathbf{x})$. Es ist nicht klar, warum die Wahrscheinlichkeit des Ereignisses $\{T > T(\mathbf{x})|H_0\}$ eine kritische Größe für die Bewertung von H_0 sein soll. Wie Hacking (1965, p. 82) anmerkt, hat Fisher hierzu nie eine Begründung gegeben. Für viele Anfänger des Studiums statistischer Verfahren ergibt sich hier eine erste Schwierigkeit, denn es ist nicht unmittelbar einsichtig, warum H_0 nach dem Kriterium "$T(\mathbf{x})$ *oder größer*" beurteilt werden soll. Jeffreys (1939/1961, p. 385) hat den Sachverhalt auf knappe Weise auf den Punkt gebracht:

> "What the use of p implies, is that a hypothesis that may be true may be rejected because it has not predicted observable results that have not occurred. This seems a remarkable procedure."

Die bei Anwendungen oft beklagte Vermischung des Fisherschen mit dem Neyman-Pearsonschen Ansatz mag hier ihre Wurzeln haben. Bei Neyman-Pearson wird ein α-Fehler festgelegt, der einem Bereich R correspondiert derart, dass unter H_0

$$P(T(\mathbf{x}) \in R) = \alpha$$

gilt. Für $\mathbf{x}_0 = \min R$ ist dann

$$P(T(\mathbf{x}) \in R) = \int_R f(\mathbf{x}|H_0)\, d\mathbf{x} = \int_{\mathbf{x}_0}^{\infty} f(\mathbf{x}|H_0) d\mathbf{x} = \alpha.$$

Es sei nun $\mathbf{x}$ der beobachtete Wert, und $R_{\mathbf{x}} = \{\xi|\xi \geq \mathbf{x}\}$, und

$$P(T \geq \mathbf{x}) = p = \int_{\mathbf{x}}^{\infty} f(\xi|H_0) d\xi.$$

Ist $\mathbf{x} \in R$ so ist $p \leq \alpha$. Man interpretiert Fishers Ansatz also, indem man davon ausgeht, dass der kleinste Wert einer Menge – hier also R – eben ein Element dieser

Menge ist, der beobachtete Wert $\mathbf{x}$ ist Element der Menge $R_{\mathbf{X}} = \{\xi | \xi \geq \mathbf{x}\}$. Gilt $P(T \in R_{\mathbf{X}}) = p \leq \alpha$, so spricht man eben von einem "signifikanten" Wert. Dies ist eine mögliche und sicherlich wohlwollende Interpretation. Man kann den gleichen Sachverhalt etwas anders pointieren:

> "... it is worthwile to discuss the main reason for the substantial difference between the magnitude of p and the magnitude of the evidence against H_0. The problem is essentially one of conditioning. The actual vector of observations is x, and $P(H_0|x)$ and l_x (Anm.: gemeint ist der Likelihood-Quotient) depend only on the evidence from the actual data observed: To calculate a p-value, one effectively replaces x by the 'knowledge" that X is in $A = \{y : T(y) \geq T(x)\}$ and then calculates $p = P_{\theta=\theta_0}(A)$. Although the use of frequentist measures can cause problems, the main culprit here is the replacing of x itself by A."
>
> Berger & Sellke (1987), p. 114. (T ist eine Teststatistik, die anhand der Daten x berechnet wird.)

Ein einfaches Beispiel illustriert den Signifikanztest.

Beispiel 3.1 Gossets (STUDENTs) t-Test Es seien die voneinander unabhängigen, normalverteilten Messungen x_{ij}, $i = 1, 2$ und $j = 1, \ldots, n$ unter zwei Bedingungen B_1 und B_2 gemacht worden. Die Erwartungswerte für die beiden Messreiehen seien μ_1 und μ_2, und der Einfachheit halber gelte für die Varianzen $\sigma_1^2 = \sigma_s^2$. Es sei $\mathbf{x} = (x_{11} - x_{21}, \ldots, x_{1n} - x_{2n})$,

$$T(\mathbf{X}) = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{s/\sqrt{n}}, \quad df = n - 1 \tag{23}$$

s_d eine Schätzung für $\sigma = \sqrt{\sigma_1^2 + \sigma_2^2} = \sigma\sqrt{2}$. Weiter sei $\theta = \mu_1 - \mu_2$, H_0: $\theta = \theta_0 = 0$, so dass

$$T(\mathbf{x}) = \frac{\bar{x}_1 - \bar{x}_2}{s_d}\sqrt{n}.$$

Ist $T(\mathbf{X}) > T(\mathbf{x})$, so ist die Differenz $\bar{x}_1 - \bar{x}_2$ groß im Verhältnis zu den Differenzen, die "normalerweise" unter H_0 auftreten, und erscheinen deshalb als unwahrscheinlich unter H_0. □

Das Beispiel liefert auch gleich eine Illustration der Tatsache, dass jede noch so kleine Abweichung von H_0 als signifikant bewertet werden kann, wenn nur die Stichprobe hinreichend groß ist. Für $n > 30$ kann man die Schätzung s für σ^2 ersetzen und von der t-Verteilung zur Normalverteilung $N(0, 1)$ übergehen. Dann hat man

$$t = \frac{\Delta\bar{x}\sqrt{n}}{\sigma} = z_n, \tag{24}$$

und für jede Differenz $\Delta\bar{x} > 0$ existiert ein n derart, dass $z_n > z_c$, mit etwa $z_c = 1.96$, d.h. wählt man den Stichprobenumfang hinreichend groß, so geht die Wahrscheinlichkeit gegen 1, dass die Differenz $\Delta\bar{x}$ als 'signifikant' bewertet wird. Man *kann* diesen Sachverhalt als einen problematischen Aspekt des Signifikanztests ansehen, denn eine Mittelwertdifferenz $0 < \Delta\bar{x} < \varepsilon$ für beliebig kleines ε kann ja ohne jede praktische Relevanz sein; wenn sich zwei Therapien nur um ein nahzu beliebig kleines ε unterscheiden, eine der beiden Therapien aber wesentlich teurer ist als die andere, wird man sich vermutlich für die billigere entscheiden und 'signifikant' nicht mit 'relevant' gleichsetzen.

3.2 Das α-Postulat und die Evidenz

Eine erste Frage, die sich stellt, ist die nach der Evidenz gegen H_0, die vom p-Wert repräsentiert wird, und der Abhängigkeit des p-Wertes und damit der Evidenz vom Stichprobenumfang.

Cornfield (1966) diskutierte die Frage, ob die Schlußfolgerungen, die aus einer Datenmenge gezogen werden, nur von den Daten selbst oder ob sie auch von der Stop-Regel, von der die Beendigung der Datenerhebung bestimmt wird, abhängt. Dabei betrachtet er die nach seiner Ansicht von vielen (Bio-)Statistikern adoptierte Annahme, dass das Signifikanzniveau α ein Index oder Maß für die Evidenz ist in dem Sinne, dass Daten, die unter H_0 hinreichend selten vorkommen, Evidenz gegen H_0 seien. Das grundlegende Postulat, dass jeder sequentiellen Analyse unterliege, sei demnach "All hypotheses rejected at the same critical level have equal amounts of evidence against them" (Cornfield, 1966; p. 19). Cornfield nennt diese Annahme das *α-Postulat* und argumentiert, dass das Postulat nicht gilt, das Signifikanzniveau mithin kein gutes Maß für die Evidenz gegen H_0 sei, und statt dessen – dem Likelihood-Prinzip zufolge – die Likelihood-Funktion ein geeignetes Maß für die Evidenz gegen H_0 sei.

Wenn es so ist, dass ein größerer Stichprobenumfang n einen existierenden Effekt leichter entdeckbar macht, könnte man argumentieren, dass ein kleiner p-Wert bei großem n eine geringere Evidenz gegen H_0 bedeutet als ein ebenso kleiner p-Wert bei kleinem n. Andererseits könnte es sein, dass der kleine p-Wert bei großem n eine solidere Evidenz gegen H_0 bedeutet, denn schließlich haben die Statistiken, die zur Abschätzung der Evidenz dienen, eine kleinere Varianz bei großem n. Tatsächlich sind beide Standpunkte vertreten worden:

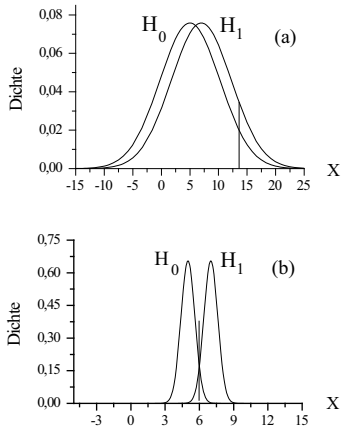
> "... the interpretation to be placed 'significant at 5%' depends on the sample size: it is more indicative of the falsity of the null hypothesis with a small sample size than with a large one. (Lindley & Scott 1984, p. 3)"
>
> "A given p-value in a large trial is usually stronger evidence that the treatments really differ than the same p-value in a small trial of the same treatments would be." (Peto et al. 1976, p. 593)

Royall (1986) argumentierte, dass beide Standpunkte gelten können. Der Kern dieser anscheinenden Widersprüchlichkeit liegt in der Interpretation des Begriffes der Evidenz. Peto et al.s Argument werde zuerst vorgestellt. In Abb. 12 (a) werden zwei Verteilungen f_0 und f_1 einer Statistik $T(\mathbf{x})$ für "kleines" n gezeigt; f_0 korrespondiert zu H_0, f_1 zu H_1. Abb. 12, (b) zeigt die Dichten f_0' und f_1' für "großes" n. Die Differenz $\vartheta = \mu_2 - \mu_1$ der Erwartungswerte ist in beiden Fällen die gleiche. Der kritische Wert T_c für $P(T > T_c) = .05$ ist in (a) $T_c = 13.63$, in (b) $T_c' = 6.0$. In beiden Fällen ist der α-Fehler gleich groß, also

$$\alpha = P(T > T_c|H_0) = P(T' > T_c'|H_0) = .05.$$

Peto et al. hatten angenommen, dass unter 100 Untersuchungen 20 sind, bei denen tatsächlich ein von Null verschiedener Effekt existiert (kleines n, der Fall (a)), und analog dazu unter 1000 Untersuchungen 200 mit einem tatsächlich von Null verschiedenen Effekt (großes n,Fall (b)). Das Verhältnis der Anzahl von Untersuchungen mit korrekter H_0 zur Anzahl mit korrekter H_1 betrage also $100/20 = 1000/200 = 5/1 = 5$.

Abbildung 1: Verteilungen einer Statistik für verschiedenes n bei identischer Differenz der Erwartungswerte; Die Erwartungswerte sind in beiden Fällen ((a) und (b)) gleich: $\mu_1 = 5$, $\mu_2 = 7$. In (a): $P(T > T_c|H_1) = .25$, d.h. die Streuung ist $s = 5.26$, $T_c = 13.63$. In (b) ist $P(T > T_c|H_1) = .95$, mit $s = .61$ und $T_c = 6.0$.



Royall (1986) betrachtet nun die Wahrscheinlichkeiten $p(H_0|\mathbf{s})$ bzw. $p(H_1|\mathbf{s})$ der Hypothesen, wenn ein signifikantes Ergebnis $\mathbf{s}$ vorliegt, d.h. wenn der p-Wert die Bedingung

$$\mathbf{s}: \ p \leq \alpha = .05 \tag{25}$$

erfüllt. Dann hat man

$$P(H_0|\mathbf{s}) = P(\mathbf{s}|H_0)P(H_0) \tag{26}$$
$$P(H_1|\mathbf{s}) = P(\mathbf{s}|H_1)P(H_1) \tag{27}$$

Dividiert man die erste Gleichung durch die zweite, so erhält man

$$\frac{P(H_0|\mathbf{s})}{P(H_1|\mathbf{s})} = \frac{P(\mathbf{s}|H_0)}{P(\mathbf{s}|H_1)} \frac{P(H_0)}{P(H_1)} \tag{28}$$

Nach den Vorgaben Peto et al.s ist aber $P(\mathbf{s}|H_0) = P(\mathbf{s}|H_1) = .05$ und bei kleinem n soll $P(\mathbf{s}|H_1) = .25$ gelten, bei großem n dagegen $P(\mathbf{s}|H_1) = .95$. Die Annahmen über die relativen Häufigkeiten von Untersuchungen mit tatsächlich existierendem Effekt implizieren $P(H_0)/P(H_1) = 5/1$. Dementsprechend erhält man

$$\frac{P(H_0|\mathbf{s})}{P(H_1|\mathbf{s})} = \frac{P(\mathbf{s}|H_0)}{P(\mathbf{s}|H_1)} \frac{P(H_0)}{P(H_1)} = \begin{cases} .05 \times 5/.25 = 1, & n \text{ klein} \\ .05 \times 5/.95 = 5/19, & n \text{ groß} \end{cases} \tag{29}$$

Für einen kleinen Wert von n erhält man also $P(H_0|\mathbf{s})/P(H_1|\mathbf{s}) = 1$, d.h.$P(H_0|\mathbf{s}) = P(H_1|\mathbf{s})$: beide Hypothesen sind nach einem signifikanten Ergebnis gleich wahrscheinlich. Für großes n folgt aber $P(H_0|\mathbf{s}) = (5/19)P(H_1|\mathbf{s})$, d.h. $P(H_0|\mathbf{s})$ ist deutlich kleiner als $P(H_0|\mathbf{s})$, das signifikante Ergebnis spricht mehr für H_1 also für H_0. Hier ist unterstellt worden, dass die Bayessche Betrachtung der bedingten Wahrscheinlichkeiten $P(H_0|\mathbf{s})$ und $P(H_1|\mathbf{s})$ erlaubt ist. Nach Fisher repräsentiert

aber die Likelihood $P(\mathbf{s}|H_0)$ oder $P(\mathbf{s}|H_1)$ die Evidenz gegen H_0. Nun (29) enthält den Likelihood-Quotienten, und man hat

$$\frac{P(\mathbf{s}|H_0)}{P(\mathbf{s}|H_1)} = \begin{cases} 1/5, & n \text{ klein} \\ 5/19, & n \text{ groß} \end{cases} \tag{30}$$

Die Likelihood für H_1 ist in der Tat deutlich größer als die für H_0, wenn n groß ist. Bei großen Stichproben ergibt sich also eine größere Evidenz gegen H_1 und damit gegen H_0.

Nun werde das Argument von Lindley & Scott betrachtet. Statt (25) kann man auf den tatsächlichen Wert von p fokussieren, und nicht darauf, dass $\mathbf{s} : p \leq \alpha$ gilt. Dann soll also gelten

$$\frac{P(H_0|p=.05)}{P(H_1|p=.05)} = \frac{f_0(T)P(H_0)}{f_1(T)P(H_1)} \text{ bzw. } \frac{P(H_0|p=.05)}{P(H_1|p=.05)} = \frac{f_0'(T)P(H_0)}{f_1'(T)P(H_1)}, \tag{31}$$

wobei f_0 und f_1 den Fall eines kleinen n-Wertes, f_0' und f_1' den Fall eines großen n-Wertes repräsentieren. $P(H_0)/P(H_1) = 5$ in beiden Fällen, aber $f_0/f_1 < f_0'/f_1'$, wie man der Abbildung 12 entnehmen kann. Dies bedeutet, dass

$$\frac{P(H_0|p=.05, n \text{ klein})}{P(H_1|p=.05, n \text{ klein})} < \frac{P(H_0|p=.05, n \text{ groß})}{P(H_1|p=.05, n \text{ groß})}, \tag{32}$$

woraus folgt, dass ein signifikantes Ergebnis bei *kleinem* n mehr Evidenz gegen H_0 bedeutet als ein signifikantes Ergebnis bei großem n.

Der Unterschied zwischen Peto al. und Lindley et al. ergibt sich, weil Peto et al. nicht den tatsächlichen p-Wert betrachten, sondern statt dessen den Wert einer Indikatorvariablen $\chi = \{0, 1\}$: $\chi = 1$, wenn $p \leq \alpha$, und $\chi = 0$ sonst. Lindley et al. dagegen betrachten den tatsächlichen p-Wert.

Damit ist gezeigt, dass das α-Postulat, demzufolge der p-Wert unabhängig vom Stichprobenumfang n zu interpretieren sei, nicht akzeptierbar ist.

Royall (1986) führt aus, dass zwar für die verschiedenen Interpretationen spezifische Rechtfertigungen gefunden werden können:

> 'But the degree of creativity and sensitivity to semantic nuances that is required seems incompatible with the popular view that the significance test is a commensensical, practical tool appropriate for wide spread use in scientific reporting. No wonder many good students and good scientists find the statistical concepts embodied in "simple" tests of significance elusive." (Royall 1986, p. 314)

Anmerkungen:

1. Die Abhängigkeit des p-Wertes vom Stichprobenumfang, – wichtig!!! Man bekommt jede Differenz "signifikant" für hinreichend großes n.

2. macht die Hypothese $\theta = 0$; allgmein $\theta = \theta_0$ Sinn? Irgendwelche Unterschiede gibt es doch immer, oder? Andererseits: Telepathie, - die gibt es oder es gibt sie nicht.

3.3 Das Lindley-Paradoxon

Lindley (1957) hat gezeigt, dass der Signifikanztest einer Hypothese H ein auf dem 5%-Niveau signifikantes Resultat liefern kann, und gleichzeitig kann die Posteriori-Wahrscheinlichkeit für H – auch für kleine Priori-Wahrscheinlichkeiten für H - bei 95% liegen. Es sei $x_1, \ldots, x_n$ eine Stichprobe aus einer $N(\theta, \sigma^2)$-verteilten Population. σ^2 sei bekannt, und es wird H_0: $\theta = \theta_0$ angenommen. Als Priori-Verteilung für θ gelte $g(\theta) = c$ für $\theta = \theta_0$ und $g(\theta) = k$ eine Konstante über einem Intervall I mit $\theta_0 \in I$. Das arithmetische Mittel $\bar{x}$ der x_i liege ebenfalls in I. Man findet, gemäß Bayes' Theorem, für die Posteriori-Verteilung

$$\tilde{c} = \frac{c}{K} \exp\left[-\frac{n(\bar{x}-\theta_0)^2}{2\sigma^2}\right], \tag{33}$$

mit

$$K = c \exp\left[-\frac{n(\bar{x}-\theta_0)^2}{2\sigma^2}\right] + (1-c)\int_I \exp\left[-\frac{n(\bar{x}-\theta)^2}{2\sigma^2}\right] d\theta, \tag{34}$$

und

$$\int_I \exp\left[-\frac{n(\bar{x}-\theta)^2}{2\sigma^2}\right] d\theta = \sigma\sqrt{\frac{2\pi}{n}}.$$

Nun werde angenommen, dass der Wert von $\bar{x}$ derart sei, dass der übliche Signifikanztest für normalverteilte Variablen mit bekannter Varianz signifikant auf dem Niveau α sei. Demgemäß kann man $\bar{x} = \theta_0 + \lambda_\alpha \sigma/\sqrt{n}$ setzen, wobei λ_α eine nur von α abhängende Zahl sei. Setzt man diesen Ausdruck für $\bar{x}$ in (33) ein, so findet man

$$\tilde{c} = \frac{ce^{-\lambda_\alpha^2/2}}{ce^{-\lambda\alpha^2/2} + (1-c)\sigma\sqrt{2\pi/n}}. \tag{35}$$

Für $n \to \infty$ folgt $\sqrt{2\pi/n} \to 0$ und damit $\tilde{c} \to 1$. Damit hat man: für jedes c existiert ein von c und α abhängiger Wert n derart, dass

(i) $\bar{x}$ ist auf dem α-Niveau signifikant verschieden von θ_0,

(ii) die Posteriori-Wahrscheinlichkeit, dass $\theta = \theta_0$ ist $(1-\alpha)$.

Diese beiden Aussagen definieren das *Lindley-Paradoxon*. Nach (i) würde man argumentieren, dass der Befund Evidenz gegen H_0 ist, dass also $\theta \neq \theta_0$ ist. Nach (ii) würde man etwa für $\alpha = .05$ die Aussage $\theta = \theta_0$ mit einer Konfidenz von .95 annehmen.

Der kritische Punkt bei dieser Argumentation ist die Wahl der Priori-Verteilung. Nach Lindley ist aber die genaue Form der Priori-Verteilung nicht wesentlich, es komme nur darauf an, dass sie eine Konzentration auf θ_0 hat. Diese Eigenschaft läßt sich häufig damit rechtfertigen, dass man über einige Vorinformationen verfügt, auf Grund derer man überhaupt den Wert θ_0 für H_0 auswählt. Lindley bringt hierfür Beispiele aus der Genetik. Ein anderes Beispiel sind Experimente zur Telepathie; Lindley bezieht sich auf Experimente von Soal & Bateman (1954). Die Experimente waren so eingerichtet, dass unter der Hypothese, dass es keine telepathischen Effekte gibt, die Erfolgsquote $\theta_0 = 1/5$ war. Die Alternativhypothese ist dann $\theta_0 \neq 1/5$. Für den Fall, dass man die Hypothese der Existenz telepathischer Effekte testen will, liegt also eine Priori-Verteilung mit einer Konzentration auf $\theta_0 = 1/5$ nahe.

Es muß allerdings angemerkt werden, dass die Konvergenz von $\tilde{c}$ gegen 1 sehr langsam ist. Lindley legt einige Rechnungen vor: für signifikanten $\bar{x}$-Wert findet man für $n = 10$ den Wert $\tilde{c} = .156$, für $n = 100$ erhält man $\tilde{c} = .369$, für $n = 1000$ hat man erst $\tilde{c} = .649$. Erst für $n = 100000$ findet man $\tilde{c} = .949$.

Auf einen zentralen Unterschied zwischen einem Signifikanztest und dem Bayesschen Vorgehen kann hier bereits hingewiesen werden. Der Signifikanztest benötigt keine Priori-Verteilung und das Resultat der Anwendung des Tests scheint deshalb nur von den Daten abzuhängen. Allerdings gehen gleichwohl nicht beobachtete Werte in den p-Wert ein: es wird ja über alle T-Werte integriert, die größer als der beobachtete $T(\mathbf{x})$-Wert sind. Das Bayessche Verfahren dagegen hängt, abgesehen von der Priori-Verteilung, nur vom beobachteten Wert ab, der die Likelihood-Funktion bestimmt. Der Signifkanztest liefert nur einen Wert, den p-Wert, für die Interpretation, wohingegen die Likelihood-Funktion eine Funktion des unbekannten Parameters θ ist.

Ein weiterer Unterschied ergibt sich, wenn ein Experimentator, der etwa die Existenz telepathischer Effekte nachweisen möchte, so lange experimentiert, bis er einen signifikanten $\bar{x}$-Wert hat. Um einen signifikanten Effekt zu berichten, ist dieses Vorgehen allerdings nicht ganz ehrlich. Für einen Bayesianer, der sich auf die Likelihood-Funktion bezieht, ist die Stop-Regel dagegen unerheblich, das Ergebnis hängt nicht davon ab, ob er den Wert n der Versuche vorher festgelegt hat oder nicht. Auf diese Eigenschaft des Bayesschen Vorgehens wird weiter unten noch weiter eingegangen.

3.4 Zur Evidenz des p-Werts

3.4.1 Die Verteilung des p-Werts

Die folgenden Betrachtungen sind der Arbeit von Hung et al. (1997) entnommen. Der p-Wert ist die Wahrscheinlichkeit $P(T(\mathbf{X}) > T(\mathbf{x})|H_0)$; da aber $T(\mathbf{x})$ eine zufällige Veränderliche ist, da ja $\mathbf{x}$ eine Stichprobe ist, folgt, dass der p-Wert ebenfalls eine zufällige Veränderliche ist. Die Verteilung des p-Wertes hängt ab entweder von der Gültigkeit der Hypothese H_0 oder der Hypothese H_1. Die Verteilung gibt dann an, mit welcher Wahrscheinlichkeit man z.B. bei einer Wiederholung eines Experiments unter gleichen Bedingungen T-Werte findet, die beide zur Beibehaltung oder Zurückweisung von H_0 führen, erhält.

Der Einfachheit halber werde zunächst ein einfaches Beispiel betrachtet: die beobachtete Meßgröße sei $N(\mu, \sigma^2)$-verteilt. Es sei $\bar{x}_n$ das arithmetische Mittel einer Stichprobe vom Umfang n, und es werde H_0: $\mu = 0$ angenommen. Die Alternativhypothese sei H_1: $\mu > 0$, und α sei das Signifikanzniveau. Die Teststatistik sei $T = \bar{x}_n/(\sigma/\sqrt{n}) = \bar{x}_n\sqrt{n}/\sigma$, und σ sei gegeben. Dann ist der p-Wert durch

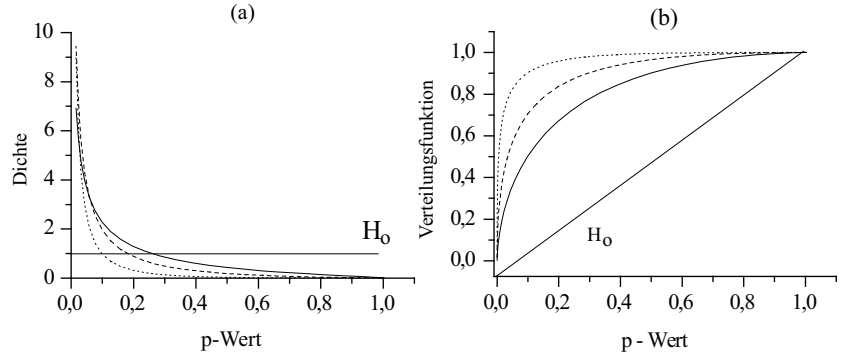
$$p = 1 - \Phi(T) \tag{36}$$

gegeben, wobei Φ die Verteilungsfunktion einer $N(0, 1)$-verteilten Variablen ist. Es sei nun $\delta = \mu/\sigma$. Es läßt sich leicht zeigen, dass dann die Vergteilungsfunktion von p durch

$$G_\delta(p) = \int_0^p g_\delta(x)dx = 1 - \Phi(Z_p - \delta\sqrt{n}) \tag{37}$$

ist, wobei Z_p das $(1-p)$-te Perzentil der $N(0, 1)$-Verteilung ist, d.h. $\Phi(Z_p) = 1-p$,

Abbildung 2: (a) Dichten und (b) Verteilungsfunktionen für p-Werte (Hung et al. 1997), durchgezogene Linie: $n = 15$; gestrichelte Linie: $n = 30$; punktierte Linie: $n = 60$. $\delta = 1/3$. Die p- Werte sind gleichverteilt, wenn H_0 gilt.



und

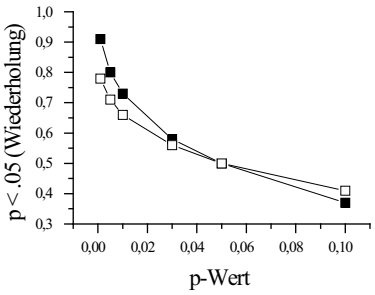
$$g_\delta(p) = \phi(Z_p - \delta\sqrt{n})/\phi(Z_p)$$

die Dichtefunktion von p, ϕ die Dichte der Standardnormalverteilung. Nach Fisher ist ein kleiner p-Wert Evidenz gegen die Hypothese H_0. Der p-Wert wird auf der Basis einer Stichprobe $\mathbf{x} = (x_1, \ldots, x_n)$ aus einer, nach Fisher, unendlich großen Population berechnet. Die Zufälligkeit von $\mathbf{x}$ bedeutet, dass die wiederholte Datenerhebung mit gleichem Stichprobenumfang und unter identischen Bedingungen verschiedene Stichproben $\mathbf{x}$, $\mathbf{x}'$, $\mathbf{x}''\ldots$ und damit verschiedene p-Werte p, p', $p''\ldots$ liefert. Die Frage ist nun, mit welcher Wahrscheinlichkeit man bei einer solchen Wiederholung der Untersuchung unter identischen Bedingungen p-Werte erhält, die ebenfalls Evidenz gegen H_0 im Sinne der Zurückweisung vonH_0 signalisieren.

Goodman (1992) ist dieser Frage nachgegangen. Er betrachtet den Fall einer normalverteilten Differenz einer Meßgröße mit H_0: $\Delta\mu = 0$ versus H_1: $\Delta\mu \neq 0$; die Varianz der Differenzen sei für beide Hypothesen identisch. Das Experiment sei so entworfen worden, dass für vorgegebene zweiseitige Wahrscheinlichkeit α und die Power $1 - \beta$ in Bezug auf eine Differenz Δ und Stichprobenumfang n einen bestimmten Wert hat. Insbesondere wurde $\alpha = .05$ angenommen; der kritische Wert ist $Z_{.025} = 1.96$. Ein Resultat ist 'signifikant', wenn $p < \alpha$. Goodman nahm insbesondere an, dass eine im ersten Experiment gefundene Differen $\Delta\bar{x}$ die 'wahre' Differenz für ein Folgeexperiment ist. Wie groß ist wieder die Wahrscheinlichkeit, dass ein signifikantes Ergebnis gefunden wird? Diese Wahrscheinlichkeit ist gleich der Power bezüglich der im ersten Experiment gefundenen Differenz $\Delta\bar{x}$, $\Phi(|Z_x| - Z_{\alpha/2})$. Für $p = .01$, $\alpha = .05$ findet man $Z_x = 2.58$, und $\Phi(2.58 - 1.96) = .73$, d.h. in diesem Fall ist die Wahrscheinlichkeit, wieder einen signifikanten p-Wert zu finden, gleich .73.

Andererseits ist die Annahme $\Delta\mu = \Delta\bar{x}$ nicht besonders plausibel. Statt dessen kann man eine lokal gleichförmige a-priori-Verteilung für μ annehmen, und als a-priori-Verteilung für das zweite Experiment die Likelihood-Funktion aus dem

Abbildung 3: Goodmans Kalkulationen. Schwarze Quadrate: $\Delta\bar{x}$ wahrer Wert für zweite Untersuchung, weiße Quadrate: Prior ist lokal gleichförmig für die erste Untersuchung, die Likelihood-Funktion der ersten ist die a-priori-Verteilung für die zweite Untersuchung. Daten: Goodman (1992), Tabelle 1, p. 877



ersten Experiment. Dann ergibt sich

$$F(z_x) = P(|Z| > Z_{\alpha/2}) = \int_{-\infty}^{\infty}\int_{Z_{\alpha/2}}^{\infty} \varphi(z - |z_x|)\varphi(y)\,dy\,dz \tag{38}$$

mit z_x dem Z-Wert des $\Delta\bar{x}$-Wertes aus dem ersten Experiment, φ die Dichte einer $N(0,1)$-verteilten Variablen. Die Ergebnisse sind bemerkenswert: für $p = .05$ ist

Tabelle 2: Goodman (1992) Wahrscheinlichkeiten für Signifikanzen bei Wiederholung eines Experiments.

Exp. I	Exp. II	
p-Wert Exp. 1	$\Delta\mu = \Delta\bar{x}$	Bayes
.10	.37	.41
.05	.50	.50
.03	.58	.56
.01	.73	.66
.005	.80	.71
.001	.91	.78

die Wahrscheinlichkeit, bei einer Replikation der Untersuchung wieder einen 'signifikanten' p-Wert zu erhalten, kleiner als 1/2. Erst für p-Werte $p < .05$ wächst diese Wahrscheinlichkeit, wobei die nach Bayes geschätzte Wahrscheinlichkeit kleiner bleibt als bei der Annahme $\Delta\mu = \Delta\bar{x}$. In anderen Worten: der p-Wert überschätzt die Evidenz gegen H_0.

Fisher argumentierte, der von ihm favorisierte p-Wert drücke induktive Evidenz aus. Goodman (1993) argumentierte dagegen: ein induktives Maß für eine Hypothese müsse eine Zahl sein, die die Glaubhaftigkeit (credibility) einer Hypothese im Licht empirischer Daten ausdrücke. Induktive statistische Evidenz könne als relative induktive Unterstützung (support) einer von zwei Hypothesen definiert werden. Da Fisher sich aber stets nur auf eine Hypothese konzentriere, könne der p-Wert kein solches Maß sein. Berkson (1942), Hacking (1965) und Goodman &

Tabelle 3: Likelihood-Quotienten für die Differenz mit 90% Power, $\alpha = .05$, für (i) $p = .05$, (ii) $p \leq \alpha$. SL = Standardisierte Likelihood; die Spalte enthält ein Maß für die minimale Evidenz für H_0, wenn $p = \alpha$. Nach Goodman (1993)

		Evidenz für H_0 versus Δ		SL
α	Z	$p = \alpha$	$p \leq \alpha$	
.10	1.64	.94	.05	.26
.05	1.96	.33	.03	.15
.03	2.17	.16	.017	.10
.01	2.58	.044	.007	.036
.001	3.28	.005	.001	.005

Royall (1988) elaborieren diesen Standpunkt. Goodman (1992) betrachtet die von Fisher eingeführte 'mathematical likelihood' als ein solches Maß; nach Fisher ist 'likelihood' nicht, wie in der Umgangssprache, ein anderes Wort für 'wahrscheinlich', sondern in Zusammenhang mit Hypothesen bedeute es das Ausmaß der Stützung einer Hypothese durch Daten (s.a. Birnbaum (1962), Hacking (1965)). Fisher:

> Mathematical likelihood is not, of course, to be confused with mathematical probability ... like mathematical probability [likelihood] can serve in a well-defined sense as a "measure of belief"; but it is a quantity of a different kind than probability, and does not obey the laws of probability.Whereas such a phrase as "the probability of A or B" has a simple meaning ... the phrase "the likelihood of A or B" is more prarallel with "the income of Peter or Paul" – you cannot know what it is ungtil you know which is meant.
>
> ... The likelihood supplies a natural order of preference among the possibilities under consideration ...
>
> Fisher (1973), p. 72 - 73

Gemeint ist der Likelihood-Quotient (8); Fisher hat den Likelihood-Quotienten als weiteres Evidenzmaß betrachtet.

Die Definition von Likelihood-Quotienten hängt davon ab, ob sie für "präzise" oder "nicht präzise" Hypothesen erklärt werden sollen. Für präzise Hypothesen gilt

$$\lambda(\mathbf{x}) = \frac{P(\mathbf{x}|H_0)}{P(\mathbf{x}|H_1)}. \tag{39}$$

Hier werden für $P(\mathbf{x}|H_i)$, $i = 1, 2$ die Dichten für $\mathbf{x}$, gegeben H_i eingesetzt, woraus folgt, dass die Likelihoods keine Wahrscheinlichkeiten sind. Für die "nicht präszise" Hypothese $p < .003$ etwa stehen die $P(\cdot|H_i$ für die entsprechenden Integrale der Dichten. Goodman (1993) betrachtet nun die Likelihoods (i) für H_0 und (ii) für H_1, wobein H_1 für eine Differenz von Mittelwerten steht, die für vorgegebenes $\alpha = .05$ jeweils eine Power von 90 % $(1 - \beta)100)$ hat. Als Beispiel betrachte man den Fall $\alpha = .05$. Der Likelihood-Quotient beträgt $.33/.03 = 11$; die Alternativhypothese $p \leq \alpha$ demnach 11-mal weniger gestützt als H_0: $p = .05$. Nimmt man $P(H_0) = P(H_1) = 1/2$ an, betrachtet man also beide Hypothesen als gleich wahrscheinlich,

Tabelle 4: p-Werte und dazu korrespondierende $P(H_0|\mathbf{x})$-Werte für $n = 1$ bis $= 1000$; a-priori-Verteilung vom Jeffreys-Typ. Nach Berger & Sellke 1987

		$P(H_0\|\mathbf{x})$ als Funktion von n						
p	t	1	5	10	20	50	100	1000
.10	1.645	.42	.44	.47	.56	.65	.72	.89
.05	1.960	.35	.33	.37	.42	.52	.60	.82
.01	2.576	.21	.13	.14	.16	.22	.27	.53
.001	3.291	.086	.026	.025	.026	.034	.045	.124

so entspricht der Quotient dem Quotienten der a-posteriori-Wahrscheinlichkeiten $P(H_0|D)/P(H_1|D)$. Ist

$$\frac{P(H_0|D)}{P(H_1|D)} = \frac{P(H_0|D)}{1 - P(H_0|D)} = \lambda,$$

so folgt

$$P(H_0|D) = \frac{\lambda}{1+\lambda}.$$

Für $\alpha = .05$ ist $\lambda = .33$, und mithin $P(H_0|D) = .33/(1 + .33) \approx .25$, d.h. $P(H_0|D) \approx P(H_0)/2$. Für $p \leq \alpha$ ist $\lambda = .03$ und $P(H_0|D) \approx .03 \approx .06P(H_0)$. Betrachtet man also $p \leq \alpha$, so erscheint die Evidenz gegen H_0 stärker ausgeprägt, als wenn man $p = \alpha$ betrachtet. Fisher hat sowohl den p-Wert als auch den Likelihood-Quotienten als Evidenzmaß propagiert, ohne dass er jemals den Vorteil des p-Wertes elaboriert hätte (Hacking (1965) bemerkt, dass Fisher die Definition des p-Werts nie gerechtfertigt hat).

Berger & Sellke 1987: Betrachtet wird eine Stichprobe $\mathbf{x} = (x_1, \ldots, x_n)$, wobei die x_i identisch und unabhängig voneinander (iid = identically and independently distributed) normalverteilt mit dem Erwartungswert μ und der Varianz σ^2. Die Teststatistik ist dann $T(\mathbf{X}) = \sqrt{n}|\bar{X} - \mu_0|/\sigma$, wobei $\bar{x}$ der Stichprobenmittelwert ist. Der p-Wert ist durch

$$p = P_{\mu=\mu_0}(T(\mathbf{X}) \geq T(\mathbf{x})) = 2(1 - \Phi(t)) \tag{40}$$

gegeben, $\Phi(t) = \int_{-\infty}^{t} \varphi(\tau)\, d\tau$, φ die Dichte der Standardnormalverteilung $N(0, 1)$, $t = T(\mathbf{x})$.

Andererseits sei die a-priori-Verteilung durch $V(\mu,\sigma^2)$ gegeben. Es läßt sich zeigen (Abschnitt 6.1.2, Seite 79), dass dann die a-posteriori-Wahrscheinlichkeit für H_0: $\mu = \mu_0$ durch

$$P(H_0|\mathbf{x}) = \frac{1}{1 + \frac{\exp(t^2/[2(1+1/n)])}{\sqrt{1+1/n}}} \tag{41}$$

gegeben ist. Für $n = 50$, $p = .05$ und also $t = 1.96$ erhält man eine a-posteriori-Wahrscheinlichkeit $P(H_0|\mathbf{x}) = .52$; Für $p = .05$ würde die Hypothese also nach Maßgabe des p-Wertes gerade "verworfen", während etwa für $n = 50$ und damit $t = 1.96$ ihre a-posteriori-Wahrscheinlichkeit durch $P(H_0|\mathbf{x}) = .52$ gegeben ist. Dieser Wert ist aber Evidenz *für* H_0. Erst für $p = .001$ ergibt sich für $n = 50$

Tabelle 5: p-Werte und dazu korrespondierende $P(H_0|\mathbf{x})$ - sowie $\inf_\mu \lambda$ - Werte

p	t	$P(H_0\|\mathbf{x})$	$\inf_\mu \lambda$
.10	1.645	.340	.258
.05	1.960	.227	.146
.01	2.576	.068	.036
.001	3.291	.0088	.0044

die a-posteriori-Wahrscheinlichkeit $P(H_0|\mathbf{x}) = .026$. Je größer der n-Wert, desto größer ist $P(H_0|\mathbf{x})$ für gegebenen p-Wert.

Natürlich hängen die a-posteriori-Wahrscheinlichkeiten von der a-priori-Verteilung ab. Eine Alternative ist die Verteilung $P(H_0) = P(H_1) = 1/2$, wobei aber $P(H_1)$ so über den Bereich $\mu \neq \mu_0$ ausgebreitet wird, dass die a-priori-Wahrscheinlichkeit für H_1 *so günstig wie möglich* ist.

Unter diesen Bedingungen ergeben sich die $P(H_0|\mathbf{x})$-Werte der Tabelle 5. Obwohl also die a-priori-Verteilung extrem unfair gegenüber H_0 ist, ergeben sich immer noch im Vergleich zu p große Werte für $P(H_0|\mathbf{x})$.

Auch eine Likelihood-Analyse liefert Evidenz, die der der p-Werte widerspricht. Es seien μ_0 und μ_1 zwei mögliche Parameterwerte. Der Likelihood-Quotient ist durch

$$\lambda(\mu_0, \mu_1) = \frac{f(\mathbf{x}|\mu_0)}{f(\mathbf{x}|\mu_1)} \tag{42}$$

gegeben. Weiß man nicht, welchen speziellen μ_1-Wert man nehmen soll, kann man eine untere Grenze für λ definieren:

$$\inf_\mu \lambda = \frac{f(\mathbf{x}|\mu_0)}{\sup_\mu f(\mathbf{x}|\mu)} = e^{-t^2/2}. \tag{43}$$

Für $p = .05$, $t = 1.96$ hat man $\inf_\mu \lambda = .146$. Während $p = .05$ bedeutet, dass man H_0 "verwirft", ist dieser untere Wert für den Likelihood-Quotienten keineswegs besonders klein, zumal die untere Grenze von λ besonders unfair gegenüber H_0 ist.

Die Ursache liegt in der Definition von p. Dieser Wert hängt nicht nur von den Daten $\mathbf{x}$ ab, sondern auch von den nicht beobachteten Stichproben $\mathbf{X}$, für die $T(\mathbf{X}) \geq T(\mathbf{x})$ (Berger & Sellke (1987), Abschnitt 4).

3.4.2 Unvereinbarkeit von p-Wert und Evidenz:

(Berger & Sellke 1987) Es wird wieder der Test der Hypothese H_0: $\theta = \theta_0$ betrachtet. Der "klassische Ansatz" besteht darin, $p = P(T(\mathbf{X}) \geq T(\mathbf{x})|\theta = \theta_0)$ zu berichten. Ist $\mathbf{x} = (x_1, \ldots, x_n)$ mit $x_i \sim N(\theta, \sigma^2)$, σ^2 bekannt, so hat man $T(\mathbf{X}) = \sqrt{n}|\bar{X} - \theta_0|/\sigma$ und $p = 2(1 - \Phi(t))$, $t = T(\mathbf{x}) = \sqrt{n}|\bar{x} - \theta_0)/\sigma$.

Es werde nun der Jeffreysche Ansatz betrachtet. Die Priori-Verteilung wird so definiert, dass H_0 und H_1 jeweils die Wahrscheinlichkeit $1/2$ zugeordnet werden, wobei die Masse nach einer $N(\theta_0, \sigma^2)$-Verteilung über H_1 verteilt wird.

In section 2: $0 < \pi_0 < 1$ die Prior für H_0. Für H_1 hat man die eben genannte Gauss-Verteilung auf $\theta \neq \theta_0$. Das Problem hierbei ist, dass θ_0 ja ein singulärer Wert

ist, und die Wahrscheinlichkeit, dass ein *bestimmter* Wert aus einem Kontinuum beobachtet wird, ist gewöhnlich gleich Null, – hier wird diesem Wert aber die Wahrscheinlichkeit 1/2 zugeordnet. Man ist kann dies als Approximation an H_0: $|\theta - \theta_0| \leq b$ verstehen, und π_0 ist die Wahrscheinlichkeit, dass ein $\theta \in \{\theta || \theta - \theta_0| \leq b\}$ beobachtet wird. Die Randdichte für X ist

$$m(x) = f(x|\theta_0)\pi_0 + (1 - \pi_0)m_g(x), \tag{44}$$

mit

$$m_g(x) = \int f(x|\theta)g(\theta)d\theta. \tag{45}$$

Für die Posterior-Dichte erhält man dann, $f(x|\theta) > 0$,

$$\begin{aligned} p(H_0|x) &= f(x|\theta)\frac{\pi_0}{m(x)} \\ &= \left[1 + \frac{1 - \pi_0}{\pi_0} \cdot \frac{m_g(x)}{f(x|\theta)}\right]^{-1}. \end{aligned} \tag{46}$$

Der Posterior-odds-ratio für H_0 zu H_1:

$$\frac{P(H_0|\mathbf{x})}{1 - P(H_0|\mathbf{x})} = \frac{\pi_0}{1 - \pi_0} \cdot \frac{f(\mathbf{x}|\theta_0}{m_g(\mathbf{x})}, \tag{47}$$

wobei $\pi_0/(1 - \pi_0)$ der *Prior-odds-Quotient* ist und

$$B_g(\mathbf{x}) = \frac{f(\mathbf{x}|\theta_0)}{m_g(\mathbf{x})} \tag{48}$$

ist der *Bayes-Faktor* H_0 versus H_1. Der Bayes-Faktor hängt nicht von den Priori-Verteilungen ab und entspricht den "odds" der Daten für die beiden Hypothesen (verwandt mit dem Likelihood-Quotienten).

Ein Beispiel erhält man, wenn π_0 beliebig ist und $g \sim N(\theta_0, \sigma^2)$. Eine suffiziente Statistik für θ ist $\bar{X} \sim N(\theta, \sigma^2/n)$. Für den Bayes-Faktor erhält man (Berger & Sellke 1987, p. 115)

$$B_g(\mathbf{x}) = \sqrt{1 + n} \exp\left(-\frac{t^2}{2(1 + 1/n)}\right) \tag{49}$$

und

$$P(H_0|\mathbf{x}) = \left[1 + \frac{1 - \pi_0}{\pi_0\sqrt{1 + n}} \exp\left(\frac{t^2}{2(1 + 1/n)}\right).\right]^{-1} \tag{50}$$

Für festes t und $n \to \infty$ strebt $P(H_0|\mathbf{x})$ gegen 1, undabhängig vom p-Wert und konsistent mit dem Lindley-Paradoxon.

Berger & Sellke argumentieren, dass die Wahl $\pi_0 = 1/2$ die Priori-Verteilung in gewisser Weise objektiv mache; der Geltung und der Nichtgeltung von H_0 werden gleiche Wahrscheinlichkeiten zugeordnet und auf diese Weise werd weder H_0 noch H_1 der Vorzug gegeben. Im Falle eines Telepathie-Experiments könnte man π_0 auch klein wählen, etwa $\pi_0 = .1$, womit man massiver Skepsis gegenüber der Existenz telepathischer Effekte ausdrücken könnte.

Weitere Literatur zu Bayesschen Tests von Punkthypothesen der Form H_0: $\theta = \theta_0$ in Berger & Sellke (1987, p. 115).

Untere Grenzen für die Posteriori-Verteilungen: $g(\theta)$ ist die Verteilung der θ-Werte, wenn H_1 wahr ist. Man betrachtet nun eine Klasse G von Verteilungen, die in einem allgemeinen Sinne "vernünftig" sind. Man kann dann den Minimalwert von $P(H_0|\mathbf{x})$ abschätzen. Betrachtet werden insbesondere G_A: die Klasser aller Verteilungen g, G_S: die Klasse aller Verteilungen, die symmetrisch in Bezug auf θ_0 sind, G_{us}: die Klasse aller Verteilungen, die unimodal und symmetrisch bezüglich θ_0 sind,und G_{nor}: die Klasse aller Verteilungen $N(\theta_0, \tau^2)$, mit $0 \leq \tau^2 < \infty$. Es sei

$$\underline{P}(H_0|\mathbf{x}, G) = \inf_{g \in G} P(H_0|\mathbf{x}) \tag{51}$$

und

$$\underline{B}(\mathbf{x}, G) = \inf_{g \in G} B_g(\mathbf{x}). \tag{52}$$

Dann

$$\underline{B}(\mathbf{x}, G) = \frac{f(\mathbf{x}|\theta_0)}{\sup_{g \in G} m_g(\mathbf{x})}$$

und

$$\underline{P}(H_0|\mathbf{x}, G) = \left[1 + \frac{\pi_0}{1 - \pi_0} \frac{1}{\underline{B}(\mathbf{x}, G)}\right]^{-1}.$$

$\sup_{g \in G} m_g(\mathbf{x})$ kann als eine obere Grenze der "Likelihood" von H_1 über den Gewichten $g \in G$ angesehen werden; $\underline{B}(\mathbf{x}, G)$ kann als untere Grenze der Likelihood von H_0 relativ zu H_1 gesehen werden. Schon in Edwards et al (1963) wird die folgende Aussage bewiesen:

Satz 3.1 *Es sei $\hat{\theta}$ eine Maximum-Likelihood-Schätzung für θ (d.h. es wird angenommen für gegbebenes $\boldsymbol{x}$, dass eine solche Schätzung existiert). Dann gilt*

$$\underline{B}(\boldsymbol{x}, G_A) = \frac{f(\boldsymbol{x}|\theta_0)}{f(\boldsymbol{x}|\hat{\theta})}, \tag{53}$$

und

$$\underline{P}(H_0|\boldsymbol{x}, G_A) = \left[1 + \frac{\pi_0}{1 - \pi_0} \frac{f(\boldsymbol{x}|\hat{\theta})}{f(\boldsymbol{x}|\theta_0)}\right]^{-1}. \tag{54}$$

Anmerkung: Numerisches Beispiel B&S, s. S. 116.

Lower Bounds für G_s (Symmetrische Verteilungen): B & S zeigen zunächst, dass die Minimalisierung von $P(H_0|\mathbf{x})$ über alle $g \in G_s$ äquivalent einer Minimalisierung über der Klasse G_{2PS}, der Menge der symmetrischen 2-Punkt-Verteilungen ist:

Satz 3.2

$$\sup_{g \in G_s} m_g(\boldsymbol{x}) = \sup_{g \in G_{2PS}} m_g(\boldsymbol{x}), \tag{55}$$

und also

$$\underline{B}(\boldsymbol{x}, G_{2PS}) = \underline{B}(\boldsymbol{x}, G_S) \tag{56}$$

und

$$\underline{P}(H_0|\boldsymbol{x}.G_{2PS}) = \underline{P}(H_0|\boldsymbol{x}, G_s). \tag{57}$$

Es läßt sich zeigen, dass

$$\lim_{t\to\infty} \frac{P(H_0|\mathbf{x}, G_s)}{P(H_0|\mathbf{x}, G_A)} = 2, \tag{58}$$

d.h. kann die Klasse der Priori-Verteilungen auf symmetrische Verteilungen beschränkt werden, so ist die Posteriori-Evidenz für H_0 noch größer als im allgemeinen Fall. Weitere Ergebnisse findet man in Berger & Sellke (1987).

Test von präzisen Hypothesen: (Berger & Delampady 1987) Betrachtet wird der Test der exakten Hypothese H_0: $\theta = \theta_0$ versus H_1: $\theta \neq \theta_0$. Die Stichprobenverteilung von $\mathbf{X}$ sei $f(\mathbf{X}|\theta)$. Es werden (i) der p-Wert, (ii) der gewogene Likelihood-Quotient (Bayes-Faktor), und (iii) die Bayessche Posteriori-Wahrscheinlichkeit von H_0 betrachtet.

Der p-Wert $T(\mathbf{X})$ sei wieder die Teststatistik, und große Werte dieser Variablen seien Evidenz gegen H_0. Es ist

$$p = P_{\theta_0}(|T(\mathbf{x})| \geq |t|). \tag{59}$$

Der Likelihood-Quotient H_0 zu H_1 oder Bayes-Faktor ist durch

$$B = \frac{f(\mathbf{x}|\theta)}{m_g(\mathbf{x})} \tag{60}$$

gegeben. Dabei ist $g(\theta)$ eine stetige Dichte auf der Menge $\{\theta|\theta \neq \theta_0\}$, und

$$m_g(\mathbf{x}) = \int f(\mathbf{x}|\theta) g(\theta) d\theta. \tag{61}$$

Es gibt zwei Interpretationen für g, für einen Bayesianer ist g die Priori-Dichte für θ unter der Bedingung, dass H_1 wahr ist, für einen Anhänger des Likelihood-Quotienten-Tests ist g einfach eine Gewichtsfunktion, über die die durchschnitliche Likelihood für H_1 berechnet werden kann. $B = 1/5$ bedeutet in jedem Fall, dass H_1 fünfmal so stark durch die Daten gestützt wird als H_0. Berger et al. wählten die Gauß-Verteilung $N(\mu, \tau^2)$. Dann ergibt sich

$$B = \sqrt{1 + 1/\rho^2} \exp\left[-\frac{1}{2}\left(\frac{(t-\rho\eta)^2}{1+\rho^2} - \eta^2\right)\right], \tag{62}$$

mit $\rho = \sigma/(\tau\sqrt{n})$; $\eta = (\theta_0 - \mu)/\tau$. Mit $\mu = \theta_0$, $\tau = \sigma$ und $\tau_0 = 1/2$ ergeben sich die Werte in Abb. 4.

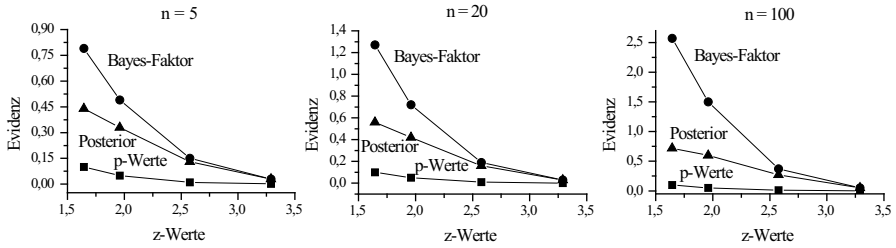
Über den Satz von Bayes kann nun die Posteriori-Wahrscheinlichkeit für H_0 berechnet werden. Ist die Priori-Wahrscheinlichkeit für H_0 durch $\pi_0 = \pi(H_0)$ gegeben, so folgt für die Posterori-Wahrscheinlichkeit

$$P(H_0|\mathbf{x}) = \left[1 + \frac{(1-\pi_0)}{\pi_0} \frac{m_g(\mathbf{x})}{f(\mathbf{x}|\theta_0)}\right]^{-1} = \left[1 + \frac{(1-\pi_0)}{\pi_0} \frac{1}{B}\right]^{-1}. \tag{63}$$

Man sieht sofort, dass $P(H_0|\mathbf{x})$ groß wird, wenn der Bayes-Faktor B groß wird.

Berger & Delampady diskutieren die Hypothese, dass Signifikanzteste und Bayessche Berwertungen übereinstimmen. Das kann nur bedingt gelten, da ja die Priori-Verteilungen verschieden gewählt werden und sich Übereinstimmung mit p-Werten erst für sehr großes n einstellt. Sie berechneten für normalverteiltes T einerseits die p-Werte, dann die zugehörigen B-Werte (Bayes-Faktoren oder Likelihood-Quotienten) sowie die Posteriori-Wahrscheinlichkeiten (Tabelle 1 in Berger et al

Abbildung 4: p-Wert, Bayes-Faktor und Posteriori Wahrscheinlichkeit als Evidenzmaße für verschiedene n-Werte (Berger & Delampady 1987)



(1987). Für die Fälle $n = 5$, $n = 20$ und $n = 100$ sind die Werte in Abbildung 4 graphisch dargestellt worden. Für $z = 1.96$ erhält man $p = .05$; nach der Logik des Signifikanztests würde H_0 auf $\alpha = .05$-Niveau verworfen. Die Posteriori-Wahrscheinlichkeit beträgt aber für $n = 5$ schon $P(H_0|\mathbf{x}) = .35$, für $n = 20$ hat man $P(H_0|\mathbf{x}) = .42$, und für $n = 100$ findet man $P(H_0|\mathbf{x}) = .6$. Je größer der Stichprobenumfang n, desto größer wird die Posteriori-Wahrscheinlichkeit des Likelihood-Quotienten (Bayes-Faktor) als auch der Posteriori-Wahrscheinlichkeit bei, und zwar um so eher, je größer der Stichprobenumfang ist. Die Frage ist nun, welche der Evidenzmaße zur Entscheidung bzw. Bewertung herangezogen werden sollen.

Man könnte argumentieren, dass die in Abb. 4 präsentierten Resultate einfach nur das Resultat der speziellen Wahl der Priori-Verteilung $g \sim N(\mu, \sigma^2)$. Dieses Argument läßt sich entkräften. Berger et al. betrachten die untere Grenze für B und $P(H_0|\mathbf{x})$ für große Klassen G von Priori-Verteilungen,

$$\underline{B} = \inf_{g \in G} B = \frac{f(\mathbf{x}|\theta_0}{\sup_{g \in G} m_g(\mathbf{x})}, \tag{64}$$

so dass

$$\underline{P}(H_0|\mathbf{x}) = \inf_{g \in G} P(H_0|\mathbf{x}) = \left[1 + \frac{(1-\pi_0)}{\pi_0} \frac{1}{\underline{B}}\right]^{-1}. \tag{65}$$

G kann im Extremfall die Klasse aller Dichten sein, oder die Klasse der konjugierten Dichten ("conjugate priors") mit dem Erwartungswert θ_0, etc. Der interessante Fall ergibt sich, wenn man ein g wählt, das H_1 favorisiert. Edwards, Lindemann und Savage (1963) haben bereits gezeigt, dass sich bei einer solchen Analyse immer noch weniger Evidenz gegen H_0 ergibt, als es der p-Wert suggeriert.

Nullhypothesen der Form H_0: $\theta = \theta_0$ stellen ein spezielles Problem dar. Oft werden Hypothesen bezüglich eines Unterschiedes zwischen Gruppen oder experimentellen Bedingungen untersucht, so dass $\theta = \mu_1 - \mu_2$, und $\theta_0 = \Delta\mu = \mu_2 - \mu_1 = 0$. Solche Hypothesen werden gelegentlich als unrealistisch bezeichnet. Will man die Hypothese testen, dass die Ostfriesen tatsächlich weniger intelligent als die Westfalen sind, so kann man leicht argumentieren, dass die Nullhypothese $\Delta\mu = 0$ kaum realistisch ist, ein bißchen intelligenter als die andere wird eine der beiden Gruppen schon sein, es komme darauf an, ob der Unterschied 'relevant' ist, d.h. ob er vernachlässigbar sei oder nicht[8]. Andererseits gibt es echte Nullhypothesen

[8]Man könnte postulieren, dass es eine dem Ostfriesen eigene genetische Disposition zu ge-

der Form $\theta_0 = 0$, wenn etwa die Möglichkeit der Existenz telepathischer Fähigkeiten unersucht werden soll. Da dem Signifikanztest nachgesagt wird, noch so kleine Abweichungen von θ_0 würden von ihm entdeckt, so müßten sich im Laufe der Zeit die Befunde akkumulieren, denen zufolge zumindest spurenhafte Telepathieeffekte existieren. Andererseits haben die Betrachtungen von Edwards et al (1963) und Berger & Delampady (1987) gezeigt, dass der p-Wert die Evidenz gegen H_0 überschätzen kann. Eine Bayes-Analyse könnte zu anderen Schlüssen kommen. Dann aber ergibt sich die Frage, welche Priori-Verteilung denn für eine Punkthypothese der Form $\theta = \theta_0$ gewählt werden soll. Berger et al. diskutieren diese Frage ausführlich, aber sie soll an dieser Stelle nicht weiter verfolgt werden, da hier Eigenschaften des Signifikanztests zur Diskussion stehen. Im Abschnitt über Bayes-Verfahren wird auf diese Frage zurückgekommen.

s.a. Goodman 1992 Comment on replication, p-values and evidence: probability of repeating a statistically significant result – the replication probability – is substantially lower than expected. Reason is interpretation of post-trial p-value as pre-trial α-error.

Schervish (1995) liefert weitere Argumente gegen eine evidentielle Interpretation von p-Werten. Er beginnt mit der Feststellung, dass in den Standardlehrbüchern für (elementare) Statistik Signifikanztests für einseitige Hypothesen und für Punkthypothesen (point-null hypotheses) behandelt werden, aber gewöhnlich nicht für Intervallhypothesen, also etwa für Hypothesen der Art $H\colon \mu_0 \leq \theta \leq \mu_1$. Nimmt man diesen Hypothesentyp mit hinzu, so läßt sich zeigen, dass der p-Wert eine stetige Funktion der Daten sind. Es folgt dann, dass der p-Wert nicht als (informelles) Maß für das Ausmaß der Stützung einer Hypothese interpretiert werden kann.

So sei $X \sim N(\mu, 1)$, also $\theta = \mu$. Die üblichen Tests sind $H_1\colon \theta = \mu_0$ versus $A_1\colon \theta \neq \mu_0$ (dies ist die point-null-Hypothese); $H_2\colon \theta \leq \mu_0$ versus $A_2\colon \theta > \mu_0$ oder $H_3\colon \theta \geq \mu_0$ versus $A_3\colon \theta < \mu_0$. Zusätzlich kann man noch $H_4\colon \theta \in [\mu_0, \mu_1]$ versus $A_4\colon \theta \notin [\mu_0, \mu_1]$ betrachten. Sowohl die point-null Hypothese als auch die einseitigen Hypothesen sind Spezialfälle von Intervallhypothesen der Form H_4, wenn entweder $\mu_0 \to \mu_1$ oder $\mu_0 \to -\infty$ oder $\mu_1 \to \infty$. Weiter sei $a \in [-\infty, \infty)$, $b \in [a, \infty]$, und der P-Wert werde mit $p_{a,b}(x)$ bezeichnet. Für $a = b$ erhält man H_1, für $a = -\infty$ erhält man H_2, für $b = \infty$ folgt H_3, und für $a < b$, a, b endlich, erhält man H_4. Φ bezeichne, wie üblich, die Verteilungsfunktion für $N(\mu, \sigma^2)$ (hier $\sigma^2 = 1$). Dann folgt

$$p_{\mu_0,\mu_1}(x) = 2\Phi(-|x - \mu_0|), \quad p_{\mu,\infty}(x) = \Phi(x - \mu_0), \quad p_{-\infty,\mu_0}(x) = \Phi\mu_0 - x).$$

Für Intervallhypothesen der Form H_4 existiert ein 'gleichmäßig bester unverfälschter Test' (UMPU = uniformly most powerful unbiased) zum Niveau α (Lehmann (1986), Abschn. 4.2) gemäß: verwerfe H_4, wenn

$$|X - .5(\mu_1 + \mu_2)| > c,$$

wobei c so gewählt wird, dass

$$\Phi[.5(\mu_1 - \mu_2) - c] + \Phi[.5(\mu_2 - \mu_1) - c] = \alpha. \tag{66}$$

ringerer Intelligenz gibt. Es könnte aber auch sein, dass es diese Disposition nicht gibt, dafür aber eine nicht genetisch, sondern ökonomisch bedingte Tendenz existiert, dass intelligentere Ostfriesen Ostfriesland verlassen, weil sich in anderen Teilen der Republik bessere berufliche Entfaltungsmöglichkeiten bieten. Im Laufe der Jahrzehnte *könnte* sich dieses Verhalten auf den Populationsmittelwert der in Ostfriesland lebenden Ostfriesen auswirken.

Diese Gleichung liefert sowohl den p-Wert p_{μ_1} (verwerfe H_4) als auch p_{μ_2} (ebenfalls verwerfe H_4), wobei p_{μ_i}, $i = 1, 2$, die bedingte Wahrscheinlichkeit, gegeben μ_i bedeutet. (66) bezeichnet eine differenzierbare, streng monoton fallende Funktion von c, die für $c = 0$ gleich 1 wird und die gegen 0 strebt für $c \to \infty$. Damit hat c eine eindeutige Lösung $c = c(\alpha; \mu_2 - \mu_1)$. Dann wird

$$\begin{aligned} p_{\mu_1,\mu_2}(x) &= c^{-1}(|x - .5(\mu_1 + \mu_2)|\mu_2 - \mu_1) \\ &= \begin{cases} \Phi(x - \mu_1) + \Phi(x - \mu_2), & x < .5(\mu_1 + \mu_2) \\ \Phi(\mu_1 - x) + \Phi(\mu_2 - x), & x \geq .5(\mu_1 + \mu_2) \end{cases}, \end{aligned} \tag{67}$$

mit $c^{-1}(y; z) = d$ derart, dass $c(d; z) = y$. Daraus läßt sich ableiten, dass p-Werte für point-null Hypothesen Grenzfälle für Intervallhypothesen sind (Schervish (1995), p. 204). Dickey (1967) diskutiert die Frage, ob die "tail area" einen Bayes-Faktor approximiert.

Eine Hypothese H impliziere nun eine andere Hypothese H'. Ein Maß für die Unterstützung (measure of support) einer Hypothese heißt nun 'kohärent', wenn eine Zurückweisung von H' impliziert, dass auch H zurückgewiesen wird ($H \Rightarrow H'$ impliziert, dass $\neg H' \Rightarrow \neg H$). Umgekehrt muß jede Stütze für H auch eine Stütze für H' sein. Es zeigt sich aber, dass p-Werte sich nicht als kohärente Maße erweisen.

3.4.3 Stichprobenumfang und Stop-Regel

Goodman (1999a) spricht von einer "p-values fallacy": einerseits ist der p-Wert ein Evidenzmaß, andererseits wird er als Fehlerrate interpretiert. Im NP-Ansatz ist α eine Fehlerrate, weil α im frequentistischen Sinn gedeutet wird: α gibt an, wie häufig man bei hinreichend häufiger Wiederholung des Experiments unter identischen Bedingungen eine falsche Entscheidung gegen H_0 trifft. Nach Fisher ist aber der p- Wert ein induktives Maß der Evidenz gegen H_0 bei einer einmaligen Untersuchung; dies ist die "short run"-Perspektive.

Beispiel 3.2 (**Test von Behandlungen nach Royall)** Es gebe 2 Behandlungen A und B. H_0 besagt, dass beide Behandlungen gleich erfolgreich sind, also $P(A) = p(B) = 1/2$. Es werden sechs Patienten untersucht: A erweist sich als besser bei den ersten 5 Patienten, B beim 6-ten. Dieses Ergebnis wird auf zwei Arten gewonnen: es werde angenommen, dass es zwei Untersucher gibt, die unabhängig voneinander handeln:

1. Der Untersucher U_1 plant von vornherein, $n = 6$ Patienten zu untersuchen.
2. Experimenter U_2 beendet das Experiment, sobald die Behandlung B erfolgreicher ist.

X sei die Anzahl der "Erfolge", d.h. gleich der Anzahl der Male, bei denen A besser als B war. Die Wahrscheinlichkeit des Resultats des U_1-Experiments ist

$$P(X = 1|H_0) = \binom{6}{1} \left(\frac{1}{2}\right)^1 \left(\frac{1}{2}\right)^5 = 6 \left(\frac{1}{2}\right)^6 = .094$$

Es sei Y die Anzahl der Misserfolge, so dass $X + Y = 6$ bzw. $Y = 6 - X$. In Fishers Signifikanztest wird die Wahrscheinlichkeit der Daten *oder extremerer* Daten

bestimmt, d.h. hier $X \leq 1$ oder $Y \geq 5$. Man findet

$$P(Y \geq 5) = P(X \leq 1) = \left(\binom{6}{5} + \binom{6}{6} \right) \left(\frac{1}{2} \right)^6 = 7 \left(\frac{1}{2} \right)^6 = .11.$$

Für den Untersucher U_2 ergibt sich ein anderes Bild, – obwohl die Daten die gleichen sind. Die Wahrscheinlichkeit der Folge von 5 Erfolgen für A und eines Erfolges für B ist

$$P(AAAAAB) = \left(\frac{1}{2} \right)^5 \left(\frac{1}{2} \right)^1 = \left(\frac{1}{2} \right)^6 = .016,$$

und ein extremeres Ergebnis wäre, wenn in sechs Fällen A erfolgreich gewesen wäre, so dass nun der p-Wert durch

$$P(\text{Daten}) + P(\text{extremere Daten}) = \left(\frac{1}{2} \right)^5 \left(\frac{1}{2} \right)^1 + \left(\frac{1}{2} \right)^6 \left(\frac{1}{2} \right)^0 = .03$$

gegeben ist. Man hat also die gleichen Behandlungen, die gleichen Patienten, nur verschiedene p-Werte, die nur von der Strategie des jeweiligen Untersuchers abhängen. □

Das Paradoxon entsteht, so Royall, weil das "long run" - Verhalten und die "short run" - Bedeutung durch die gleiche Zahl beschrieben werden sollen. Die *long run*-Ergebnisse der beiden Arten von Untersuchung sind völlig verschieden und haben nur (i) das beobachtete Resultat und (ii) das nicht beobachtete Resultat AAA$\cdots$A gemeinsam (zur Berechnung des p-Werts). Dieser Befund illustriert die *p-value fallacy*. Bleibt man bei der *short run*-Perspektive, so resultieren gleiche Daten in gleichen p-Werten, unabhängig von den Strategien verschiedener Untersucher. (s.a. Goodman II)

3.5 Einige kritische Betrachtungen

Kritische Betrachtungen zum p-Wert hat es schon relativ früh gegeben, etwa bei Berkson (1942). Rozeboom (1960) stellte fest, dass der Signifikanztest zu einem inferenzstatistischen Dogma geworden sei, das zumindest unter Psychologen den Status einer religiösen Überzeugung gewonnen habe, vergl. etwa Gigerenzer (1989, 1993a, 1993b).

In der Tat hat Melton (1962) in einem Resumé seiner Tätigkeit als Editor des *Journal of Experimental Psychology* festgestellt, dass unter seiner Regie nur Ergebnisse mit $P(T > T(\mathbf{x})|H_0) < .01$ zur Veröffentlichung akzeptiert worden seien, denn nur so könne gesichert werden, dass vernünftige Aussagen gemacht werden können. Meltons Ansatz ist deutlich kritisiert worden, worauf hier nicht im Detail eingegangen werden muß. Sterling (1959) über die religiöse Bedienung des Signifikanztests und die Konsequenzen für die Wissenschaft. Schon Rozeboom (1960, p. 417) hatte festgestellt, dass eine Hypothesenevaluation nach dieser Art

> "... is based on a fundamental misunderstanding of the nature of rational inference, and is seldom if ever appropriate to the aims of scientific research."

Natürlich gab es auch 1960 bereits alternative Verfahren (etwa den Ansatz von Neyman-Pearson, oder Bayes-Verfahren), aber

> "'... the perceptual defenses of psychologists are particularly efficient when dealing with matters of methodology, and so the statistical folkways of a more primitive past continue to dominate the local scene."

Weitere kritische Anmerkungen zum Signifikanztest findet man bei Berkson (1942), insbesondere mit Bezug zum Evidenzbegriff. Goodman et al (1988) argumentieren, dass "p-values miss the evidence", und Huberty (1993) beklagt die unordentliche Behandlung der Ansätze von Fisher einerseits und Neyman-Pearson andererseits in Lehrbüchern. Es ist also sinnvoll, die Eigenschaften des Signifikanztests etwas näher zu betrachten.

Ein unmittelbares, mit der Definition (22) des Signifikanztests verbundenes Problem ist, dass die Entscheidung über H_0 nicht nur von den beobachteten Daten $\mathbf{x}$ abhängt, sondern von allen Werten T, die größer als $T(\mathbf{x})$ sind. Verstehen wird dann durch Gewöhung bzw. durch den Druck, eine Klausur bestehen zu müssen, ersetzt. Während in Lehrbüchern der Mathematischen Statistik der Fishersche Signifikanztest nicht oder nur am Rande erwähnt wird und statt dessen der Ansatz von Neyman-Pearson als kanonisch präsentiert wird (Lehmann (1959), Kendall & Stuart (1973), Pruscha (2000), und viele andere), verzichtet die Darstellung des Signifikanztests in vielen Lehrbüchern der Statistik für Psychologen, Sozialwissenschaftler und Biologen im Allgemeinen auf eine Darstellung der Unterschiede zwischen Fishers Ansatz und dem von Neyman-Pearson und präsentiert eine Art Hybrid-Ansatz: Neyman-Pearson (1933) wird die H_0 entsprechende Hypothese verworfen, wenn $T(\mathbf{x}) \in R$, und $P(T(\mathbf{x}) \in R|H_0) = \alpha$, d.h.

$$\alpha = P(T(\mathbf{x}) \in R) = \int_R g(T|H_0)dT. \tag{68}$$

Andererseits ist

$$p = P(T > T(\mathbf{x})|H_0) = \int_{T(\mathbf{x})}^{\infty} g(T|H_0)dT,$$

so dass für $T(\mathbf{x}) \in R$ bzw. $T(\mathbf{x}) > \inf\{R\}$ die Beziehung $p \leq \alpha$ folgt. Führt man nun die Regel ein, dass für $p \leq \alpha$ die Hypothese H_0 verworfen wird, so hat man durch impliziten Bezug auf Neyman-Pearson die Definition des p-Wertes gerechtfertigt. Andererseits ist diese Interpretation weder im Sinne Fishers, der ja den p-Wert als Maß der Evidenz gegen H_0 verstanden haben wollt, und gegen die Interpretation im Rahmen des NP-Modell spricht, dass R gar nicht im Sinne der Minimierung des β-Fehlers bestimmt wird, sondern einfach nur auf einen Wert von α Bezug genommen wird. Siehe aber Hubbard & Bayarri (2003) über die Inkompatibilität von Fisher und NP.

Konfusion von p- und α-Werten: (Hubbard & Bayarri, 2003) Hubbard et al. gehen von der Vermischung des Fisher- und des NP-Ansatzes aus. Insbesondere betrachten sie die Überinterpretation der Evidenz des p-Wertes gegen die Nullhypothese.

Fishers (1925, 1935a) Ziel war, eine objektive Art des induktiven Schließens zu erreichen. Deshalb konnte er dem Bayes-Ansatz nicht folgen, der auf eine Berechung der Posteriori-Wahrscheinlichkeiten $P(H|\mathbf{x})$ hinausgelaufen wäre, da die

Priori-Verteilungen als subjektiv gelten. Es sei möglich, "to argue from consequences to causes, from observations to hypotheses" (Fisher 1966, p. 6). p-Werte sind, so Fisher, "constitutive evidence against the null hypothesis". Die Neyman-Pearson Theorie dagegen ist explizit nicht-evidentiell. Fisher 1959, p. 100 stellte fest, dass es eine grundsätzliche Differenz in der Interpretation von Signifikanztesten gäbe: deren Uminterpretation als Entscheidung zwischen Handlungen (entscheide für A und nicht-B, oder für A und nicht-A) sei numerisch falsch, und Entscheidungen seien endgültig, während die Meinung (state of opinion), die auf der Basis von Signifikanztests gewonnen würde, vorläufig sei und revidiert werden könne.

Neyman (1950) dagegen hat Schwierigkeiten mit dem Fisherschen Induktionsbegriff. Die Bedeutung des Begriffs "inductive reasoning" bleibe obskur, es sei unklar, ob damit irgendeine Art von definiertem Begriff bezeichnet werden könne. Statt dessen könne man von *induktivem Verhalten* sprechen. Solch ein Verhalten könne dazu dienen, die Anpassung des Verhaltens an bestimmte, begrenzte Informationen zu erklären. Diese Anpassung sei partiell bewußt, partiell aber auch unbewußt. Der bewußte Teil basiere auf bestimmten Regeln (wenn ich dieses beobachte, tue ich jenes), und diese Regeln seien die des induktiven Verhaltens; im Übrigen war er der gleichen Meinung wie Popper: alles Argumentieren sei deduktiv. Dem Neyman-Pearsonschem Modell zufolge wird eine Nullhypothese in der Weise entweder akzeptiert oder verworfen, derart, dass "in the long run of experience, we shall not be too often wrong" (N& P (1933, p. 291). Hier wird der frequentistische Aspekt der NP-Theorie deutlich.

In der NP-Theorie wird der Wert von α vor der Untersuchung festgelegt. Bei Fisher dagegen wird der p-Wert erst anhand der Daten berechnet. Da diese eine Zufallsstichprobe $\mathbf{x}$ sind, gilt $p = p(\mathbf{x})$, d.h. p ist eine zufällige Veränderliche, die unter der Nullhypothese gleichförmig auf dem Intervall $[0, 1]$ verteilt ist. Während nach Fisher der p-Wert für einzelne Experimente interpretierbar ist, erlauben Neyman & Pearson keine Schlußfolgerungen in Bezug auf eine spezifische Hypothese:

> "We are inclined to think that as far as a particular hypothesis is concerned, no test based upon the theory of probability can by itself provide any valuable evidence of the truth or falsehood of that hypothesis." (N & P, 1933, p. 290)

Man kann also mit Hubbard & Bayarri (2003) folgern, dass die NP-Theorie *nicht evidentiell* ist. Das Fishersche Signifikanzniveau α ist *keine* Fehlerrate, es wird von Fisher nicht frequentistisch interpretiert (anders als die Fehler vom Typ I Wahrscheinlichkeit α bei N & P). sondern wie eine Carnapsche Wahrscheinlichkeit P_1. Viele Wissenschaftler und auch Statistiker sprechen vom p-Wert als einer "beobachteten Fehlerrate", – aber diese Beobachtung entpsricht weder den Intentionen Fishers noch Neyman & Pearsons. (s.a.Berger & Sellke (1986), Shervish (1996)).

DeGroot (1973) stellte einige interessante Beziehungen zwischen p-Wert, Likelihood-Quotienten und a posteriori-Wahrscheinlichkeiten her. Der p-Wert entspricht dem Integral

$$P(T \geq T_0|H_0) = \int_{T_0}^{\infty} f(x|h_0)dx = 1 - F(T_0).$$

DeGroot (1973) hat gezeigt, dass sich $1 - F(T_0)$ im Prinzip als eine a posteriori-Wahrscheinlichkeit für H_0 interpretrieren läßt. Es sei f_θ eine Klasse von Wahrscheinlichkeitsdichten, die durch den Parameter θ indiziert sind, $\theta \in \Theta$. Für einen

beobachteten Wert x läßt sich dann der Likelihood-Quotient

$$\lambda(x) = \frac{f(x)}{\sup_{\theta\in\Theta} f_\theta(x)} \tag{69}$$

definieren. Es sei $\lambda = .001$. Dies bedeutet, dass ein Parameterwert θ existiert, für die die Likelihood von x 100 mal größer ist als unter der Hypothese H_0.

Es läßt sich nun eine Familie von Dichten definieren, für die $\lambda(x) = 1 - F(x)$. Es sei Θ derart, dass $0 < F(\theta) < 1$ für alle $\theta \in \Theta$. Dann sei

$$f_\theta(x) = \begin{cases} f(x)/(1 - F(\theta)), & x \geq \theta \\ \\ 0, & x < \theta \end{cases} \tag{70}$$

Demnach ist f die Dichte einer zufälligen Veränderlichen Y, und f_θ ist die bedingte Dichte von Y, gegeben $Y \geq \theta$. K_1 sei die Klasse der alternativen Dichten f_θ; mit $\theta \in \Theta$. Es sei weiter F_θ die zu f_θ korrespondierende Verteilungsfunktion. Dann gilt

$$F_\theta(x) \leq F(x), \quad -\infty < x < \infty$$

und für $\theta_1 < \theta_2$ ist $F_{\theta_2} \leq F_{\theta_1}(x)$. Dies bedeutet, dass die Dichten aus K_1 *stochastisch größer* sind als die Dichten f (vergl.Lehman (1994)). Für einen beobachteten Wert x hat man nun

$$\lambda(x) \sup_{\theta\in\Theta} f_\theta(x) = \frac{f(x)}{\inf_{\theta\leq x}[1 - F(\theta)]} = \frac{f(x)}{1 - F(x)}. \tag{71}$$

Dann folgt $\lambda(x) = 1 - F(x)$. Fürdie Klasse K_1 von Algternativdichten ist also der Likelihood-Quotient gleich $1 - F(x)$, gegeben H_0.

Es sei nun $\xi(\theta)$ die a priori-Verteilung für θ.Die durchschnittliche Likelihood für x, gegeben eine Dichte aus K_1, ist nun

$$L_\theta(x) = \int_\Theta f_\theta(x)\xi(\theta)d\theta.$$

Der Likelihood-Quotient $\lambda_\xi(x)$ für die a priori-Dichte ξ ist dann

$$\lambda_\xi(x) = \frac{f(x)}{L_\xi(x)}. \tag{72}$$

Es sei $b = \inf\Theta$; dann folgt

$$\lambda_\xi(x) = \left[\int_b^x \frac{\xi(\theta)}{1 - F(\theta)}d\theta\right]^{-1} > 1 - F(x). \tag{73}$$

Tatsächlich kann also, abhängig von der gewählten a priori-Verteilung für $\theta < x$, $\lambda_\xi(x)$ deutlich größer sein als $1 - F(x)$, korrespondierend zu dem allgemeinen Befund der Bayesschen Statistik, dass ein Wert $1-F(x) = .01$ zu einem Likelihood-Wert größer als .01 korrespondiert und deswegen weniger gegen H_0 spricht, als der Wert .01.

Die Konstruktion von $1-F$ als a posteriori-Verteilung erfordert die Einführung einer weiteren Klasse von apriori-Verteilungen; die Darstellung kann hier übergangen werden.

Casella & Berger (1987) zeigen ebenfalls, wie frequentistische und Bayessche Methoden zusammengebracht werden können; Voraussetzung sei, dass der Nullhypothese H_0 und der Alternativhypothese H_1 gleiche a priori-Wahrscheinlichkeiten zugeordnet werden müssen. Als Bayessches Maß für die Evidenz für H_0 nehmen die Autoren $\inf P(H_0|x)$ an, wobei sich inf auf eine Klasse von a priori-Verteilungen bezieht. Die Autoren kommen zu etwas anderen Resultaten als Berger & Selke (1987), Sellke, Bayarri & Berger (2001). Die Details gehen allerdings zu sehr auf Fragen der Wahl von a priori-Verteilungen ein, als dass sie an dieser Stelle ausführlich behandelt werden können.

Rozeboom (1960)[9] stellte eine Reihe von Punkten zur Problematik von Signifikanztests zusammen:

1. Wissenschaft ist keine Akkumulation von Entscheidungen über Hypothesen, sondern eine systematisierte Menge von wahrscheinlichem Wissen. Das Endprodukt einer wissenschaftlichen Untersuchung ist ein Grad an Konfidenz in eine bestimmte Menge von Aussagen, und diese Menge dient als Basis für Entscheidungen,

2. Als Entscheidungstheorie sei der Signifikanztest "elendiglich inadäquat"; für eine effektive Entscheidung man müsse einerseits die Wahrscheinlichkeit der Hypothese, gegeben die Daten, kennen, als auch den Nutzen der verschiedenen Resultate der Entscheidungen. Aber Signifikanztests stellen Nutzenfragen nicht in Rechnung. Weiter wird die inverse Wahrscheinlichkeit einer Hypothese, also $P(H|D)$ nicht berücksichtigt, weshalb der Test auch nicht als Entscheidungshilfe funktioniert.

3. Es werden nur zwei Alternativen betrachtet: Bestätigung oder Nichtbestätigung der Nullhypothese. Darüber hinaus ist der Übergang von Bestätung zu Nichtbestätigung *unstetig:* eine winzige Änderung des Wertes von $T(\mathbf{x})$ kann die Beurteilung als "signifikant" oder "nicht signifikant" ändern. Der Punkt, an dem dieser Übergang geschieht, ist willkürlich: die Theorie des Signifikanztests impliziert nicht den Wert: das 95%-Niveau, oder das 94%- oder das 96%- oder auch das 99%-Niveau.

4. Der Signifikanztest impliziert einen starken Bias für *eine* aus einer großen Menge von Alternativen. Wird die Stichprobe aus einer Population mit unbekanntem Erwartungswert $\mathbb{E}(X) = \mu$ gezogen, so gibt es unendlich viele alternative Nullhypothesen über den Wert von μ. Die Wahl einer spezielen Hypothese gibt dieser Hypothese einen Vorteil gegenüber den anderen Hypothesen. Rozeboom spricht von einem *double standard*: die begünstigte Hypothese werde gewissermaßen für unschuldig gehalten, bis sie für "schuldig" befunden wird, und alle übrigen Hypothesen gelten so lange als "schuldig", bis keine Wahl mehr bleibt, sie als "unschuldig" zu deklarieren.

   Rozeboom hebt eine wichtige Implikation dieses Sachverhaltes hervor: die differentielle Gewichtung der Hypothesen sei der alles-oder-nichts-Effekt einer *persönlichen Entscheidung*, und darüber hinaus impliziert die Methode, dass die Bevorzugung einer Methode notwendiger Bestandteil der Methode ist. In der klassischen Methode der inversen Wahrscheinlichkeit erhält jede Hypothese – via Priori-Verteilung – ein individuelles Gewicht, dass die

[9] "one can hardly avoid polemics when butchering a sacred cow"

Glaubhaftigkeit der Hypothese aufgrund anderer als der erhobenen Daten repräsentiert.

5. Am Ende fragt Rozeboom, welcher Wissenschaftler denn jemals eine Hypothese aufgegeben habe, nur weil ein Signifikanztest für die Daten *eines* Experiments dies nahegelegt hätte, und welcher Wissenschaftler "in his right mind" es als einen beachtenswerten Unterschied halte, wenn die Daten $p = .04$ implizieren oder wenn die Daten $p = .06$ implizieren, wenn als kritischer Wert vorher $P_c = .05$ für eine Entscheidung festgelegt wurden. Da die Daten $\mathbf{x}$ zufällig sind, ist auch $p = p(\mathbf{x})$ eine zufällige Veränderliche, und der evidentielle Unterschied zwischen $p = .04$ und $p = .06$ ist zu gering, um Entscheidung für oder gegen eine Hypothese zu rechtfertigen.

Rozeboom hegt offenbar Sympathien für einen Bayesschen Ansatz der Hypothesenevaluation, hat aber eine Reihe offener Fragen, z.b. ob die Wahrscheinlichkeiten, mit denen man große Theorien wie etwa die Spezielle Relativitätstheorie für wahr hält, mit den Wahrscheinlichkeiten, die man etwa beim Münzwurf betrachtet, in ein und demselben System betrachten kann. Das ist eine offensichtlich wichtige Frage, auf die später zurückgekommen wird.

Rejection Trials: Royall's 1997/2000 "illogic of rejection trials": Es sei $\text{Bin}(n, \theta)$ eine Binomialverteilung mit den Parametern $p = \theta$ und n. H_0: $\theta = 1/2$. X Anzahl der "Erfolge". H_0 wird auf dem Niveau α verworfen, wenn $P(X \geq x|H_0) \leq \alpha/2$. Nun sei H_0: $\theta \leq 1/2$. Der Befund x ist hinreichend starke Evidenz gegen H_0 wenn $P(X \geq x|H_0) \leq \alpha$. Damit hat man: wenn

$$\frac{\alpha}{2} \leq P(X \geq x|H_0) \leq \alpha,$$

so hat man hinreichend starke Evidenz, um die zusammengesetzte Hypothese

$$H_0\text{: } \theta \leq 1/2 \equiv \theta < 1/2 \text{ or } \theta < 1/2$$

zurückzuweisen, nicht aber die einfache Hypothese

$$H_0\text{: } \theta = 1/2.$$

Man kann also schließen: weder A noch B, nicht aber: $\neg A$. Royall: dieser Befund sei "odd". Alternative Formulierung: Zurückweisung von H_0: $\theta = 1/2$, wenn entweder $\theta < 1/2$ oder $\theta > 1/2$. Dieser Sachverhalt gilt insbesondere, wenn Evidenz für $\theta > 1/2$. Wenn also die Evidenz zu dem Schluß, A sei wahr, führt, so sollte dies ebenfalls die Aussage: 'Entweder A oder B ist wahr' rechtfertigen. (Vergl. Paradoxon of the Raven nach Hempel). Aber ein Rejection Trial erlaubt diese Folgerung nicht! (Royall, p. 77).

4 Neyman-Pearson Tests und Evidenz

4.1 Allgemeine Charakteristika von Neyman-Pearson-Tests

Der Fishersche Signifikanztest fokussiert auf eine Hypothese, die Nullhypothese H_0; H_0 wird verworfen, wenn die Wahrscheinlichkeit p, dass Teststatistik $T(\mathbf{x}) \geq T_c$ bei Geltung von H_0 hinreichend klein ist, etwa $p \leq .05$; für $T(\mathbf{x}) < T_c|H_0$ wird H_0 vorläufig beibehalten.

Schon Gossett, der Erfinder des t-Tests, argumentierte, dass die Likelihood der Daten unter H_0 sehr klein sein könne, ohne dass dies ein Beweis gegen die Gültigkeit von H_0 sei, der Zufall könne es nun einmal so wollen, dass auch bei Gültigkeit von H_0 die Likelihood klein sei. Erst im Vergleich mit einer Alternativhypothese H_1 könne man sich für oder gegen H_0 entscheiden (vergl. Hacking (1965), p. 83). H_1 kann aber nicht einfach $\neg H_0$ sein, weil $\neg H_0$ oft unspezifiziert sei. Hacking (p. 89) rät dann auch: " Don't reject something unless you have something better." Diese Überlegung war der Ausgangspunkt für Neyman & Pearson, den Fisherschen Signifikanztest zu verbessern:

> "The problem of testing a statistical hypothesis occurs when circumstances force us to make a choice between two courses of action: either take step A or step B." (Neyman (1950), p. 258, zitiert nach Royall (1997/2000, p. 35)

Am Ende resultierte mit ihrem Hypothesentest eine Alternative zum Fisherschen Signifikanztest und nicht nur eine Verbesserung. Eine ausführlichere Darstellung des Neyman-Pearsonschen Ansatzes findet man im Anhang, Abschnitt 7.1, p. 128.

Das Schema des Hypothesentests: Es wird angenommen, dass die Entscheidung für die Handlungen A oder B von der unbekannten Wahrscheinlichkeitsverteilung einer zufälligen Veränderlichen X abhängen: A wird gewählt, wenn die Verteilung zu einer aus einer bestimmten Menge von Verteilungen gehört, und B, wenn sie zu einer anderen Menge gehört; eine 'statistische Hypothese' wird demnach getestet, wenn $X = x \in R$ auf A führt, und $X = x \notin R$ auf B. Dabei korrespondiert A zu einer Hypothese H, und B zu einer Hypothese $\bar{H}$:

> "The choice between the two actions A and B is interpreted as the *adoption* or the *acceptance* of one of the hypotheses H or $\bar{H}$ and the *rejection* of the other. Thus, if the application of an adopted rule ... leads to action A, we say that the *hypothesis H is accepted* (and therefore the hypothesis $\bar{H}$ is rejected). On the other hand, if the application of the rule leads to action B, we say that the *hypothesis H is rejected* (and, therefore, the hypothesis $\bar{H}$ is accepted). Frequently it is convenient to concentrate our attention on a particular one of the two hypotheses H and $\bar{H}$. To do so, one of them is called the *hypotheis tested*. The outcome of the test is then reduced to either accepting or rejecting the hypothesis tested. Plainly itis immaterial which of the two hypotheses H and $\bar{H}$ is labelled the hypothesis tested." (Neyman, 1950, p. 259)

Neyman warnt vor weitergehenden Interpretationen: es handele sich eben nur um Entscheidungen für eine der Handlungen A oder B, aber nicht notwendig darum, H oder $\bar{H}$ für wahr zu halten, wenn eine Hypothese akzeptiert wird, oder sie für falsch zu halten, wenn sie zurückgewiesen wird.

Die Bewertung einer Entscheidungsregel habe dementsprechend in Bezug auf die Fehlerwahrscheinlichkeiten α und β zu erfolgen. Ein Test ist gut dann, wenn diese Wahrscheinlichkeiten klein sind. Haben zwei Tests die gleiche Typ I-Fehlerwahrscheinlichkeit α, so ist derjenige besser, der die kleinere Typ II-Fehlerwahrscheinlichkeit β hat. Das Neyman-Pearsonsche Fundamentallemma (1933) zeigt, wie man

im Falle zweier einfacher Hypothesen für vorgegebenen Wert von α den Test mit dem kleinsten β-Wert findet.

Das Testen von Hypothesen wird insbesondere in wissenschaftstheoretischen Diskussionen oft mit der Frage nach der Induktion in Verbindung gebracht. Neyman formuliert dazu:

> 'If a [decision] rule unambiguously prescribes the selection of action for each possible outcome ..., then it is a rule of inductive behaviour', (Neyman 1950, p. 10)

und damit definiert Neyman die Bedeutung des Ausdrucks 'Mathematische Statistik':

> "Mathematical statistics is a branch of the theory of probability. It deals with problems relativn to performance characteristics of rules of indcutive behaviour based on random experiments." (Neyman 1950, p. 11)

Wesentlich am Neyman-Pearsonschen Ansatz ist, dass nicht nur eine Hypothese (H_0), sondern eine Hypothese gegen eine andere getestet wird. Selbst wenn Die Daten $\mathbf{x}$ unter der Hypothese H_0 sehr unwahrscheinlich sind, ist damit H_0 nicht widerlegt, denn unwahrscheinliche Ereignisse sind keine unmöglichen Ereignisse. Der Vergleich der Likelihoods der Daten in Bezug auf zwei verschiedene Hypothesen erlaubt die Evaluation der Gültigkeit der Hypothesen relativ zueinander.

Die Rede von der Evaluation der Gültigkeit ist einerseits vage, andererseits soll nicht zu früh eine spezielle Interpretation der Daten suggeriert werden. Die Notwendigkeit, eine Entscheidung treffen zu müssen – als Therapeut über eine anzuwendende Therapie, als Manager über eine Investition, als Sportler über ein Trainingsprogramm, als Musiker über anzuschaffendes Instrument, etc – zwingt zu einer anderen Interpretation als die Situation eines Wissenschaftlers, dessen primäres Interesse wenn nicht die Wahrheit, so doch die Angemessenheit einer Hypothese ist.

Die Frage ist, was die Forderung nach einer Entscheidung bedeutet. Es widerspricht der wissenschaftlichen Praxis, sich aufgrund eines experimentellen Befundes, d.h. eines Datensatzes, für eine und gegen eine andere Hypothese zu entscheiden. Man hält allenfalls die eine Hypothese, gegeben die Daten, für wahrscheinlicher als die andere. Formal ist diese Einschätzung natürlich falsch, so lange die Betrachtung sich nur auf den Vergleich der Likelihoods bezieht, da eine Likelihood ja gerade nicht die Wahrscheinlichkeit einer Hypothese, gegeben $\mathbf{x}$, bedeutet.

Bei der Einführung des Neyman-Pearson-Ansatzes definiert Lehmann (1986, p. 1) die *statistical inference* als diejenige Methodik, die aus Beobachtungen der Werte zufälliger Veränderlicher X Informationen über die Verteilungen von X und deren Parameter erlangt. Die Notwendigkeit statistischer Analyse ergibt sich, nach Lehmann, aus der Tatsache, dass die Verteilung von X unbekannt sei. Daraus ergebe sich Ungewißheit bezüglich der optimalen Weise, sich zu verhalten, d.h. sich für die beste Art, sich zu entscheiden. Es müsse eine Regel δ gefunden werden, die die Bestimmung der Entscheidung erleichtere bzw. optimiere. Für gegebene Daten $\mathbf{x}$ soll dann δ eine Regel d liefern, $d = \delta(\mathbf{x})$, nach der verfahren werden soll. Ist die Verteilung von X durch P_θ gegeben, so ist mit jedem "Wert" d von δ ein *Verlust* $L(\theta, d) \geq 0$ verbunden. Betrachtet wird dann der durchschnittliche Verlust, der

sich bei Anwendung von d "in the long run" ergibt, wenn P_θ die wahre Verteilung von $\mathbf{x}$ ist. Der durchschnittliche Verlust ist die Erwartung $R(\theta, \delta) = \mathbb{E}[L(\theta, \delta\mathbf{x})]$; $R(\theta, \delta)$ heißt auch *Risikofunktion* von δ. Das Ziel der Statistik, so Lehmann, sei demnach die Auswahl einer Entscheidungsfunktion, die das resultierende Risiko (die Risikofunktion) minimiere.

Die Lösung eines Testproblems besteht in der Wahl einer Regel, nach der zwischen Handlungsalternativen entschieden wird. Diese Regel definiert das 'induktive Verhalten' und wird nach ihren 'performance charakteristics' bewertet, d.h. nach ihren probabilistischen Eigenschaften. Diese wiederum werden spezifiziert durch die Eigenschaften, die sie 'in the long run' im Durchschnitt zeigen. Diese Eigenschaften beziehen sich auf die Schätzungen von Parametern: ein 'Schätzer' (estimator) ordnet jeder Beobachtung x eine Schätzung $\hat{\theta}$ eines Parameters θ zu. Die Eigenschaften eines Schätzers sind (i) der erwartete Fehler (Bias), repräsentiert durch $\mathbb{E}[t(X) - \theta]$, (ii) die Varianz $\mathbb{V}(t)$, (ii) der erwartete quadrierte Fehler $\mathbb{E}[t(X) - \theta]^2$. Darüber hinaus wird mit jedem x ein Konfidenzintervall definiert, $(\hat{\theta}_u, \hat{\theta}_0)$, $P(\hat{\theta}_u < \theta < \hat{\theta}_o)$, mit der erwarteten Breite $\mathbb{E}[\hat{\theta}_o - \hat{\theta}_u]$.

Es ergibt sich die Frage, ob es ein bestes Verfahren für die Schätzung eines Parameters gibt. Dazu hat Royall (1997/2000, p. 39) ein Urnenbeispiel konstruiert. Geschätzt werden soll der Anteil θ weißer Kugeln in einer Urne. X sei die Anzahl weißer Kugeln in einer Stichprobe. Man kann etwa drei Schätzungen betrachten:

1. Es sei $t_1(X) = X/n$. Dann ist der Bias

$$\mathbb{E}(X/n - \theta) = \frac{1}{n}\mathbb{E}(X) - \frac{1}{n}\theta = \frac{1}{n}(\theta - \theta) = 0,$$

   d.h. die Schätzung X/n ist biasfrei oder erwartungstreu. Für die Varianz und den MSE ergibt sich

$$\mathbb{V}(t_1) = \mathrm{MSE}(t_1) = \theta(1 - \theta)/n.$$

2. $t_2(X) = 1/2$. Der Schätzer ist unabhängig von X. Der erwartete Fehler ist

$$\mathbb{E}(t_2(X) - \theta) = \frac{1}{2} - \theta,$$

   und die Varianz ist

$$\mathbb{V}(t_2(X)) = \mathbb{E}[(t_2(X) - \mathbb{E}(t_2(X))^2] = \mathbb{E}[(1/2 - 1/2)^2] = 0,$$

   und

$$\mathrm{MSE}(t_2) = (1/2 - \theta)^2.$$

3. $t_3(X) = (1 - w)t_1(X) + w1/2$, mit $w = 1/(1 + \sqrt{n})$. Dieser Schätzer ist ein Kompromiss zwischen t_1 und t_2.

   Diese Schätzung hat den Bias, Varianz und quadrierten Fehler

$$\mathbb{E}[t_3(X) - \theta) = w(1/2 - \theta),\ \mathbb{V} = \theta(1 - \theta)w^2,\ \mathrm{MSE}(t_3) = (w/2)^2.$$

Will man die Schätzer bewerten, so findet man, dass t_1 bezüglich des Bias der beste ist; für t_1 ist er gleich Null. Bezüglich der Varianz ist t_2 der beste, und bezüglich des quadrierten Fehlers MSE ist t_1 der beste, wenn (für $n = 10$) $\theta < 1.7$ oder $\theta > .83$, t_2 ist der beste für $.38 < t_2 < .62$ und sehr schlecht sonst, und t_3 ist der beste für $.17 < \theta < .38$ und $.62 < t_3 < .83$. Offenbar ist die Definition eines 'besten Schätzers' keine einfache, da nicht eindeutig zu beantwortende Frage.

4.2 Optimaler Stichprobenumfang und Evidenz

Der NP-Theorie zufolge kann für vorgegebenen Wert von α und vorgebenen Wert von β ein Bereich R bestimmt werden derart, dass für eine geeignet definierte Statistik $T(\mathbf{X})$ und $\mathbf{X} = \mathbf{x}$

$$P(T(\mathbf{x}) \in R|H_1) = \alpha, \quad P(T(\mathbf{x}) \in R^c|H_2) = 1 - \beta.$$

R^c ist das Komplement zu R. Damit diese Bedingungen erfüllt sind, muß der Stichprobenumfang einen bestimmten, minimalen Wert n_s haben (s für **S**tichprobenumfang), d.h. man kann diese Bedingungen benutzen, um n_s zu bestimmen. Diese Bestimmung des Stichprobenumfangs wird in vielen Lehrbüchern empfohlen, um zu optimalen Entscheidungen über Hypothesen zu gelangen. Das Verfahren wird kurz vorgestellt und anhand einiger Befunde Royalls (1997) diskutiert.

Es sei etwa $X \sim N(\theta, \sigma^2)$, und H_1: $\theta = \mu$, H_2: $\theta = \mu + \delta$. $\mathbf{X}$ sei eine Stichprobe vom Umfang n; das arithmetische Mittel $\bar{X}$ ist eine Schätzung für θ mit der Varianz $\mathbb{V}(\bar{X}) = \sigma^2/n$. Es sei $\bar{x}_c$ ein kritischer Wert, d.h. es gelte $P(\bar{X} > \bar{x}_c|H_1) = \alpha$, so dass $R = \{\bar{X}|\bar{X} > \bar{x}_c\}$. Man hat unter H_1

$$z_{1-\alpha} = \frac{(\bar{x}_c - \mu)\sqrt{n}}{\sigma}, \quad z_\beta = \frac{\bar{x}_c - (\mu + \delta))\sqrt{n}}{\sigma},$$

oder

$$\bar{x}_c = z_{1-\alpha}\frac{\sigma}{\sqrt{n}} + \mu = z_\beta\frac{\sigma}{\sqrt{n}} + \mu + \delta,$$

so dass

$$(z_{1-\alpha} - z_\beta)\frac{\sigma}{\sqrt{n}} = \delta,$$

so dass und es folgt

$$n = n_s = (z_{1-\alpha} - z_\beta)^2\left(\frac{\sigma}{\delta}\right)^2 = (z_{1-\alpha} + z_{1-\beta})^2\left(\frac{\sigma}{\delta}\right)^2 \tag{74}$$

$\varepsilon = \delta/\sigma$ ist auch als *Effektgröße* bekannt; offenbar wird n_s für vorgegebene Werte von α und β durch die Effektgröße ε bestimmt.

Die Frage ist, wie oft eine Untersuchung bezüglich H_1 und H_2 Evidenz etwa für H_2 liefert. Notwendig für eine solche Evidenz ist

$$\lambda(\bar{x}) = \frac{P(\bar{x}|H_2)}{P(\bar{x}|H_1)} > k,$$

wobei k hinreichend groß sein muß, damit von 'großer Evidenz' gesprochen werden kann. Der Likelihood-Quotient ist

$$\frac{\exp(-(\bar{x} - \theta - \delta)^2 n/2\sigma^2)}{\exp(-(\bar{x} - \theta)^2 n/2\sigma^2)} = \exp[((\bar{x} - \theta)^2 - (\bar{x} - \theta - \delta)^2))n/2\sigma^2]$$

Hieraus folgt[10]

$$\lambda(\bar{x}) = \exp[(\bar{x} - (\theta + \delta/2))n\delta/\sigma^2]. \tag{75}$$

[10] Es ist

$$(\bar{x} - \theta)^2 = (\bar{x} - \theta - \delta + \delta)^2 = (\bar{x} - \theta - \delta)^2 + \delta^2 + 2\delta(\bar{x} - \theta - \delta),$$

also

$$(\bar{x} - \theta)^2 - (\bar{x} - \theta - \delta)^2) = \delta^2 + 2\delta(\bar{x} - \theta - \delta) = 2\delta(\bar{x} - \theta) + \delta^2,$$

und da $2\delta(\bar{x} - \theta) + \delta^2 = 2\delta\bar{x} - (2\theta + \delta)\delta$, ergibt sich (75).

Tabelle 6: Wahrscheinlichkeiten für unerwünschte Ergebnisse, wenn der Stichprobenumfang n so bestimmt wird, dass für $\alpha = .05$ der Fehler II-ter Art $\beta = .20$ bzw $\beta = .05$. $P_1(\bar{x}) \geq k)$: große Evidenz für die falsche Hypothese, $P_2(\bar{x}) < k)$: keine große Evidenz für die wahre Hypothese.

	$P_1(\bar{x}) \geq k)$		$P_2(\bar{x}) < k)$	
k	$\beta = .20$	$\beta = .05$	$\beta = .20$	$\beta = .05$
8	.019	.011	.342	.156
16	.009	.006	.449	.212
32	.004	.003	.560	.277

Nun ist $\lambda > k$ wenn $\log \lambda > \log k$, so dass

$$(\bar{x} - (\theta + \delta/2))n\delta/\sigma^2 > \log k. \tag{76}$$

Dies ist die Bedingung für 'große Evidenz' für H_2. Gesucht ist nun die Wahrscheinlichkeit, mit der die Bedingung (76) erfüllt ist. Dazu muß diese Ungleichung in eine äquivalente Ungleichung umgeformt werden, auf deren einen Seite eine zufällige Veränderliche mit bekannter Wahrscheinlichkeitsverteilung steht. Man findet

$$\frac{n^{1/2}(\bar{x} - \theta)}{\sigma} > \frac{\delta\sqrt{n}}{2\sigma} + \frac{\sigma}{\delta\sqrt{n}} \log k; \tag{77}$$

offenbar ist $(\bar{x}-\theta)\sqrt{n}/\sigma$ der standardisierte $\bar{x}$-Wert, so dass $(\bar{x}-\theta)\sqrt{n}/\sigma \sim N(0,1)$ gefolgert werden kann.

In Gleichung (74) wurde ein Ausdruck für den Stichprobenumfang $n = n_s$ gegeben, der nötig ist, um eine Power $1-\beta$ zu erhalten. Setzt man diesen Ausdruck in (77) ein, so ergibt sich nach kurzer Rechnung

$$\frac{\delta\sqrt{n}}{2\sigma} + \frac{\sigma}{\delta\sqrt{n}} \log k = \frac{1}{2}(z_{1-\alpha} + z_{1-\beta}) + \frac{\log k}{z_{1-\alpha} + z_{1-\beta}} = \frac{c}{2} + \frac{\log k}{c} \tag{78}$$

mit $c = z_{1-\alpha} + z_{1-\beta}$. Für die Wahrscheinlichkeit, dass der Likelihood-Quotient größer k ist, also eine Entscheidung für H_2 getroffen werden soll, obwohl H_1 wahr ist, erhält man dann

$$P_1(\lambda(\bar{x}) \geq k) = 1 - \Phi\left(\frac{c}{2} + \frac{\log k}{c}\right). \tag{79}$$

Diese Wahrscheinlichkeit für eine Entscheidung für H_2, wenn H_2 wahr ist, ist

$$P_2(\lambda(\bar{x}) \geq k) = 1 - \Phi\left(\frac{\log k}{c} - \frac{c}{2}\right). \tag{80}$$

Dann ist die Wahrscheinlichkeit, dass für H_1 entschieden wird, wenn H_2 wahr ist,

$$P_2(\lambda(\bar{x}) < k) = \Phi\left(\frac{\log k}{c} - \frac{c}{2}\right). \tag{81}$$

Die Tabelle 6 gibt die Wahrscheinlichkeiten (79) und (81) sowohl für $\beta = .20$ wie für $\beta = .05$ unter der Annahme, dass $n = n_s$ gewählt wurde. Die ersten

beiden Spalten zeigen, dass die Erzeugung irreführender Evidenz zumindest für die ausgewählten k-Werte von kleiner Wahrscheinlichkeit ist.

Andererseits zeigen die dritte und die vierte Spalte der Tabelle, dass die Beziehung (74) offenbar zu kleine n_s-Werte liefert, denn die Wahrscheinlichkeit einer Entscheidung für H_1, obwohl H_2 korrekt ist, ist relativ hoch. Man kann sagen, dass der Wert von n_s keine hinreichend große Evidenz für H_2 garantiert, wenn H_2 wahr ist. Für $\beta = .20$ wird in mehr als 1/3 der Fälle ($P(\cdot) = .342$) Evidenz für H_1 suggeriert, obwohl H_2 korrekt ist. Für $\beta = .05$ ist die Wahrscheinlichkeit gleich .156, dass keine Evidenz für H_2 erzeugt wird, wenn H_2 wahr ist: Der Wert ist dreimal so groß als mit der Voraussetzung $\alpha = \beta = .05$ angenommen wird!

Royall (1997), p. 53, diskutiert die Möglichkeit, dass die k-Werte in der Tabelle 6 einfach zu radikal gewählt wurden und zeigt, dass dies *nicht* der Fall ist; die Details können hier übergangen werden. Der Punkt der Betrachtungen ist, dass die NP-Theorie zwar stets zu Entscheidungen über H_1 und H_2 führt, die evidentielle Interpretation dieser Entscheidungen aber keineswegs auch adäquat sein muß.

4.3 Neyman-Pearson-Entscheidungen und Evidenz

4.3.1 Evidentielle Anwendungen eines Tests

Gerade bei wissenschaftlichen Untersuchungen ist man weniger daran interessiert, zu welcher Entscheidung die NP-Theorie führt, sondern zu wissen, 'was die Daten (bezüglich der Hypothesen) sagen'. Man nennt deswegen oft nicht nur das Resultat der Entscheidung, sondern weitere Aspekte, die zu der Entscheidung geführt haben: etwa die Größe α des Tests und, wenn möglich, die Power $1 - \beta$ des Tests, die Schätzung des Standardfehlers der Schätzung eines Parameters und/oder das Konfidenzintervall. Dies sind die *probabilistischen* Eigenschaften des Verfahrens, das zur Entscheidung über die H_i geführt hat, und diese Eigenschaften sind Ausdruck der 'Evidenz' in den Daten. Die 'Zurückweisung' einer Hypothese bedeutet nur, dass man in den Daten Evidenz sieht, die gegen die Hypothese spricht. Die Fehlerwahrscheinlichkeiten α und β als Maß für diese Evidenz interpretiert. Konfidenzintervallen wird ebenfalls eine 'evidentielle" Bedeutung zugemessen; manche Autoren empfehlen, die Intervalle für mehrere Konfidenzkoeffizienten $(1 - \alpha)100$ anzugeben: etwa für das 80%-, das 90%-, das 95%- und das 99%-Intervall. Diese Interpretationen sind streng genommen außerhalb der NP-Theorie, werden aber gleichwohl von Vertretern dieser Theorie nahegelegt: im kanonischen NP-Lehrbuch Lehmanns (1986/1994), p. 4, liest man, dass Konfidenzintervalle 'indicate, what information is avaiblable concerning the unknown parameter', und Neymann (1976) schreibt:

> my own preferred substitute for "do not reject H" is "no evidence against H is found".

Es muß also ein Unterschied gemacht werden zwischen dem statistischen Verfahren einerseits und 'evidentiellen' Anwendungen des Verfahrens. In der Tabelle 7 werden NP-Verfahren und ihre jeweiligen Resultate sowie die entsprechenden evidentiellen Interpretationen zusammengestellt. Die NP-Theorie reicht gewissermaßen nur bis zur Spalte 'Resultat', die Spalte 'Evidenz' gibt die Interpretation dessen, "was die Daten sagen". So sei $\delta(x) = 1$. Dann wird dies in der Anwendung als "die Daten sind Evidenz für H_1 und gegen H_2" interpretiert. Die Frage ist, ob $\delta(x) = 1$ auch

Tabelle 7: NP-Verfahren und evidentielle Interpretation für $X = x$

Verfahren	Eigenschaft	Resultat	Evidenz
Konfidenzintervall $I_\theta(X) = (\theta_u(X), \theta_o(X))$	$P(\theta_u < \theta < \theta_o) = 1 - \alpha$	$I_\theta(x)$	$x \in I_\theta$ Evidenz für $\theta \in I_\theta$
Schätzung $\hat{\theta}(X)$	$\mathbb{E}(\hat{\theta}(X)) = \theta$ $\mathbb{V}(\hat{\theta}_\theta) = \sigma^2$	$\hat{\theta}(x)$	x ist Evid., dass θ nahe bei $\hat{\theta}(x)$
Hypthesentest $H_\delta(X)$	Typ I $\to \alpha$ Typ II $\to \beta$	$H_\delta(x)$ $(\delta(x) = 1 \Rightarrow H_1$ $\delta(x) = 2 \Rightarrow H_2)$	x Evid. für $H_\delta(x)$ α, β klein,

wirklich diese Interpretation zuläßt. In analoger Weise läßt sich fragen, ob für eine gegebene Schätzung $\hat{\theta}$ die Interpration "θ liegt nahe bei $\hat{\theta}$" sinnvoll ist, und ob $x \in I_\theta$ tatsächlich den Schluß nahelegt, dass $\theta \in I_\theta$.

Neyman & Pearson (1933) haben im Falle einfacher Hypothesen argumentiert, dass für vorgegebenen Wert von α ein Ablehnungsbereich R gefunden werden kann derart, dass für $\mathbf{x} \in R$ mit maximaler Wahrscheinlichkeit $1 - \beta$ die Entscheidung für H_2 korrekt ist. Der Bereich R wird dabei durch den Likelihood-Quotienten gemäß

$$R = \left\{ \mathbf{x} \,\middle|\, \lambda(\mathbf{x}) = \frac{P(\mathbf{x}|H_1)}{P(\mathbf{x}|H_2)} \leq k \right\}$$

bestimmt; die Regel ist: "Entscheide für H_1, wenn $\lambda(\mathbf{x})$ ist mindestens gleich k, sonst für H_2", und der Wert von k wird durch den von α bestimmt.

Beispiel 4.1 Zur Illustration werde ein Versuch mit $n = 30$ Bernoulli-Versuchen durchgeführt, d.h. es werde $\text{Bin}(\theta, n)$ betrachtet, und es sei H_1: $\theta = 3/4$ versus H_2: $\theta = 1/4$. X sei die Anzahl von "Erfolgen". Der Likelihood-Quotient ist dann für $X = x$

$$\lambda(x) = \frac{\binom{30}{x}(3/4)^x(1/4)^{30-x}}{\binom{30}{x}(1/4)^x(3/4)^{30-x}} = 3^{2x-30}.$$

Es sei

$$r(j) = \sum_{k=j}^{30} (1/4)^k (3/4)^{30-k}.$$

Man findet $r(12) \approx .0507$, d.h. $P(X \geq 12|H_1) \approx .00507$. Dies bedeutet $\lambda(x) = \lambda(12) = \lambda_\alpha = 1/729 = .00137$. Für $X > 12$ ist $\lambda(x) \leq \lambda_\alpha$.

Nach dem NP-Lemma entscheidet man sich für H_2, wenn die Evidenz H_2 relativ zu H_1 favorisiert. Im betrachteten Fall ist der kritische Faktor λ kleiner als 1. Man betrachte nun den Likelihood-Quotienten für $X = 12$. Man findet $\lambda(12) = 729$, d.h. die Evidenz für H_1 ist 729-mal so groß wie die für H_2, obwohl man sich nach dem NP-Lemma für H_2 entscheiden soll! Weiter findet man $\lambda(13) = 81$, $\lambda(14) = 9$, d.h für $X = 13$ und $X = 14$ hat man immer noch starke Evidenz für H_1, obwohl nach dem NP-Lemma für H_2 entschieden würde, denn $x \in R = \{k|k \geq 12\}$. Für $x = 15$ schließlich findet man $\lambda(15) = 1$, d.h. beide Hypothesen erscheinen als gleich

wahrscheinlich, und die (Maximum-Likelihood-)Schätzung ist $\hat{\theta}_{ML} = 15/30 = 1/2$, liegt also gerade in der Mitte zwischen 1/4 und 3/4. □

Das Beispiel zeigt, dass die NP-Theorie einerseits auf dem Likelihood-Quotienten basiert, aber zu anderen Entscheidungen führt als eine Betrachtung, die sich nur am Wert des Likelihood-Quotienten für $X = x$ orientiert. Für eine Reihe von Werten aus R signalisiert das NP-Lemma eine Entscheidung für H_2, während die Daten H_1 favorisieren. Das NP-Lemma minimiert die Fehlerwahrscheinlichkeit β für vorgegebenen Wert von α "in the long run", der Likelihood-Quotient repräsentiert die Evidenz für eine Hypothese anhand der vorliegenden Daten.

Ein anderes Problem der Dateninterpretation wird erzeugt, wenn Randomisierung eine Rolle spielt (s.a. Abschnitt 7.1.1, Seite 128, Anhang). Ein weiteres Beispiel illustriert das Problem:

Beispiel 4.2 Eine Urne enthalte weiße und schwarze Kugeln. θ sei der Anteil weißer Kugeln. Es gebe zwei Hypothesen: $H_1\colon\ \theta = 1/2$, und $H_2\colon\ \theta = 3/4$. Es werden nun fünf Kugeln gezogen und jeweils nach Identifizierung der Farbe wieder zurückgelegt. X sei die Anzahl der weißen Kugeln, die gezogen worden sind. Da die Stichprobe mit Zurücklegen gebildet wird, spielt die Gesamtzahl der Kugeln in der Urne keine Rolle. Also ist

$$P(X = 5|H_1) = \left(\frac{1}{2}\right)^5 = \frac{1}{2^5} = \frac{1}{32} = .03125.$$

Würde man aufgrund des Befunds $X = 5$ die Hypothese H_1 *verwerfen*, so hätte man einen Fehler erster Art der Größe $\alpha = .03125$. Weiter ist

$$P(X = 4|H_1) = \binom{5}{4}\left(\frac{1}{2}\right)^4\left(\frac{1}{2}\right) = 5\left(\frac{1}{2}\right)^5 = \frac{5}{32} = .15625.$$

Führte man die Regel ein, gegen H_1 zu entscheiden, wenn $X = 4$ oder $X = 5$, so hätte man

$$\alpha = P(X = 4|H_1) + P(X = 5|H_1) = \frac{6}{32} = .1875.$$

Für $X = 5$ ist der zugehörige α-Wert kleiner als .05 ($\alpha \approx .03$), für $R = \{4, 5\}$ ist $\alpha \approx .19 > .05$. Für einem solchen Fall führen Neyman & Pearson den randomisierten Test ein. Demnach wird H_1 zurückgewiesen, wenn (i) $X = 5$, und (ii) *manchmal*, wenn $X = 4$. Dazu wählt man ein "Gewicht" g derart, dass

$$\alpha = P(X = 5|H_1) + gP(X = 4|H_1) = .05.$$

Löst man die Gleichung nach g auf, so findet man

$$g = \frac{.05 - P(X = 5|H_1)}{P(X = 4|H_1)} = .12.$$

Die Entscheidungsregel lautet nun: Weise H_1 zurück, wenn $X = 5$. Ist $X = 4$, so wähle eine Zufallszahl $\xi \in (0, 1)$. Ist $\xi \leq g$, so entscheide ebenfalls gegen H_1. Nach dem NP-Lemma existiert kein besserer Test für gegebenen Wert von $\alpha = .05$, der einen kleineren β-Wert hat. □

Die hier erklärte Regel ist optimal im Sinne der NP-Theorie; *auf lange Sicht*, d.h. bei hinreichend häufiger Wiederholung des Experiments, führt der randomisierte

Test zu einer falschen Ablehnung von H_1 mit der Wahrscheinlichkeit α und minimaler Wahrscheinlichkeit β, H_2 irrtümlich abzulehnen. Andererseits stelle man sich vor, man habe das Experiment durchgeführt und $X = 4$ gefunden. Dieser Befund ist die Evidenz bezüglich der Hypothesen. Es stellt sich die Frage, warum eine vom experimentellen Befund und damit von der Evidenz unabhängige Zufallszahl ξ eine Entscheidung bezüglich H_1 – und damit auch H_2 – beeinflussen soll, warum man sich also im Falle $\xi \leq g$ gegen H_1, andernfalls für H_1 entscheiden soll.

Ein zweites Beispiel liefert eine weitere Illustration:

Beispiel 4.3 Gelegentlich hängen die Beobachtungen, die man macht, von einem nicht weiter zu kontrollierenden Zufallsfaktor ab: mit einer Wahrscheinlichkeit p macht man Beobachtungen oder Messungen der Art A, und mit der Wahrscheinlichkeit $1-p$ macht man Beobachtungen der Art B. In beiden Fällen sollen aus den Beobachtungen Schlußfolgerungen bezüglich eines Parameters θ gezogen werden. Cox (1958) hat hierfür eine Formalisierung gegeben, die von Royall (1997, 46) in der folgenden Weise modifiziert bzw. vereinfacht wurde.

Demnach wird das Experiment in zwei Stufen ausgeführt. In Stufe I wird ein Zufallsexperiment, etwa ein Münzwurf, durchgeführt. Liegt "Kopf" oben, so wird im zweiten Teil eine $N(\theta, \sigma^2)$-verteilte Variable X gemessen. Liegt die "Zahl" oben, so werden k unabhängig und identisch verteilte Variable $X_1, \ldots, X_k$ beobachtet, mit $X_j \sim N(\theta, \sigma^2)$. Der Wert von σ^2 sei bekannt. Gewünscht ist das kürzeste 95 %-Konfidenzintervall für den unbekannten θ-Wert. Es gibt zwei Vorgehensweisen:

A: Für den Fall "Kopf" ist für $X = x$ das Intervall durch $x \pm 1.96\sigma$ gegeben, und im Fall "Zahl" durch das Intervall $\bar{x} \pm \sigma/\sqrt{k}$, wobei $\bar{x}$ das arithmetische Mittel der $x_1, \ldots, x_k$ ist.

B: Es sei $k = 100$. Für den Fall "Kopf" werde nun das Intervall $I_1 = (x \pm 1.68\sigma)$ bestimmt, und im Falle "Zahl" das Intervall $I_2 = (\bar{x} \pm 2.72\sigma/10)$. Die Wahrscheinlichkeit, dass I_1 den Parameter θ enthält (überdeckt), ist .91, und für I_2 ist die Wahrscheinlichkeit .99.

Die Überdeckungswahrscheinlichkeit beim Vorgehen **A** ist

$$\frac{1}{2}.95 + \frac{1}{2}.95 = .95.$$

Beim Vorgehen **B** hat man

$$\frac{1}{2}.91 + \frac{1}{2}.99 = .95,$$

d.h. beie Vorgehensweisen liefern eine durchschnittliche Wahrscheinlichkeit von .95, dass das jeweils bestimmte Intervall den Parameter θ überdeckt. Nun kann aber noch die erwartete Breite der Intervalle betrachtet werden. Im Falle **A** hat man

$$\frac{1}{2}1.96\sigma + \frac{1}{2}1.96/10 = 2.1\sigma.$$

Im Fall **B** hat man

$$\frac{1}{2}1.68\sigma + \frac{1}{2}2.72\sigma/10 = 1.95\sigma,$$

d.h. im Falle des Vorgehens nach **B** erhält man im Durchschnitt ein kürzeres Intervall! Dies gilt für alle $k > 8$.

Das Problem: Im Falle **B** und "Kopf" etc

Es sind wieder die Standpunkte (i) *in the long run* versus (ii) der konkrete Fall. Warum soll der Zufall bestimmen, wie im konkreten Fall vorzugehen ist? □

Beispiel 4.4 Ein Test habe kleine Fehlerwahrscheinlichkeiten α und β und habe in einer speziellen Anwendung zu einer Entscheidung für H_2 relativ zu H_1 geführt. Die Frage ist, ob diese Information Evidenz für die Gültigkeit von H_2 (relativ zu H_1) bedeutet. Das Argument ist, dass das, was üblicherweise geschieht, auch tatsächlich eingetreten ist. Nach dem Likeliood-Prinzip könnte so argumentiert werden.

Royall (1997), p. 49, argumentiert, dass in diesem Fall *nicht* von Evidenz für H_2 gesprochen werden könne. Denn es werden ja nicht die Daten $X = x$ beurteilt, sondern nur eine Indikatorfunktion $Z(x)$:

$$Z(x) = \begin{cases} 1, & x \in R, \ (\rightarrow H_1), \\ \\ 2, & x \in \bar{R}, \ (\rightarrow H_2), \end{cases} \qquad P(x \in R|H_1) = \alpha, \ P(x \in \bar{R}|H_2) = 1 - \beta,$$

Der Übergang von x zu $Z(x)$ bedeutet eine Datenreduktion, bei der Evidenz verloren geht. Ist nun $Z(x) = 2$, so hat man den Likelihood-Quotienten

$$\lambda(Z) = \frac{P(Z = 2|H_2)}{P(Z = 2|H_1)} = \frac{1 - \beta}{\alpha}. \tag{82}$$

Die Evidenz $Z(x) = 2$ für H_2 bedeutet noch nicht, dass auch $X = x$ Evidenz für H_2 ist. Dazu betrachte man den Fall einer Wahrscheinlichkeitsverteilung $f(X|\theta)$, und H_1: $\theta = \theta_1$, H_2: $\theta = \theta_2$, $\theta_1 \neq \theta_2$. Für $X = x$ gelte

$$f(x|\theta_1) = f(x|\theta_2) = \frac{1}{20}.$$

Dann ist

$$\frac{f(x|\theta_1)}{f(x|\theta_2)} = 1$$

und x ist gegenüber beiden Hypothesen neutral.

Nun werde die zusammengesetzte Hypothese $H_c : \theta = \theta_1 \vee \theta_2$ betrachtet. Dann ist ebenfalls

$$f(x|\theta_1 \vee \theta_2) = 1/20,$$

und der Likelihood-Quotient für H_1 versus H_c ist

$$\frac{f(x|H_1)}{f(x|\theta_1 \vee \theta_2)} = 1,$$

d.h. die mit dem Likelihood-Quotienten verbundene Evidenz ist keine besondere Evidenz für H_c, obwohl H_c die Hypothese H_1 logisch impliziert. Royall (1997), p. 16, argumentiert, dass man es hier mit zwei Arten von Evidenz zu tun hat: die eine ist statistisch und besteht in der Beobachtung $X = x$, die andere ist logisch, nämlich $H_c \Rightarrow H_1$. H_c ist zwar "glaubwürdiger" (weil allgemeiner), aber dieser Sachverhalt bedeutet offenbar nicht, dass die statistische Evidenz für H_c größer als die für H_1 ist. Das Gesetz der Likelihood bezieht sich nur auf die statistische, nicht auf die logische Evidenz. □

Tabelle 8: Howsons Urne

	Farbe		
Zahl	s	w	$\sum$
0	0	949	949
1	1	50	51
$\sum$	1	999	1000

4.3.2 Evidenz und Grundquote: Howsons Urne

Nach Neyman & Pearson ist ein Test durch die Fehlerraten $\alpha = P(T > T_c|H_0)$ und $\beta = P(T \leq T_c|H_1)$ bestimmt, wobei $T = T(\mathbf{X})$ eine auf den möglichen Stichproben definierte Statistik ist, und $T_c = T(\mathbf{x})$ ist die in einer Untersuchung erhobene Stichprobe $\mathbf{x} = (x_1, \ldots, x_n)$; wird der Stichprobenumfang so gewählt, dass für gewählten kleinen α-Wert der β-Wert ebenfalls hinreichend klein ist, so hat man bei häufiger Wiederholung des Experiments entsprechend kleine Fehlerraten. Es ist oft angemerkt worden, dass Tests dieser Art wohl für Qualitätskontrollen angemessen sind, selten aber für wissenschaftliche Fragestellungen, denn die Anzahl der Wiederholungen eines Experiments ist im Allgemeinen gering. Der Test orientiert sich an den Ereignissen $A = \{T > T_c\}$ bzw. $\neg A = \{T \leq T_c\}$, aber diese sind nicht notwendig Evidenz für eine Hypothese und gegen die andere. Wie sehr ein nach den Prinzipien Neyman & Pearson konstruierter Test in die Irre führen kann, wird zunächst an einem von Howson konstruierten Urnenmodell illustriert. Das Modell wird dann auf übliche Untersuchungen übertragen und diskutiert.

Urnenmodell von Howson (1997, 2000) betrachtet: gegeben sei eine Urne mit N Kugeln, von denen 1 schwarz und die anderen weiß seien. Die Kugeln repräsentieren Personen, weiß bedeutet "gesund" (H_0), schwarz bedeutet "krank" (H_1). Weiter sind die Kugeln mit jeweils einer Zahl beschriftet, die das Resultat eines Tests signalisieren: eine 0 zeigt ein negatives Resultat an, d.h. die Person ist dem Test zufolge nicht krank, und ein 1 zeigt ein positives Resultat an, d.h. die Person ist dem Test zufolge krank. Tests sind üblicherweise nicht perfekt, und so gibt es Fehler: falsch-positive, wenn eine 1 resultiert, obwohl die Person gesund ist, und falsch-negative, wenn also eine 0 resultiert, obwohl die Person krank ist. Konkret gelte: $N = 1000$, $N_s = 1$ schwarze Kugel, $N_w = 999$ weiße Kugeln, auf 50 weißen Kugeln steht eine 1, und auf der einzigen schwarzen Kugel steht ebenfalls eine 1. Es gibt also keine falsch-negativen Fälle, aber 50 falsch-positive. Wird zufällig eine Kugel gezogen, so erhält man in 50 von 999 Fällen eine Kugel, die weiß ist und damit eine gesunde Person repräsentiert, aber eine 1 zeigt und damit durch den Test falsch kategorisiert wird. In einem von 1000 Fällen erhält man die schwarze Kugel, die die kranke Person repräsentiert und deren Krankheit durch den Test korrekt angezeigt wird. Offenbar hat man

$$P(1|H_0) = \alpha = 50/999 \approx .05, \quad P(0|H_1) = \beta = 0. \tag{83}$$

Es werde zunächst angemerkt, dass nach dem Bayesschen Satz die Gleichungen

$$P(H_0|1) = P(1|H_0)\frac{P(H_0)}{P(1)} \tag{84}$$

$$P(H_1|0) = P(0|H_1)\frac{P(H_1)}{P(0)} \tag{85}$$

gelten. Wegen $\beta = P(0|H_1) = 0$ ist hier $P(H_1|0) = 0$. Die a-priori-Wahrscheinlichkeiten $P(H_0)$ und $P(H_1)$ sowie $P(0)$ und $P(1)$ haben hier offenbar eine *objektive Bedeutung*; sie sind die Wahrscheinlichkeiten, zufällig auf eine gesunde bzw. auf eine kranke Person bzw. auf eine positiv oder negativ klassifizierte Person zu treffen.

Man kann die a-posteriori-Wahrscheinlichkeiten als Maß der Evidenz für H_0 bzw. H_1 auf der Basis des Testergebnisses betrachten. Für ein negatives Testresultat erhält man

$$P(H_0|0) = \frac{P(H_0 \wedge 0)}{P(0)} = \frac{949}{949} = 1, \quad P(H_1|0) = 1 - P(H_0|0) = \beta = 0.$$

Die Evidenz "0" führt zu der korrekten Einschätzung, dass die Person gesund ist.

Andererseits erhält man für $P(H_1|1)$

$$P(H_1|1) = P(1|H_1)\frac{P(H_1)}{P(1)} = \frac{P(H_1)}{P(1)} = \frac{1}{51} \approx .02, \tag{86}$$

aus $\beta = P(0|H_1) = 0$ folgt $P(1|H_1) = 1 - \beta = 1$. Korrespondierend dazu erhält man

$$P(H_0|1) = 1 - P(1|H_1) \approx .98. \tag{87}$$

Ein positives Testergebnis ("1") spricht trotz $\beta = 0$ mit unerwartet *geringer* Wahrscheinlichkeit ($\approx .02$) *für* die Erkrankung der getesteten Person, und mit hoher Wahrscheinlichkeit ($\approx .98$) gegen die Erkrankung. Die Fehlerraten α und β implizieren keineswegs eine adäquate Interpretation der Evidenz; diese hängt, wie (86) zeigt, stark von den a-priori-Wahrscheinlichkeien bz. Grundquoten ab.

Man mag argumentieren, dass hier ein extremes Beispiel vorgestellt wurde, das die allgemeine Gültigkeit des Neyman-Pearsonschen Tests nicht wirklich in Frage stelle. Darüber hinaus sei es bei wissenschaftlichen Untersuchungen, bei denen zwei konkurrierende Hypothesen zur Diskussion stehen, ja nicht der Fall, dass H_0 und H_1 mit bestimmten Anteilen zuträfen: entweder sei H_0 wahr oder H_1. Es ist gleichwohl nicht schwer, das Modell auf reale Situationen zu übertragen.

Übertragung des Urnenmodells auf Standarduntersuchungen:[11] Es werde eine Variable X gemessen, für die das Modell

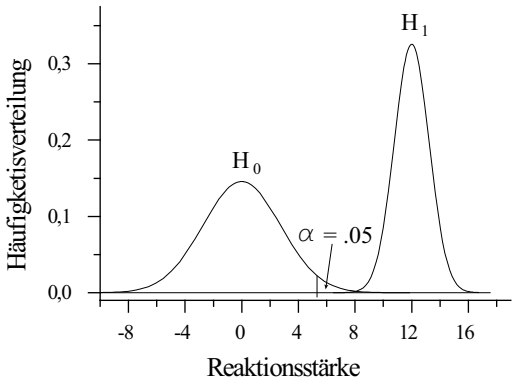
$$X = \mu + \varepsilon \tag{88}$$

angenommen wird; der Erwartungswert von X sei $\mathbb{E}(X) = \mu$, sei. Varianz sei $\mathbb{V}(X) = \sigma^2$. Es soll entschieden werden, ob $H_0 \colon \mu = 0$ oder $H_1 \colon \mu = 12$ gilt. Die Verteilung der Messwerte sei wie in Abb. 5: unter H_0 ist $\mu = 0$, $\sigma_0^2 = 7.5$, unter H_1 sei $\mu = 12$, $\sigma_1^2 = 1.5$. Der Einfachheit halber sei $T(\mathbf{X}) = X$, und $\alpha = P(X > x_c|H_0) = .05$ mit $x_c = 5.37$, aber $\beta \approx 0$; dies sind die Fehlerwahrscheinlichkeiten, die in Howsons Urnenbeispiel angenommen wurden. Gilt also H_1, so ist die Wahrscheinlichkeit, dass Messwerte $X < x_c$ auftreten, praktisch gleich Null.

In vielen Untersuchungen sind die Elemente der Stichprobe (Versuchs-)Personen, Mäuse, oder Zellen (etwa Neurone), etc. Jeder Messung x_i entspricht ein Element e_i, $i = 1, 2, \ldots, n$ der Stichprobe; der Fall, dass mehrere Messungen an einem

[11] Diese Übertragung wurde nicht von Howson vorgeschlagen, – eventuelle Fehler sind also nicht ihm anzulasten!

Abbildung 5: Messwerteverteilungen mit Howsons Fehlerraten



Elemente vorgenommen werden, werde hier der Einfachheit halber vernachlässigt. Im einfachsten Fall gilt $\mathbb{E}(x_i) = \mu$ für alle i, wenn alle Elemente unter den gleichen Bedingungen gemessen werden; realistischer ist vermutlich der Fall, dass $\mathbb{E}(x_i) = \mu_i \neq \mu$. Dann kann trotzdem (88) angenommen werden, wenn nämlich die Differenzen $\delta_i = \mu_i - \mu$ in die Fehler ε_i absorbiert werden: werden die Elemente zufällig aus der Population gezogen, können die δ_i als zufällige Größen betrachtet werden.

Es werde nun angenommen, dass die Messungen unter einer experimentellen Bedingung erhoben werden, die den Test H_0 versus H_1 erlaubt; hat die Bedingung keinen Einfluss auf die Messungen, soll H_0: $\mu = 0$ gelten. Der Effekt der Bedingung sei so, dass Messwerte $< x_c|H_1$ mit extrem kleiner Wahrscheinlichkeit – also $\beta \approx 0$ – auftreten. Es wird zunächst die Indikatorfunktion

$$\chi = \begin{cases} 0, & x < x_c \\ \\ 1, & x \geq x_c. \end{cases} \tag{89}$$

eingeführt. Zur Vereinfachung wird $P(0|H_0)$ statt $P(\chi = 0|H_0)$, $P(1|H_0)$ statt $P(\chi = 1|H_0)$ etc geschrieben, ebenso $P(0)$ statt $P(\chi = 0)$, $P(1)$ statt $P(\chi = 1)$. Es gelte also, entsprechend dem Howsonschen Urnemmodell,

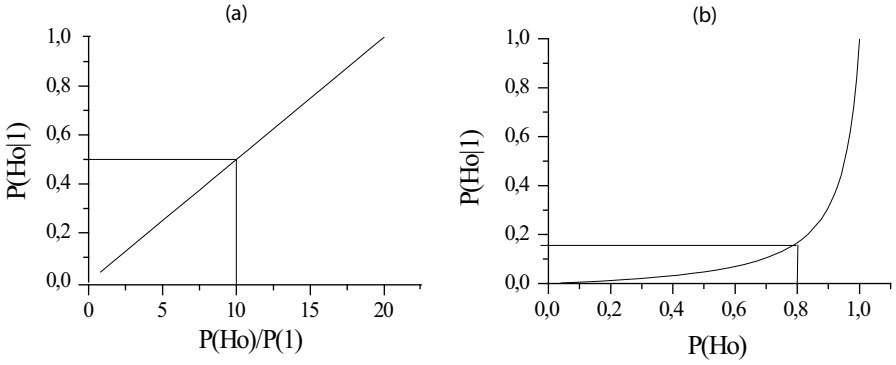
$$P(1|H_0) = \alpha = .05, \quad P(0|H_1) = \beta \approx 0. \tag{90}$$

Das "≈"-Zeichen wurde geschrieben, weil z.B. im Fall der Normalverteilung die Annahme $\beta = 0$ kaum streng gelten kann. Dann gilt jedenfalls

$$P(H_0|1) = P(1|H_0)\frac{P(H_0)}{P(1)} = \alpha\frac{P(H_0)}{P(1)}, \quad P(H_1|1) = 1 - P(H_0|1). \tag{91}$$

Der Howson-Fall $P(H_0|1) = .98$, $P(H_1|1) = .02$ ist dann gegeben, wenn $.98 = \alpha P(H_0)/P(1)$, d.h. wenn $P(H_0)/P(1) = 19.6$. Abb. 6 zeigt $P(H_0|1)$ als Funktion des Quotienten $P(H_0)/P(1)$. Für $P(H_0)/P(1) > 10$ ist $P(H_0|1) > .5$ und $P(H_1|1) < .5$. Die Bedingung $\beta \approx 0$ kann als erfüllt angesehen werden, wenn die Verteilung von X derart ist, dass stets $P(X < x_c|H_1) = P(0|H_1) \approx 0$ gilt. Die

Abbildung 6: (a) $P(H_0|1)$ als Funktion von $P(H_0)/P(1)$. $P(H_0)$ Anteil der Personen/Mäuse/Zellen mit $\mu = \mu_0$, $P(1)$ Anteil der Messungen $\{x \geq x_c\}$. $P(1|H_0) = \alpha = .05$. (b) $P(H_0|1)$ als Funktion von $P(H_0)$ unter der Bedingung, dass $\alpha = .05$ und $\beta \approx 0$; siehe Text für weitere Erläuterung.



Bedingung $\beta \approx 0$ kann aber auch "eingebaut" werden: Nach dem Satz der Totalen Wahrscheinlichkeit gilt

$$P(1) = P(1|H_0)P(H_0) + P(1|H_1)P(H_1) = P(1|H_0)P(H_0) + 1 - P(0),$$

da $P(1|H_1) = 1$ wegen der Forderung $\beta = P(0|H_1) = 0$ und $P(1) = 1 - P(0)$. Daraus folgt

$$\frac{P(H_0)}{P(1)} = \frac{P(H_0)}{1 - (1-\alpha)P(H_0)},$$

so dass $P(H_0|1)$ in der Form

$$P(H_0|1) = \frac{\alpha P(H_0)}{1 - (1-\alpha)P(H_0)},$$

also als Funktion von $P(H_0)$ ausgedrückt werden kann; sie wird in Abb. 6 (b) gezeigt. Für $P(H_0) > .8$ und $P(H_0|1) = .17$ ($P(H_1) < .2$) geht $P(H_0|1)$ rapide gegen 1, d.h. aber, die Evidenz für H_1 wird unterdrückt.

Entscheidungen nach Neyman & Pearson entsprechen offenbar dann nicht der Evidenz in den Daten, wenn der Anteil der Elemente einer Stichprobe, die nicht auf die experimentelle Bedingung reagieren, groß ist im Vergleich zum komplementären Anteil derjenigen Elemente, die auf die Bedingung reagieren.

Ein Beispiel sind Untersuchungen zur Wirksamkeit einer Therapie. Es ist denkbar, dass eine Therapie bei einem Patienten nur mit einer bestimmten Wahrscheinlichkeit, die der a-priori-Wahrscheinlichkeit $P(H_1)$ entspricht, wirksam ist, oder dass die Therapie bei der Mehrzahl der Patienten gar nicht anspricht, bei einer kleinen Minderheit aber perfekt wirkt. "Wirksam" soll hier bedeuten, dass, wenn X den Effekt der Therapie mißt, stets nur Werte von X größer als x_c auftreten. ($\beta = 0$). Ist $P(H_1)$ klein, so wird die Entscheidung anhand der Daten einer Versuchsgruppe gegen H_1 und für H_0 ausfallen. Für die Krankenversicherung ('quality control') ist das in Ordnung, die Therapie wird nicht zugelassen und die Versicherung muß Behandlungen mit der Therapie nicht bezahlen. Für die Wissenschaft

ist die Entscheidung aber nicht in Ordnung, zumindest solange sie an den in der Therapie ablaufenden Prozessen interessiert ist. Denn folgt man der Entscheidung, so geht man üblicherweise vom Modell $x = \mu + \varepsilon$ aus mit $\mu = \mu_0$ *für alle* Patienten; dass es einen Anteil von Patienten mit $\mu = \mu_1$ gibt, für die überdies $P(0|H_1) \approx 0$ gilt, bei denen die Therapie also in jedem Fall anschlägt, wird von der Neyman-Pearson-Entscheidung unterschlagen.

Die Relevanz dieses Sachverhalts für die Wissenschaft werde am Beispiel einer HIV-Therapie illustriert. Bei einem großen Anteil $P(H_0)$ der Bevölkerung wirke die Therapie nicht, bei einem kleinen Anteil $P(H_1) = 1 - P(H_0)$ aber wirke sie perfekt. Bei diesem Anteil wirken bestimmte Nebenbedingungen, die in der Untersuchung nicht weiter kontrolliert worden sind, z.B. eine bestimmte genetische Disposition, bestimmte Lebensumstände, etc. Könnten diese Nebenbedingungen auch für den Rest der Bevölkerung wirksam gemacht werden, hätte man eine perfekte Therapie. Die Ausrichtung an einem Neyman-Pearson-Test unterdrückt diese Möglichkeit.

Ein weiteres, für viele provokantes Beispiel ist die Hypothese H_1, dass Telepathie existiert. X ist hier ein objektives Maß, etwa von Muskelspannnung; untersucht wird, ob es eine telepathische Beeinflussung der Muskelspannung gibt. Denkbar ist, dass nur ein kleiner Teil von Personen über telepathische Fähigkeiten verfügt, überdies ist denkbar, dass die Fähigkeit nicht permanent, sondern nur in selten auftretenden Situationen vorhanden ist. Untersuchungen, die Daten liefern, die dem Modell (88) entsprechend interpretiert werden unter der Zusatzannahme, dass für alle Personen einer Stichprobe konstant entweder $\mu = \mu_0$ oder $\mu = \mu_1$ gilt, werden in der überwiegenden, wenn nicht in der Gesamtzahl der Fälle in Entscheidungen zugunsten H_0 – es gibt keine Telepathie – münden. Könnte man den Nachweis führen, dass Telepathie tatsächlich gelegentlich vorkommt, so käme dies einer wissenschaftlichen Revolution gleich. Die Anwendung von Neyman-Pearson-Tests stellt sicher, dass diese Revolution nicht stattfindet.

Wie man es anstellt, Personen, die auf eine Therapie ansprechen oder die zumindest gelegentlich telepathische Fähigkeiten zeigen, aus der Gesamtpopulation herauszufiltern, ist eine andere Frage, die an dieser Stelle nicht beantwortet werden muß. Es wurde nur gezeigt, dass kleine Fehlerraten ($\alpha = .05$, $\beta = 0$) nicht nur nicht notwendig zu korrekten Entscheidungen führen, sondern systematisch relevante Evidenz unterdrücken können.

Das Argument überträgt sich auch auf den Fall, dass nur eine Person (Maus, Zelle, etc) untersucht wird, wie es oft in psychophysischen Untersuchungen der Fall ist. Die Annahme, dass stets $\mu = \mu_0$ oder $\mu = \mu_1$ gilt und unterschiedlliche Messwerte nur wegen der zufälligen Störungen ε zustande kommen, muß nicht gelten: denkbar ist ja, dass die Person (Maus, Zelle, etc) nur in einem bestimmten Prozentsatz der Fälle (d.h. Messungen) im Zustand $A : \mu = \mu_1$ ist, und im Rest der Fälle im Zustand $\neg A : \mu = \mu_0$ ist. Dies ist der Fall, wenn die gemessene Merkmalsausprägung über die Zeit fluktuiert und der Verlauf der Ausprägung den Trajektorien eines stochastischen Prozesses entsprechen. Auch deterministische Fluktuationen könnten eine Rolle spielen; die Ausprägung des Merkmals hängt dann von den Messzeitpunkten ab, und auch konstante Zeiten zwischen den Messungen können dann eine Folge von A- und $\neg A$-Zuständen implizieren, die zufällig wirkt. Die "Mess"fehler ε und die Effekte, die eine Person entweder in den Zustand A oder $\neg A$ bringen, überlagern sich und Entscheidungen insbesondere gegen H_1 transportieren nicht die Evidenz, die die Daten zugunsten H_1 enthalten.

Es ist hier nur mit Anteilen argumentiert worden; es wäre noch zu zeigen, dass

sich die Argumentation auf die Anwendung von Teststatistiken $T(\mathbf{x})$ überträgt, mit $P(T(\mathbf{x}) \geq T_c|H_0) = \alpha$, $P(T(\mathbf{x}) < T_c|H_1) = \beta$. Weiter müsste die Tatsache diskutiert werden, wie sich Stichprobenfluktuationen auf die Entscheidungen auswirken, wenn man nur Schätzungen $\hat{P}(H_1)$ von $P(H_1)$ hat. Da es vorerst aber nur um grundsätzliche Betrachtungen geht, wird die Diskussion dieser Fragen zunächst einmal verschoben.

4.4 Weitere kritische Anmerkungen zur Neyman-Pearson-Theorie

Das Buch *Testing statistical hypotheses* von E. L. Lehmann gilt die kanonische Darstellung des Neyman-Pearson-Ansatzes überhaupt. Pratt (1961) hat das Buch in seinem Review der 1959-er Ausgabe ausführlich gewürdigt. Aber dann schreibt er:

> "But this book, by its very excellence, its theroughness, lucidity and precision, intensifies my growing feeling that nevertherless the theory is arbitrary, be it however "objective", and the problems it solves, however precisely it mayx solve them, are not even simplified theoretical counterparts of the real problems to which it is applied."
> Pratt (1961, 164).

Das ist ein hartes Urteil. Es istdeshalb von Interesse, die einzelnen Punkte zu betrachten, die Pratts Unbehagen ausmachen. Es beginnt damit, dass die Neyman-Pearson-Theorie jeweils auf eine Entscheidung hinausläuft, – aber eine Entscheidung ist nicht dasselbe, wie Schlußfolgerungen zu ziehen.

Pratt bemängelt, dass das Signifikanzniveau α eine einerseits zentrale, andererseits beliebige Rolle in der Neyman-Pearson-Theorie spiele, im Unterschied zur Entscheidungstheorie (etwa der von Wald). Man betrachte eine zusammengesetze Hypothese $H : \theta \in \Theta_0$. Man könnte dann daran interessiert sein, einen Typ-I-Fehler $\alpha(\theta)$, also als eine Funktion des Parameters θ, zu betrachten, und nicht nur in einem oberen Wert für α. Sei insbesondere H_0: $|\theta| > 0$. Man könnte die Werte $\alpha(-1) = .1$, $\alpha(1) = .05$ interessanter als einen UMP-Test[12] finden, bei dem $\alpha(-1) = \alpha(1) = .1$ ist.

Ein weiteres Problem ergibt sich aus der Definition von Konfidenzintervallen. Nach Lehman liefern diese Intervalle eine nützliche Zusammenfassung der Daten, bzw. zeigen die Information an, die bezüglich eines unbekannten Parameters vorliegen. Aber nach der Definition eines Konfidenzintervalles gibt ein solches Intervall nicht mehr die Wahrscheinlichkeit an, mit der der gesuchte Parameterwert innerhalb des Intervalles liegt: hat man numerische Werte in die Ausdrücke für die Intervallgrenzen eingesetzt, so können nur noch die speziellen Werte 0 oder 1 für diese Wahrscheinlichkeit angenommen werden, denn entweder liegt der Wert des Parameters innerhalb des Intervalls, oder er liegt nicht darin. Nur weiß man nicht, welcher dieser beiden Fälle zutrifft. Die Theorie des Konfidenzintervalles umgehe die Frage, welche Information über den Parameterwert vorliege.

Tests zur Überprüfung von Hypothesen werden oft für Zwecke der Inferenz durchgeführt. Die Frage ist, wie stark die Daten der jeweiligen Nullhypothese widersprechen, wenn der Test ein auf dem α-Niveau signifikantes Resultat liefert.

[12] UMP = Uniformly Most Powerful

Daα-Niveau hat verschiedene Bedeutung für verschiedene Stichprobenumfänge, – man denke an die Hypothese $\theta = .5$ versus $\theta = .6$ in einem Bernoulli-Experiment, wenn einmal der Stichprobenumfang $n = 10$ und ein anderes Mal $n = 1000$ ist. Neyman & Pearson sind aber die Testgröße α und die Power $1-\beta$ die zentralen Größen, anhand derer die Entscheidung gefällt wird. Weiter werde ein UMP-Testeiner einfachen Hypothese gegen eine einfache Alternative mit dem Niveau $\alpha = .05$ und der Power $1-\beta = .99$. Die Frage ist, wie der Fall interpretiert werden kann, dass ein Resultat *gerade signifikant* ist, denn wenn die Rollen der Hypothesen vertauscht werden, ist der Test auch gerade signifikant auf dem $\alpha = .01$-Niveau (dieses Argument zielt u. A. die Willkür, mit der eine Hypothese als H_0 und die andere als H_1 betrachtet wird). Generell gilt dann, dass, je größer die Power, desto mehr spricht ein gerade signifikantes Resultat *für die Nullhypothese*. Dann fragt sich aber, in welcher Weise ein Test Inferenzen erlaubt.

Es werde ein Experiment betrachtet, bei dem die möglichen Resultate $a, b, \ldots, z$ beobachtet werden können. Es gebe zwei mögliche 'Registrierungsgeräte' für das beobachtete Resultat: das Gerät 1 melde genau das Resultat, das Gerät 2 signalisiere nur, ob das Resultat d vorliegt oder nicht. Das Resultat sei nun d; beide Geräte signalisieren dieses Ergebnis und führen also zu den gleichen Schlußfolgerungen. Andererseits wird das Ergebnis eines Signifikanztests davon abhängen, welches Gerät benutzt wird. Andererseits sollte eine Inferenz oder auch eine Entscheidung nur vom beobachteten Ergebnis und nicht von der Art des Gerätes abhängen. Das Neyman-Pearson-Modell kann zu Inkonsistenzen führen.

Pratt hat in seinem Review nur aufgezählt, das der Neyman-Pearson-Ansatz nicht notwendig optimal ist, ist aber der Ansicht, dass es bei richtig verstandener Anwendung zu vernünftigen Ergebnissen führen kann. Damit hat er zumindest aufgezeigt, dass der Ansatz nicht als ein kanonisches, in jedem Fall anzuwendendes Modell zur Inferenz und Entscheidungsfindung gelten kann.

Andere Autoren zielen direkt auf eine Zurückweisung der Neyman-Pearson-Theorie.

Gillies Kritik: Gillies (1971) beginnt mit Poppers in dessen *Logik der Forschung* gemachten Feststellung, dass Wahrscheinlichkeitsaussagen einer strengen Falsifikation widerstehen, dass aber gute Wissenschaftler – Physiker zum Beispiel – wissen, wann sie eine solche Aussage als falsifiziert ansehen können. Daraus ergibt sich für Gillies das Problem, eine F.R.P.S. (Falsifying Rule for Probability Statements) zu finden, die dieses "Wissen" in einer transparenten Weise kodifiziert. Er erarbeitet eine solche Regel auf der Basis des Prinzips der kleinen Wahrscheinlichkeit: hat ein Ereignis ω, dass unter der betrachteten Hypothese H eine kleine Wahrscheinlichkeit p hat mit $p < p_0$, wobei p_0 geeignet gewählt werden muß, dann kann H *von einem praktischen Standpunkt aus* als falsifiziert gelten.

Zentral für den Ansatz von Neyman & Pearson ist dagegen die Forderung, dass für den Test einer Hypothese H stets eine explizit definierte Alternativhypothese zur Verfügung stehen muß. Dass diese Forderung sinnvoll ist, ist nicht schwer einzusehen: auch wenn H gilt, können Daten beobachtet werden, die unter H sehr unwahrscheinlich sind, so dass unwahrscheinliche Daten nicht *zwingend* zur Falsifikation von H herangezogen werden können. Hat man eine Alternativhypothese, die mehr mit den Daten kompatibel ist, so liegt es nahe, sich eher für diese Alternative zu entscheiden. Andererseits ist nicht klar, aus welchem Grundprinzip es zwingend folgt, dass zur Diskussion einer Hypothese stets eine Alternativhypothese zum Vergleich herangezogen werden *muß*.

So verhalten sich Wissenschaftler im Allgemeinen nach Maßgabe einer F.R.P.S., einfach weil oft gar keine explizit formulierte Alternative zur Verfügung steht. Gillies elaboriert seinen Ansatz, worauf an dieser Stelle nicht weiter eingegangen werden muß. Wichtig ist, dass der Gilliessche Ansatz nicht mit dem von Neyman & Pearson übereinstimmt, die ja eine wohldefinierte Alternative fordern, wenn eine Hypothese H getestet werden soll. Dieses von Gillies so genannte *principle of alternative hypotheses* ist es denn auch, das von Gillies einer Kritik unterzogen wird. Dabei geht es nicht darum, das *principle* grundsätzlich anzuzweifeln, schon weil die Geschichte der Wissenschaft genügend Beispiele liefert, die zeigen, dass sich das Prinzip bewährt hat. Es geht aber um die Grundsätzlichkeit und Ausschließlichkeit, mit der es von Neyman & Pearson und deren Anhängern gefordert wird.

Der Punkt ist, dass Alternativen zur betrachteten Hypothese H oft erst nach empirischen Untersuchungen zu H entwickelt werden (Gillies Beispiel: die Theorie des Kopernikus, die nach der des Ptolemäus entwickelt wurde). Neyman & Pearson müssen aber, um die Power zur vorgegebenen Size bestimmen zu können, die Alternative bereits vor der Untersuchung kennen bzw. spezifizieren. Tatsächlich hat jede(r) praktizierende Wissenschaftler(in) erfahren, dass erst ein Experiment, dessen Ergebnisse nicht oder nur wenig mit einer gegebenen Hypothese übereinstimmen, Ideen für eine Alternativhypothese entwickelten werden. Hier kann nun darauf hingewiesen werden, dass es oft nicht qualitativ andere Hypothesen sind, die im Neyman-Pearson Paradigma als Alterniven betrachtet werden, sondern nur verschiedene Parameterwerte, also etwa statt eines Wertes $theta = \mu_0$ der alternative Wert $\theta = \mu_1$. Gillies diskutiert ein Beispiel, dass von Neyman (1952) selbst gegeben wurde, – und das seiner eigenen Theorie widerspricht: ein Feld wird in kleine Quadrate unterteilt und in jedem Feld wird die Anzahl einer bestimmten Art von Larven ausgezählt. Die betrachtete Hypothese H ist, dass die Anzahlen Poisson-verteilt sind, $P(X = k) = \exp(-\lambda)\lambda^k/k!$, und der freie Parameter λ muß aus den Daten geschätzt werden. Die Hypothese kann mit einem gewöhnlichen χ^2-Test überprüft werden und wurde falsifiziert, d.h. die Daten erschienen, gegeben H, als äußerst unwahrscheinlich. Neyman entwickelte *dann* ein Alternativmodell, dass den Daten sehr viel besser entsprach. Die formalen Aspekte der Gillieschen Argumentation können hier übergangen werden, der wesentliche Punkt ist, dass die Neyman-Pearsonsche Forderung nach einer a priori formulierten Alternativhypothese kaum jemals realistisch ist. Nimmt man nun noch den Cohenschen Ansatz, eine Alternative nach Maßgabe einer irgendwie Effektgröße zu spezifizieren hinzu, so fragt sich allerdings, warum man dann nicht gleich zu einem Bayesschen Ansatz wechselt, denn einerseits enthält die Spezifikation der Effektgröße mit Sicherheit subjektive Komponenten, und andererseits erweist sich die Interpretation von a-posteriori-Wahrscheinlichkeiten als deutlich mehr evidentiell orientiert als eine dichotome Entscheidung für oder gegen eine Hypothese.

Spielmans (1973) Kritik: Versucht, über Gillies Kritik hinauszugehen und den Neyman-Pearson Ansatz, der auf der Bestimmung von *size* und *power* eines Tests (und damit eben auch auf der Forderung nach einer Alternativhypothese) beruht, als in sich invalide zu kennzeichnen. Dies hat schon Hacking (1965) getan, wenn der auch auf etwas bizarre Beispiele zur Illustration seiner Argumente zurückgriff. Nach Hacking jedenfalls sind *size* und *power* nach Durchführung des Experiments irrelevant für eine Bewertung der Entscheidung, die der Test diktiert. Kleine α- und β-Werte seien keine intrinsisch wünschenswerten Größen. Spielman will die-

ses Argument ausbauen: *size* und *power* seien 'dangerously misleading concepts', "beste Tests" im Sinne von Neyman & Pearson können Entscheidungen führen, die relativ zum Resultat des Experiments schlicht dumm sein können.

Neymans Tuberkulose-Beispiel: es soll auf das Vorliegen von Tuberkulose geschlossen werden. Dazu wird eine Anzahl von Röntgenbildern einer Person angefertigt, die in zufälliger Folge – gemischt mit den Bildern anderer Patienten – dem Arzt gezeigt werden, der ein Bild als 'positiv' oder 'negativ' klassifiziert. Annahme der Klinik: ist ein Patient frei von TBC, so ist die Wahrscheinlichkeit einer 'positiven' Klassifikation gleich .01. Ist der Patient moderat betroffen, so sei die Wahrscheinlichkeit für eine solche Klassifikation gleich .6. Kann Unabhängigkeit der Beurteilungen angenommen werden, so ist für eine TBC-freien Person die Wahrscheinlichkeit von r 'positiven' Beurteilungen binomialverteilt mit dem Parameter $\theta = .01$. Für eine moderat betroffene Person hat man eine Binomialverteilung mit dem Parameter $\theta = .6$. Die endgültige Diagnose wird von der Anzahl k 'positiver' Beurteilungen abhängig gemacht. Neyman nimmt an, dass der schwerste Fehler, der von der Klinik gemacht werden kann, derjenige ist, dass eine 'negative' falsche Diagnose getroffen wird, d.h. wenn die Hypothese, dass eine Person TBS hat, verworfen wird, obwohl sie richtig ist. Deswegen wird die Hypothese 'Person hat TBC' als Nullhypothese definiert. Die Alternativhypothese H_1 ist dann 'Person hat nicht TBC'. Nur diese beiden Hypothesen werden betrachtet, wobei die Annahme gemacht wird, dass Personen mit schwerer TBC nicht mehr zum Check-up gehen, weil sie schon in Behandlung sind. Neyman nimmt an, dass für eine Person fünf Röntgenaufnahmen gemacht werden müssen und dass eine finale 'negative' Diagnose nur dann gemacht werden sollte, wenn $k = 0$, k die Anzahl 'positiver' Urteile. Die Tabelle 9 zeigt, dass die Wahrscheinlichkeit, dass die gete-

Tabelle 9: Wahrscheinlichkeiten für TBC-Erkrankung einer Person Y

	k					
	0	1	2	3	4	5
H_0^y	.01 (size)	.077	.234	.346	.259	.078
H_1^y	.951 (power)	.048	.001	.00001	≈ 0	≈ 0

stete Hypothese zurückgewiesen wird, obwohl sie wahr ist, gleich .01 ist, und die Power ist gleich .951. Darüber hinaus gibt es keinen anderen, auf Röntgenbildern beruhenden Test, der eine kleinere Size und größere, mindestens gleiche Power hat hat. Daher ist der Test ($k = 0$) der "beste" Test dieser Art (im Sinne von Neyman & Pearson). Die Interpretation von Neyman & Pearson ist nun:

> "The interpretation [of the error probabilities of the test] in operational terms is as follows: If all the assumptions made are approximately true, and if the described procedure of multiple X-ray examinations is used by the clinic, then the final diagnosis based on five X-ray examinations will be "negative" for about one percent of all tuberculosis patients and about ninety five percent of all those who have not contracted the desease." (nach Spielman)

Die Interpretation bezieht sich auf das, was "in the long run" geschieht. Wie Spielman ausführt, haben Neyman & Pearson nie genauer ausgeführt, was genau darunter zu verstehen ist. Das Problem ist, dass bei dem Beispiel diese Argumentation

auf ein Individuum bezogen werden muß. Hat das Individuum TBC, so heißt dies, dass es in 1 % der Fälle als TBC-frei diagnostiert wird, und in 95 % der Fälle als TBC-frei diagnostiert wird, wenn es nicht unter TBC leidet. Die Frage ist aber, diese Art von Test in der Tat zu "besten" Tests führt.

Nach Spielman muß gefragt werden, ob ein Test wie der Röntgen-Test *reliabel* genug ist, um das Risiko einer möglicherweise falschen Entscheidung zu rechtfertigen. Die Reliabilität eines Tests muß genauer definiert werden. Dazu wird zunächst eine Referenzklasse von Personen definiert, etwa alle Erwachsenen,die zhu einer bestimmten sozio-ökonomischen Schicht in einer bestimmten Region gehören. Weiter sei y der Prozentsatz der Personen, für die H_0^y zutrifft, d.h. die TBC in zumindest nachweisbarer Form haben. z sei der Prozentsatz der Personen, die keine TBC haben, für die also H_1^y zutrifft. Es sei nun $n_0 = .01y + .951z$ die erwartete An-

Tabelle 10: Durchschnittliche Anzahl Personen per 100 mit TBC bzw ohne TBC; $y \rightarrow$TBC, $z \rightarrow$ keine TBC

	k					
	0	1	2	3	4	5
H_0	$.01y$ (size)	$.077y$	$.234y$	$.346y$	$.259y$	$.078y$
H_1	$.951z$ (power)	$.048z$	$.001z$	$.00001z$	≈ 0	≈ 0

zahl von Personen, für die $k = 0$ ist. n_0 setzt sich also zusammen aus denen, die tatsächliche keine TBC haben und für die $k = 0$ gefunden wurde, und denen ($= .01y$), die unter TBC leiden, für die aber gleichwohl $k = 0$ gefunden wurde. Der Anteil dieser Personen an der Gesamtzahl der Personen mit $k = 0$ ist

$$\frac{.01y}{.01y + .951z}.$$

Die *Reliabilität* des Tests ist dann

$$\rho = 1 - \frac{.01y}{.01y + .951z}. \tag{92}$$

Für $y = 0$ folgt $\rho = 1$, d.h. die Entscheidung gegen die Hypothese, dass TBC bei einer Person vorliegt, ist perfekt reliabel genau dann, wenn es gar keine Erkrankten gibt. Umgekehrt sei $z = 0$; dann sind alle erkrankt, und man hat $\rho = 0$, d.h. die Entscheidung, es liege keine TBC vor, ist stets falsch. Selbst wenn man nun $z = 100 - y$ annimmt – was man der Neyman-Pearson Theorie zufolge tun müßte – könnte die Reliabilität einer Zurückweisung der Hypothese, dass TBC vorliegt (also der Entscheidung, dass keine TBC vorliegt) nicht berechnet werden, ohne dass y bekannt ist. Eine kleine *size* des Tests (.01 im gegebenen Fall) langt also nicht aus, um reliable Entscheidungen zu fällen, – es sei denn, die *size* ist gleich Null. Der Wert der *power* spielt dabei gar keine Rolle. Die *size* eines Tests und seine Reliabilität können völlig verschiedene Werte haben!

Spielman elaboriert diesen Befund und liefert weitere Beispiele, auf die hier aber verzichtet werden kann. Es genügt, noch einmal festzuhalten, dass der Anspruch Neymans & Pearsons, dass *size* und *power* hinreichende Größen für 'induktives Verhalten' sind, nicht allgemein gelten kann. Die 'Evidenz' eines Befundes läßt sich durch diese beiden Größen nicht unter allen Umwständen ausdrücken.

5 Likelihood-basierte Dateninterpretationen

5.1 Das Prinzip

Die Idee, Schlußfolgerungen und eventuell Entscheidungen nach Maßgabe der Likelihood der Daten,gegeben eine Hypothese zu fällen, wurde schon von Barnard (1949), aber auch von Birnbaum (1962) und Edwards (1972) diskutiert, vergl. auch Fraser (1991). Glover & Dixon (2004) argumentieren, dass Likelihood-Quotienten einfacheund flexible Statistiken für den empirischen Psychologen liefern. Im Folgenden werden insbesondere die Überlegungen von Royall (1997) dargestellt, der sich ausführlich für die Likelihood als Maß für die Evaluation von Hypothesen ausläßt. Eine neuere Arbeit von Zhang (2009) über das *Law of Likelihood* für zhusammengesetzte Hypothesen wird hier nur hingewiesen; Fitelsons (2008) kritische Überlegungen (s. Abschnitt 5.3) liefern einen Übergang zu Bayesiansichen Verfahren.

Royalls Theorie Die Theorie ist Non-Bayesian. Ausgangspunkt: Zweck einer wissenschaftlichen Studie sei die Produktion von Evidenz. Fishers Ansatz ist unbefriedigend, und Neyman-Pearsons Ansatz ebenfalls, da dieser Ansatz auf Entscheidungen für Handlungen zielt, nicht auf die Repräsentation der Evidenz.

Royalls Behauptung/Definition: der Begriff der Evidenz sei im *Law of Likelihood* verkörpert (embodied). Das Vorgehen ist zunächst analog zu dem von Neyman-Pearson:

1. Es werden zwei alternative Hypothesen H_1 und H_2 betrachtet, für die Wahrscheinlichkeitsdichten f_1 und f_2 für eine zufällige Veränderliche X korrespondieren.

2. Es werden i.i.d. Realisationen $X_1, \ldots, X_n$ von X bestimmt; es resultiert ein Datenvektor $\mathbf{x} = (x_1, \ldots, x_n)$. Die Daten werden nach dem *Law of Likelihood* interpretiert, und zu diesem Zweck wird der Likelihood-Quotient

$$\lambda(\mathbf{x}) = \frac{L_2}{L_1} = \prod_{i=1}^{n} \frac{f_2(x_i)}{f_1(x_i)} \tag{93}$$

   bestimmt

3. Je größer $\lambda(\mathbf{x})$, desto größer die Evidenz für H_2, je kleiner, desto größer ist die Evidenz für H_1.

Royall kontrastiert den Neyman-Pearson-Ansatz und den Likelihod-Ansatz:

> **Neyman-Pearson:** Nach Neyman-Pearson ist ein Experiment ein Verfahren, um zwischen H_1 und H_2 zu *entscheiden*. Die Wahrscheinlichkeit einer korrekten Entscheidung – $1-\alpha$ für H_1, wenn H_1 korrekt ist, $1-\beta$, wenn H_2 korrekt ist – soll maximiert werden.
>
> **Likelihood:** Das Experiment ist ein Verfahren zur Erzeugung von Evidenz über H_1 relativ zu H_2. Das Experiment kann große Evidenz für eine falsche Hypothese generieren, oder schwache Evidenz erzeugen, wenn nämlich die Daten keine der beiden Hypothesen stützen.

(Schwache Evidenz zeigt möglicherweise, dass keine der betrachteten Hypothesen eine gute Erklärung der Daten liefert, d.h. es gibt Anlaß, sich neue Hypothesen auszudenken.) Mit den Wahrscheinlichkeiten $1-M, 1-W$ soll weder ein falsches Resultat mit großer Evidenz noch eines mit schwacher Evidenz eintreten.

Die Begriffe 'große Evidenz' und 'schwache Evidenz' lassen sich spezifizieren:

$$\lambda(\mathbf{x}) = \begin{cases} \geq k, & \text{große Evidenz für } H_2 \\ \leq 1/k, & \text{große Evidenz für } H_1 \\ > 1/k \text{ und } < k, & \text{schache Evidenz} \end{cases} \tag{94}$$

Die Spezifikation von großer und schwacher Evidenz ist natürlich erst vollständig, wenn k definiert wird. Eine Definition von k könne, so Royall, analog zu einem 'signifikanten p-Wert' geschehen. Dem entpräche ein Wert $k = 8$. Dem entspricht ein Likelihood-Quotient für drei aufeinander folgende weiße Kugeln in einem Experiment mit Zurücklegen und einer Urne, in der entweder alle Kugeln weiß sind oder nur die Hälfte. Ein Wert $k = 32$ entspricht ein Signifikanzniveau $\alpha = .01$, entsprechend fünf weißen Kugeln.

Die Definition (94) kann verallgemeinert werden zu $\lambda(\mathbf{x}) \geq k_1$ für große Evidenz für H_2, $\lambda(\mathbf{x}) \leq 1/k_1$ für große Evidenz für H_1, $k_1 \neq k_2$. Dann kann die Wahrscheinlichkeit $P(1/k_1 \leq \lambda(\mathbf{x}) \leq k_2)$ berechnet werden, keine hinreichend große Evidenz für entweder H_1 oder H_2 zu erhalten; damit diese Wahrscheinlichkeit klein wird, kann ein entsprechender Stichprobenumfang n_s bestimmt werden. Ein solcher Wert existiert stets, da sowohl $P(\lambda(\mathbf{x}) \leq 1/k_1)$ wie auch $P(\lambda(\mathbf{x}) \geq k_2)$ mit n_s gegen 1 streben.

Beispiel 5.1 Es werde noch einmal das Beispiel einer $N(\theta, \sigma^2)$-verteilten Variablen betrachtet, mit H_1: $\theta = \mu$, H_2: $\theta = \mu + \delta$. Der Likelihood-Quotient wurde in (75) mit

$$\lambda(\bar{x}) = \exp[(\bar{x} - (\theta + \delta/2))n\delta/\sigma^2]$$

angegeben. Die Daten $\mathbf{x}$ werden einen Likelihood-Quotienten mit

$$1/k_1 < \lambda(\mathbf{x}) < k_2$$

ergeben, wenn der Stichprobenmittelwert in einem Intervall $(\bar{x}_u, \bar{x}_o)$ liegt, wobei

$$\bar{x}_u = \mu + \frac{\delta}{2} - \frac{\sigma^2 \log k_1}{n\delta}, \quad \bar{x}_o = \mu + \frac{\delta}{2} + \frac{\sigma^2 \log k_1}{n\delta}.$$

Es sei W_1 die Wahrscheinlichkeit einer schwachen Evidenz, gegeben H_1 ist wahr, und W_2 die Wahrscheinlichkeit einer schwachen Evidenz, gegeben H_2 ist wahr. Dazu sei

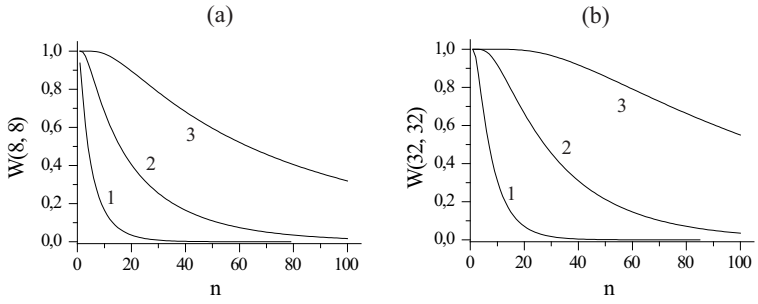
$$W(k_1, k_2) = \Phi\left(\frac{\delta\sqrt{n}}{2\sigma} + \frac{\sigma \log k_1}{\delta\sqrt{n}}\right) - \Phi\left(\frac{\delta\sqrt{n}}{2\sigma} - \frac{\sigma \log k_2}{\delta\sqrt{n}}\right). \tag{95}$$

Dann folgt

$$W_1 = P_1(\bar{x}_u < \bar{X} < \bar{x}_o) = W(k_2, k_1) \tag{96}$$

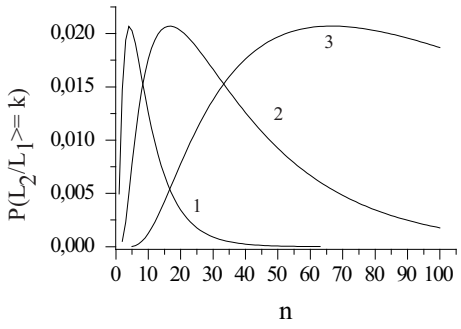
$$W_2 = P_2(\bar{x}_u < \bar{X} < \bar{x}_o) = W(k_1, k_2) \tag{97}$$

Abbildung 7: Wahrscheinlichkeiten für schwache Evidenz. Effektstärken: (1) $\delta/\sigma = 1$, (2) $\delta/\sigma = .5$, (3) $\delta/\sigma = .25$. (a): $k_1 = k_2 = 8$, (b): $k_1 = k_2 = 32$



Es folgt sofort, dass $W_1, W_2 \to 0$ für $n \to \infty$, und dies bedeutet, dass man für hinreichend großen Wert von n_s stets $\max(W_1, W_2) \leq W$ hat. Für $k_1 = k_2$ folgt $W_1 = W_2$. Die Abb. 8 zeigt die Wahrscheinlichkeiten für $\lambda(\mathbf{x}) \in (1/k, k)$ für (a) k = 8 und (b) für $k = 32$ und verschiedene δ/σ-Werte, in Abhängigkeit vom Stichprobenumfang n_s. Angenommen, H_1 sei wahr; trotzdem seien die Daten so

Abbildung 8: Wahrscheinlichkeiten ($\lambda(\mathbf{x}) \geq 8$) für irreführende Evidenz zugunsten H_2 in Abhängigkeit vom Stichprobenumfang $n = n_s$. Effektstärken: (1) $\delta/\sigma = 1$, (2) $\delta/\sigma = .5$, (3) $\delta/\sigma = .25$. (a): $k_1 = k_2 = 8$, (b): $k_1 = k_2 = 32$



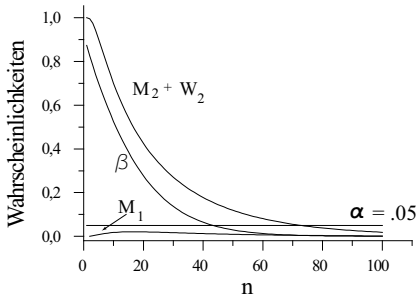
ausgefallen, dass die Evidenz gleichwohl für H_2 spricht. Dies geschieht, wenn $\bar{x} > \bar{x}_0$, und die Wahrscheinlichkeit für dieses Ereignis ist

$$M_1(k_2) = \Phi\left(-\frac{\delta\sqrt{n}}{2\sigma} - \frac{\sigma \log k_2}{\delta\sqrt{n}}\right) \tag{98}$$

Ist umgekehrt H_2 wahr, so tritt der Fall $\bar{x} > \bar{x}_0$ mit der Wahrscheinlichkeit

$$M_2(k_1) = \Phi\left(-\frac{\delta\sqrt{n}}{2\sigma} - \frac{\sigma \log k_1}{\delta\sqrt{n}}\right) \tag{99}$$

Abbildung 9: Unerwünschte Ergebnisse: M_1: H_1 wird verworfen, obwohl wahr. α und β sind die Wahrscheinlichkeiten des Fehlers erster Art bzw. zweiter Art (als Funktion von n)(Neyman & Person), M_2+W_2 : große Evidenz für H_2 wird verfehlt, obwohl H_2 korrekt ist (vergl. Royall (1997), p. 93-94.



auf. Generell läßt sich zeigen, dass

$$M(k) = \Phi\left(-\frac{\delta\sqrt{n}}{2\sigma} - \frac{\sigma \log k}{\delta\sqrt{n}}\right) \leq \Phi(-\sqrt{2\log k}). \tag{100}$$

Dies bedeutet, dass für $k = 8$, $M(k) \leq .021$, und für $k = 32$, $M(k) \leq .004$.

Die Wahrscheinlichkeit einer irreführenden Evidenz M_1 kann analog zum Fehler erster Art der Größe α im NP-Ansatz gesehen werden. Denn die Zurückweisung von H_1 wird oft fälschlicherweise als Evidenz gegen H_1 gesehen. $M_2 + W_2$ ist die Wahrscheinlichkeit, keine starke Evidenz für H_2 zu finden, wenn H_2 wahr ist, und korrepsondiert zum Fehler zweiter Art der Größe β. Im Gegensatz zum NP-Fehler besteht aber $M_2 + W_2$ aus zwei verschiedenen Komponenten, nämlich der Komponente M_2, der Wahrscheinlichkeit, irreführende Evidenz zugunsten von H_1 zu finden, obwohl lH_2 wahr ist, und der Komponente W_2, der Wahrscheinlichkeit, nur schwache Evidenz zu finden. Der Punkt hierbei ist, dass man nach Abschluß einer Untersuchung weiß, ob die Daten nur schwache Evidenz liefern; hat man große EVidenz gewonnen, so weiß man nicht, ob sie irreführend ist.

Unterschied zwischen (α, β) und $(M_1, M_2 + W_2)$. (M_1: irreführende Evidenz zugunsten von H_2, $M_2 + W_2$: keine klare Evidenz für H_2, wenn H_2 wahr ist.) Wie Royall ausführt, werden im Likelihood-Paradigma keine Fehler begangen, da keine Entscheidungen ('no acts of will') begangen werden; das Resultat des Experiments bzw. der Untersuchung ist eine Datenmenge, die zu interpretieren ist:

1. Die Daten $\mathbf{x}$ seien beobachtet worden. Für $\lambda(\mathbf{x}) = L_2/L_1 > k_2$, so wird dies gemäß
$$\lambda(\mathbf{x}) = \exp([(\bar{x} - (\mu + \delta/2)]n\delta/\sigma^2]$$
als Evidenz zugunsten von H_2 interpretiert.

2. Für $1/k_1 < \lambda(\mathbf{x}) < k_2$ hat man keine große Evidenz zugunsten der einen oder der anderen Hypothese.

3. Für $\lambda(\mathbf{x}) \cdots$

□

Das hier vorgestellte Likelihood-Paradigma macht von einem Ansatz Gebrauch, der dem von Neyman & Pearson ähnlich ist. Wie der NP-Ansatz werden Hypothesen nicht – wie bei Fisher – absolut betrachtet, sondern relativ zu einer bestimmten Alternativhypothese. Es wird, anders als bei Neyman & Pearson, kein besonderes Wahrscheinlichkeitskonzept vorausgesetzt: man *kann* ein frequentistisches Modell wählen, muß es aber nicht.

Ein weiterer, wesentlicher Unterschied zwischen dem Likelihood-Paradigma und einem Signifikanz- bzw. Hypothesentest ist, dass beim Likelihood-Paradigma einerseits die Größe der Evidenz über den Likelihood-Quotienten ausgedrückt wird, andererseits die Wahrscheinlichkeit ausgedrückt werden kann, dass zumindest eine spezifizierte Evidenz unter bestimmten Bedingungen ausgedrückt werden kann. Diese Wahrscheinlichkeiten hängen vom Stichprobenumfang ab und die Ausdrücke dafür – M_1, M_2, W_1 und W_2 – können zur Planung der Studie herangezogen werden.

5.2 Royalls Betrachtungen (Paradoxa und ihre Auflösung)

Die Rolle von α und β Eine der ersten Fragen ist, warum man bezüglich H_1 eine Fehlerrate $\alpha = .05$ festlegt, für den Fehler zweiter Art aber eine Fehlerrate aber $\beta = .20$ zuläßt. Es wäre doch auch denkbar, dass man $\alpha = .20$ und $\beta = .05$ wählt.

Tatsächlich beginnt man meistens mit dem Wunsch, $\alpha = \beta = .05$ setzen zu können. Es zeigt sich dann, dass sich ein notwendiger Stichprobenumfang n ergibt, der unter den üblicherweise vorgegebenen Randbedingungen nicht erreichbar ist (etwa wegen zu großer Zeitdauer bzw. zu großen Kosten der Erhebung). Man erhöht dann sukzessive den β-Wert, bis man zu praktikablen Größen kommt.

Dass die Umkehrung $\alpha = .20$, $\beta = .05$ nicht vorkommt, liegt daran, dass der NP-Ansatz gewählt wird. In diesem Ansatz spielt α eine doppelte Rolle, nämlich (i) als Wahrscheinlichkeit einer irrtümlichen Zurückweisung von H_1, wenn H_1 wahr ist, und (ii) als Maß für die Evidenz gegen H_1. Ein kleiner α-Wert soll große Evidenz gegen H_1 reflektieren. β wird nur als Wahrscheinlichkeit, H_2 fälschlich zurückzuweisen, gesehen. β wird in diesem Interpretationsschema zum Ausdruck schwacher Evidenz gegen H_1. Das NP-Schema kann dieser Argumentation zufolge die evidentiellen Aspekte der Daten nicht erfassen, im Gegensatz zum Likelihood-Paradigma, in dem $M_2 + W_2$ die Wahrscheinlichkeit angeben, große Evidenz für H_2 relativ zu H_1 zu finden, wenn H_2 tatsächlich wahr ist.

Der flexible Stichprobenumfang: Cornfields (1966) Argument Ein Wissenschaftler sei daran interessiert, nachzuweisen, dass eine bestimmte Hypothese H_0 (H_1, etwa einer Mittelwertsdifferenz $\Delta\mu$) nicht korrekt ist. Also führt er eine Studie mit einem Stichprobenumfang n durch, testet die Hypothese, – und findet, dass keine Zurückweisung von H_0 möglich ist, das Ergebnis ist nicht signifikant. Also veröffentlicht er sein Ergebnis nicht, denn die Herausgeber der in Frage kommenden Fachzeitschriften sind der Ansicht, dass nur signifikante Ergebnisse von wissenschaftlicher Bedeutung sind, – denn der Herausgeber-Folklore entsprechend repäsentieren nicht signifikante Ergebnisse ja nur "Rauschen".

Wäre aber nun der Stichprobenumfang n größer als der vom Wissenschaftler

gewählte, so wäre die beobachtete Differenz $\Delta\bar{x}$ *könnte* signifikant. Also fragt er sich, wie viel mehr Beobachtungen er benötigt, um ein signifikantes Ergebnis zu erhalten; dabei macht er stillschweigend die Annahme, bei größerem n wieder eine Differenz $\Delta\bar{x}$ zu beobachten, die der Größenordnung nach der bereits gefundenen entspricht. Es ist ja denkbar, dass die Nullhypothese tatsächlich korrekt ist und die mit dem größeren Wert von n verbundene stabilere Schätzung der Differenz von geringerer Größenordnung als die bereits gefundene ist und sich wieder kein signifikantes Resultat bekannt geben läßt.

Im "wirklichen Leben" liegt es nahe, wenn es irgend geht, weiter Daten zu sammeln und, soweit es irgend geht, damit die ursprüngliche Stichprobe zu vergrößern. Die Annahme ist, dass mit wachsendem Stichprobenumfang die Wahrscheinlichkeit steigt, ein signifikantes Ergebnis zu erhalten, zumal die Wahrscheinlichkeit für $\bar{x} = 0$ gleich Null[13] ist auch wenn $\mu = 0$.

Eine seriösere Alternative zum einfacheren post-hoc-Vergrößern der Stichprobe ist, eine weitere Stichprobe zu erheben und diese nicht einfach zur ursprünglichen zu addieren, sondern sie für sich zu analysieren. Es werde angenommen, dass H_0 gilt. Die Wahrscheinlichkeit, dass sich für eine gegebene Stichprobe ein signifikantes Resultat ergibt, beträgt α. Die Wahrscheinlichkeit, dass sich bei zwei Stichproben eine Signifikanz ergibt, ist $p_1 = \alpha$ für den Fall, dass sich schon in der ersten eine Signifikanz ergibt, – dann ist keine zweite mehr nötig, oder in der ersten Stichprobe hat sich keine Signifikanz ergeben *und* sie ergibt sich bei der zweiten; dieses Ereignis hat die Wahrscheinlichkeit $(1 - p_1)\alpha$, so dass die Gesamtwahrscheinlichkeit durch

$$p_{ges} = \alpha + (1 - \alpha)\alpha = \alpha(2 - \alpha) = .0975$$

gegeben ist. Die Gesamtwahrscheinlichkeit ist also größer als .05; allgemein hat man $\alpha(2 - \alpha) > \alpha$ für $\alpha < 1$. Ergibt sich wieder keine Signifikanz, kann man eine dritte Stichprobe hinzuziehen, und die Gesamtwahrscheinlichkeit erhöht sich noch weiter über .05 hinaus. Es gelingt auf diese Weise nicht, die Hypothese insgesamt auf dem α-Niveau zurückzuweisen. Für Cornfield folgt daraus, dass nicht alle Hypothesen, die auf dem $\alpha = .05$-Niveau zurückgewiesen werden, den gleichen Betrag an Evidenz gegen die jeweilige Hypothese reflektieren. Wie Royall (1997), p. 112, ausführt, ergibt sich die hier angedeutete Schwierigkeit aus der Tatsache, dass der aus einem Test resultierende p-Wert zwei Aspekte repräsentieren soll: einerseits soll der die Wahrscheinlichkeit angeben, mit der man fälschlicherweise die korrekte H_0 verwirft, zweitens soll der p-Wert die Evidenz gegen H_0 ausdrücken. Die Auflösung des hier aufscheinenden Paradoxons ergibt sich, so man Royall folgen will, wenn man von der Evidenz gegen eine Hypothese in Relation zur Evidenz gegen eine Alternativhypothese spricht. Dazu betrachtet man den Likelihood-Quotienten

$$\frac{L(\delta)}{L(0)} = \exp\left[\frac{n\delta^2}{\sigma^2}\left(\bar{x}_n - \frac{\delta}{2}\right)\right] > k$$

($\bar{x}_n$ der Mittelwert von n Werten). Man stoppt dann, wenn

$$\bar{x}_n > \frac{\delta}{2} + \frac{\sigma^2 \log 8}{n\delta}$$

Dieser Ausdruck, nicht $\bar{x} > 1.645\sigma/\sqrt{n}$, drückt (relativ) schwache Evidenz gegen H_0 (d.h. $\delta = 0$) aus. Wenn sich nach n Beobachtungen nicht zeigt, dass $\bar{x}_n >$

[13] Wenn $\bar{x}$ eine auf einer stetigen Skala variierende Größe ist. Die Wahrscheinlichkeit, einen bestimmten Wert x auf einer stetigen Skala zu beobachten, ist bekanntlich Null, wegen $\lim \Delta x(F(x + \Delta x) - F(x)) = 0$, wenn F die Verteilungsfunktion von X.

$\delta/2 + \sigma^2(\log 8)/(n\delta)$ ist, können weitere Daten erhoben und nach der Evidenz in der kombinierten Stichprobe gesehen werden. Nach m zusätzlichen Messungen ergibt sich

$$\frac{L(\delta)}{L(0)} = \exp\left[\frac{(m+n)\delta}{\sigma^2}\left(\bar{x}_{m+n} - \frac{\delta}{2}\right)\right].$$

Wenn $\bar{x}_{m+n} > \delta/2 + \sigma^2 \log 8/((m+n)\delta)$ hat man Evidenz gegen H_0. Die Wahrscheinlichkeit, dass diese Evidenz irreführend ist, beträgt

$$1 - \Phi\left(\frac{\delta\sqrt{n}}{2\sigma} + \frac{\sigma \log 8}{\delta\sqrt{n}}\right) \leq \Phi(-\sqrt{2\log 8}),$$

für jede Kombination von b, δ und σ.

Angenommen, die Hypothese H_0 ist korrekt. Die Wahrscheinlichkeit, dass sie irrtümlich zurückgewiesen worden wäre, beträgt $\alpha = .05$. Sie ist aber nicht zurückgewiesen worden. Cornfield (1966) argumentierte, dass es keine zusätzlichen Beobachtungen gäbe, um in diesem Fall die Hypothese zurückzuweisen. Er unterscheidet zunächst zwischen *Sequentieller Analyse* und *sequentiellen Versuchen* (sequential trials). Bei sequentiellen Versuchen hängt die Entscheidung, ob weitere Versuche durchgeführt werden sollen, davon ab, welche Information man bis zu einem bestimmten Zeitpunkt gesammelt hat. Bei der sequentiellen Analyse hängt das Ergebnis nicht nur von den Daten, sondern auch von der Stop-Regel ab.

5.3 Fitelsons Überlegungen

Das Likelihood-Paradigma scheint die Problematik von Fishers Signifikanztests und Neyman & Pearsons *inductive behaviour* zu umgehen, ohne wie der Bayessche Ansatz auf a-priori-Verteilungen zurückgreifen zu müssen. Gleichwohl, es regt sich Kritik.

Fitelson (2008) beginnt mit einer Darstellung von "evidential concepts" mit probabilistischen Explikationen, insbesondere mit Carnaps (1962) Taxonomie in klassifikatorische oder qualitative, komparative oder relationale Konzepte, und quantitative Konfirmation. Carnap:

> Das komparative Konzept der Konfirmation wird üblicherweise in Aussagen der Form "H_1 wird durch E_1 stärker bestätigt als H_2 durch E_2".

Dieser Begriff kommt dem des "relational support" nahe. Dies ist der Begriff, auf den im gegenwärtigen Zusammenhang fokussiert wird. Allerdings gibt es zwei wichtige Unterschiede: (i) in Carnaps Ansatz gibt es eine 'evidentielle Evidenz' E und zwei Hyothesen H_1 und H_2, (ii) macht Carnap die stillschweigende Annahme, dass die relationale Aussage "die Evidenz E stützt H_1 relativ zu H_2" kann auf einen Vergleich nicht-relationaler, quantitativer, bestätigender Größen zurückgeführt werden, d.h. einmal auf die Bestätigung von H_1 durch E und einmal auf eine Bestätigung von H_2 durch E; die Bestätigung von H_1 ist dann größer als die von H_2. Einige Likelihoodisten (Sober (1994), Royall (1997)) akzeptieren diesen Ansatz nicht und ziehen den nicht-reduktionistischen Ansatz über die Likelihood vor, wohingegen moderne Bayesianer eine Art Carnapschen Ansatz vorziehen. Relationale Bestätigung wird als 'abgeleitetes Konzept' eingeführt, die zugrundeliegenden primitiven Ausdrücke sind nicht-relational. Hier liegt der Hauptgrund für den Unterschied zwischen Likelihoodisten und Bayesianern. Nach Carnap: 'E favorisiert H_1 gegenüber H_2' wird übersetzt in 'E stützt H_1 mehr als H_2'.

Likelihoodisten sind *kontrastive Empiriker*: Evidenz wirkt stützend in einer relationalen, aber nicht auf die von Carnap intendierte Weise. Der Test von Theorien wird als eine "kontrastierende Aktivität" betrachtet (Sober (1994)). Der Test einer Theorie setzt, diesem Ansatz zufolge, eine Spezifikation dessen voraus, gegen das bzw. gegen welche Theorie denn getestet werden soll. Um zu sehen, ob eine Theorie T plausibel ist, muß man herausfinden, ob sie plausibler ist als Nicht-T. Andererseits: wenn T gegen T' getestet werden soll, kann es sein, dass eine Evidenz E hinreichend ist, wenn aber gegen T'' getestet werden soll, so wird eine andere Evidenz E' benötigt.

Die Likelihoodisten schlagen einen einfachen Ansatz vor:

> **LL:** (Law of Likelihood) Die Evidenz E favorisiert H_1 gegenüber H_2 dann und nur dann, wenn H_1 die Daten bzw. die Evidenz E als wahrscheinlicher erscheinen läßt als H_2. (Royall, 1997)

Man kann dies direkt über die Likelihoods ausdrücken: E favorisiert

> Man habe ein Ass aus einem Kartendeck gezogen; dies sei die verfügbare Evidenz E. Die Hypothesen seien: H_1: Die Karte ist ein Herz-Ass versus H_2: Die Karte ist entweder ein Pik-Ass oder ein Kreuz-Ass. Der Aufbau eines Kartenspiels impliziert nun, dass $P(H_1|E) = 1/4$, denn es gibt nur vier Asse, und aus dem gleichen Grund $P(H_2|E) = 1/2$. Andererseits gilt aber $P(E|H_1) = P(E|H_2) = 1$. Die Likelihoods für E sind identisch, die a-posteriori-Wahrscheinlichkeiten sind es nicht. Nach **LL** würde die Evidenz für beide Hypothesen gleichermaßen sprechen; man wird aber geneigt sein, die Evidenz für die Hypothesen über die a-posteriori-Wahrscheinlichkeiten zu bewerten. (Leeds, nach Sober 2005)

Zumindest intuitiv betrachtet hebelt das Beispiel das Likelihood-"Gesetz" aus. Darüber hinaus gilt

$$P(E|\neg H_1) > P(E|\neg H_2).$$

Was würde ein Likelihoodist zum Leedschen Beispiel sagen? Er würde darauf hinweisen, dass der Unterschied zwischen den a-posteriori-Wahrscheinlichkeiten $P(H_1|E)$ und $P(H_2|E)$ nur auf die unterschiedlichen a-priori-Wahrscheinlichkeiten zurückginge; dieser Unterschied erkläre auch $P(E|\neg H_1) > P(E|\neg H_2)$. Die a-priori-Größen seien aber nicht-relatinale Eigenschaften der Hypothesen H_i. Gleichwohl, die Möglichkeit besteht, das die Aussage 'E favorisiert H_1 gegenüber H_2' als Aussage über a-posteriori-Wahrscheinlichkeiten verstanden wird.

Zweites Gegenbeispiel zu **LL**:

> Eine einzige Karte wird aus einem gut gemischten Kartendeck gezogen. E sei eine Pik-Karte. H_1: Die Karte ist ein Pik-Ass, H_2: Die Karte ist schwarz. Sicherlich ist $P(E|H_1) = 1$, denn wenn die Karte ein Pik-Ass ist, muß sie auch ein schwarzes Zeichen haben. Dagegen ist $P(E|H_2) = 1/2$, denn unter der Bedingung, dass die Karte schwarz ist, gibt es nur zwei Arten von Karten: Pik-Karten und Kreuz-Karten, die jeweils mit gleicher Häufigkeit vorkommen. (Fitelson (2007), p. 477)

Es ist $P(E|H_1) > P(E|H_2)$, also müßte E, dem "Law of Likelihood" **LL** entsprechend, H_1 favorisieren. Andererseits garantiert E, dass H_2, aber nicht H_1 zutrifft. Fitelson (2007) schlägt dementsprechend ein anderes Kriterium vor:

(*) Wenn E schlüssige Evidenz für H_1 liefert, aber nicht-schlüssige Evidenz für H_2, dann favorisiert E die Hypothese H_1 gegenüber H_2.

Sober (2005) schlägt ein reduziertes LL vor:

> **Reduziertes LL:** The Law of Likelihood should be restricted to cases in which the probabilities of hypotheses are not under consideration (perhaps because they are not known or are not even "well-defined") and one is limited to information about the probability of the observations given different hypotheses.

LL soll also nur gelten, wenn die a-priori-Verteilungen nicht bekannt sind, und nur die Likelihoods zur Verfügung stehen. Damit ist das **LL** natürlich kein "Gesetz" im üblichen Sinne mehr, – es ist eigentlich falsch. Fitelson argumentiert, dass es nicht die a-priori-Verteilungen sind, sondern die *catch-alls* $P(E|\neg H_1)$ und $P(E|\neg H_2)$, die unbekannt sein müssen, da es diese catch-alls sind und nicht die a-priori-Verteilungen, die die jeweilig favorisierenden Aussagen bedingen.

Bayessche Ansätze

> **BA:** Die Evidenz E favorisiert die Hypothese H_1 relativ zur Hypothese H_2 genau dann, wenn E die Hypothese H_1 mehr bestätigt als die Hypothese H_2.

Die Idee hinter dieser Charakterisierung ist, dass E die jeweiligen Hypothesen nicht-relational, individuell zu einem unterschiedlichen Maß bestätigt, und Relation der Favorisierung einer Hypothese nach Maßgabe dieser individuellen Bewertung vorgenommen wird. Bestätigung ist dabei für Bayesianer eine Sache der *probabilistischen Relevanz*. Dazu muß ein Relevanzmaß $\mathbf{c}(H, E)$ eingeführt werden, das den Grad,in dem E die Wahrscheinlichkeit von H anhebt, angibt. Mögliche Maße für *non-relationale Konfirmation*:

- *Differenz:*
$$d(H, E) = P(H|E) - P(H)$$
- *Verhältnis:*
$$r(H, E) = \frac{P(H|E)}{P(H)}$$
- *Likelihood-Quotient:*
$$\lambda(H, E) = \frac{P(E|H)}{P(E|\neg H)}.$$

Setzt man eines der Maße in **BA** ein, so erhält man eine bestimmte Bewertung:

> **BA(x)**: Die Evidenz E favorisiert die Hypothese H_1 relativ zur Hypothese H_2 gemäß dem Maß $\mathbf{c}$ genau dann, wenn $\mathbf{c}(H_1, E) > \mathbf{c}(H_2, E)$.

Wird das Verhältnismaß eingesetzt, erhält man ein Kriterium, das "likelihoodistisch" ist. Die beiden anderen führen zu einem eher Bayes'schen Kriterium. Das Verhältnismaß erweist sich als logisch äquivalent dem **LL**-Kriterium (Fitelson 2007, p. 478). Joyce (2004) führte ein Kriterium ein:

Weak Law of Likelihood, WLL: Die Evidenz E favorisiert die Hypothese H_1 relativ zur Hypothese H_2, wenn sowohl $P(E|H_1) > P(E|H_2)$ als auch $P(E|\neg H_1) \leq P(E|\neg H_2)$ gilt.

Dass WLL heißt 'schwaches Gesetz der Likelihood', weil es logisch schwächer als LL ist, – es kann nicht konsistent als falsch bezeichnet werden. Fitelson argumentiert, dass (i) dass WLL nur eine hinreichende, nicht aber eine notwenidge Bedingung für die Favorisierung einer Hypothese gegenüber einer anderen repräsentiere, und (ii), dass das oben eingeführte Maß $r(H, E)$ das WLL impliziere, wobei er aber keinen Beweis gibt. Die folgende Betrachtung zeigt, dass der hier von Fitelson benützte Implikationsbegriff (er spricht von *entailment*) allerdings nicht ganz klar ist:

Es gelte $r(H, E)$, insbesondere $r(H_1, E) > r(H_2, E)$, d.h. die Evidenz favorisiere H_1 gegenüber H_2. Dann soll also

$$\frac{P(H_1|E)}{P(H_1)} > \frac{P(H_2|E)}{P(H_2)}$$

gelten. Es ist $P(H|E) = P(E|H)P(H)/P(E)$, so dass

$$\frac{P(E|H_1)}{P(E)} > \frac{P(E|H_2)}{P(E)}$$

folgt, d.h. $P(E|H_1) > P(E|H_2)$. Gilt andererseits $P(E|H_1) > P(E|H_2)$, so folgt umgekehrt wegen $P(E|H) = P(H|E)P(E)/P(H)$ sofort

$$\frac{P(H_1|E)P(E)}{P(H_1)} > \frac{P(H_2|E)P(E)}{P(H_2)}, \text{ d.h. } r(H_1, E) > r(H_2, E).$$

Dem WLL zufolge impliziert

$$P(E|H_1) > P(E|H_2) \ \textit{und} \ P(E|\neg H_1) \leq P(E|\neg H_2)$$

die Aussage, dass die Evidenz E die Hypothese H_1 gegenüber der Hypothese H_2 favorisiere. Wie gerade gezeigt, impliziert bereits der erste Teil – $P(E|H_1) > P(E|H_2)$ – die Favorisierung von H_1 gegenüber H_2, die *catch-all*-Bedingung $P(E|\neg H_1) \leq P(E|\neg H_2)$ wird nicht benötigt. Dieser Sachverhalt illustriert die Bemerkung Fitelsons, dass das WLL nur eine hinreichend, nicht aber auch eine notwendige Bedingung für die Favorisierung von H_1 gegenüber H_2 darstellt. □

Carnap (1962) unterschied zwischen zwei Konzepten nicht-relationaler Konfirmation: i) *confirmation as firmness*, und (ii) *confirmation as increase of firmness*. Die Letztere entspricht der Bayesschen *confirmation als probability-rising*, während confirmation as firmness als ein Schwellenkonzept verstanden werden kann: E bestätigt H entspricht $P(H|E) > \rho$, ρ ein bestimmter Schwellenwert. Nach diesem Kriterium wird die a-posteriori-Wahrscheinlichkeit ein direktes Maß für non-relationale Bestätigung, und die Definition der Hypothesenfavorisierung wäre durch

$$E \text{ favorisiert } H_1 \text{ dann und nur dann, wenn } P(H_1) > P(H_2|E). \tag{101}$$

Allerdings gilt diese Charakterisierung des Bayesschen Favorisierens mittlerweie als inadäquat, weil der zugrundeliegende Konfirmationsbegriff die *probabiitsitische Relevanz* nicht berücksichtigt. Dazu betrachte man den Fall, dass die Evidenz E die Wahrheit von H_1 und die Falschheit von H_2 nahelegt. Dementsprechend könnte man entsprechend (**x**) sagen, dass E die Hypothese H_1 gbegenüber H_2 favorisiert. Andererseits ist das LL äquivalent zum Verhältniskriterium

$r(H, E)$, so dass der Likelihood-Ansatz dieselbe Konsequenz hat, – aber nicht (101), denn nach diesem Kriterium kann E auch H_2 gegenbüber H_1 favorisieren, da ja einerseits $P(H_1|E)/P(H_1) > P(H_2|E)/P(H_2)$ sein kann, aber gleichwohl $P(H_2|E) > P(H_1|E)$ möglich ist. Man darf den Grad der Bestätigung nicht mit dem Grad der Wahrscheinlichkeit (degree of belief) verwechseln; für eine Bayessche Theorie ist der Begriff der Bestätigung zentral.

Royalls Kritik der Bayesschen Favorisierung: (Royall (1997), Kap.1) betrachtet den Fall eines Arztes, der eine Krankheit diagnostizieren soll. Der Arzt verwendet einen Test, dessen Charakteristika in der folgende Tabelle angegeben werden: Für einen gegebenen Patienten – Mr. Doe – falle der Test positiv (+) aus.

Tabelle 11: Test für eine Krankheit; K vorhanden, $\bar{K}$ Krankheit nicht vorhanden

	Testergebnis (E)	
	+	-
K	.95	.05
$\bar{K}$	.02	.98

Nach Royall hat der Arzt die folgenden Optionen:

1. Mr. Doe hat die Krankheit K wahrscheinlich nicht.
2. Mr. Doe sollte gegen die Krankheit K behandelt werden.
3. Das Testresultat repräsentiert Evidenz E, dass Mr. Doe an K leidet.

Zu 1: Die Feststellung, Mr. Doe habe die Krankheit K wahrscheinlich nicht, beruht auf einer Annahme bezüglich der a-priori-Wahrscheinlichkeit $P(K)$ für K, und natürlich auf den Testcharakteristika, wie sie in der Tabelle 11 angegeben wurden. Nach dem Bayesschen Theorem hat man

$$P(K|+) = \frac{P(+|K)P(K)}{P(+)} = \frac{P(+|K)P(K)}{P(+|K)P(K) + P(+|\bar{K})P(\bar{K})},$$

wobei $\bar{K}$ für "Hat die Krankheit K nicht" steht. Setzt man die Werte aus der Tabelle ein, erhält man

$$P(K|+) = \frac{.95P(K)}{.95P(K) + .02(1 - P(K))}$$

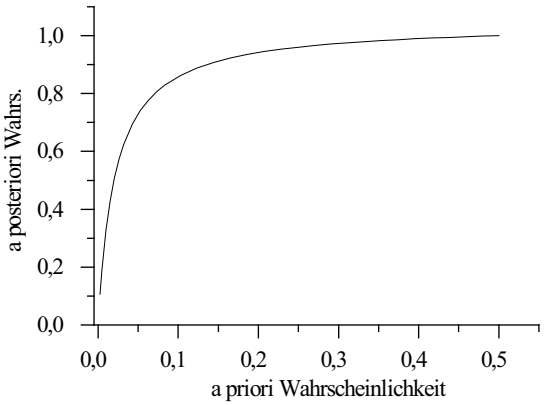
Für $P(K) = .001$ erhält man eine A-posteriori-Wahrscheinlichkeit $P(K|+) = .046$, für $P(K) = .2$ folgt $P(K|+) = .942$. Wie man der Abbildung 10 entnimmt, strebt die A-posteriori-Wahrscheinlichkeit sehr schnell gegen 1, gleichwohl ist Abhängigkeit der A-posteriori- von der A-priori-Wahrscheinlichkeit deutlich.

Zu 2: Selbst wenn $P(K|+)$ klein ist – etwa $P(K|+) = .045$ – kann die Schlußfolgerung 2. korrekt sein, einfach weil die Konsequenz einer Nichtbehandlung desaströs sein kann. Um die Option 2. diskutieren zu können, muß eine Reihe von Nebenbedingungen betrachtet werden.

Zu 3: Hier schlägt Royall vor, nicht nach Bayes vorzugehen, sondern nach dem, was er das *Law of Changing Probability* nennt, das er so einführt:

(i) Eine Beobachtung $X = x$ ist Evidenz für eine Hyxpothese H, wenn x die Wahrscheinlichkeit für H erhöht, d.h. wenn $P(H|X = x) > P(H)$ ist, und

Abbildung 10: A-posteri-Wahrscheinlichkeit als Funktion der a-priori-Wahrscheinlichkeit



(ii) Das Verhältnis $P(H|X = x)/P(H)$ ist ein Maß für die Stärke der Evidenz.

Royalls 'Law of Changing Probability' $P(H|X = x)/P(H)$ ist aber gerade die Regel $r(H, E)$, mit $E = x$, – die er aber über den Likelihood-Quotienten

$$\lambda = \frac{P(+|K)}{P(+|\bar{K})}$$

ausdrückt (mit $K = H$, und $+$ für E):

$$\begin{aligned} r(K,+) &= \frac{P(K|+)}{P(K)} = \frac{P(+|K)}{P(+)} \\ &= \frac{\lambda P(+|\bar{K})}{P(+|K)P(K) + P(+|\bar{K})P(\bar{K})} = \frac{\lambda}{\lambda P(K) + 1 - P(K)} \end{aligned}$$

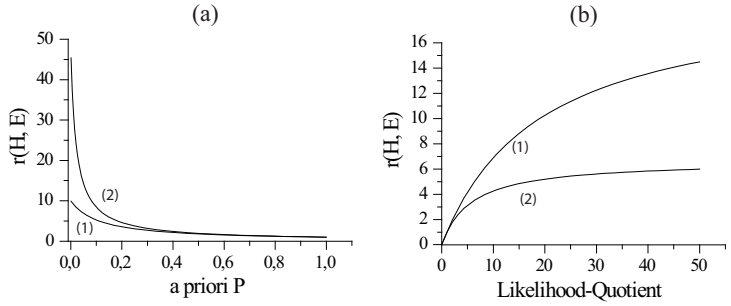
Royall betrachtet nun den Ausdruck $\lambda/(\lambda P(K) + 1 - P(K))$ als Funktion von des Likelihood-Quotienten λ; $r(K, +)$ ist eine monoton wachsende Funktion von λ und kann als Evidenz für das Vorhandensein der Krankheit K interpretiert werden. Hier ist $\lambda = .95/.02 = 47.5$, während $r(K, +)$ noch von $P(K)$ abhängt; diese Größe stützt die Hypothese K nur dann, wenn $\lambda > 1$. Abbildung 11 zeigt das Verhalten von $R(H, E)$ für verschiedene Werte von λ und $P(K)$. Royall ist der Ansicht, das 'Law of Changing Probability' sei von nur geringem Nutzen, da es eben von der a priori-Wahrscheinlichkeit $P(K)$ bzw allgemein von $P(H)$ abhänge. Die Evidenz E solle unabhängig von jeder subjektiven Größe bewertet werden.

Fitelson (2007) argumentiert, dass Royall allerdings mehr als nur das LL (*Law of Likelihood*) annehme, nämlich eine Art Verallgemeinerung des LL, das die Form

> **LL$^+$:** E favorisiert H_1 relativ zu H_2 genau dann, wenn $P(E|H_1) > P(E|H_2)$ *und* das Ausmaß, in dem E die Hypothese H_1 relativ zur Hypothese H_2 favorisiere, sei durch den Likelihood-Quotienten $\lambda = P(E|H_1)/P(E|H_2)$ gegeben.

Nach Fitelson wird nun aber dem Bayesianer ermöglicht, seinerseits eine Verallgemeinerung vorzunehmen:

Abbildung 11: Das *Law of Changing Probability* (äquivalent $r(H, E)$) (a) als Funktion der a priori Wahrscheinlichkeit, (1) $\lambda = 10$, (2) $\lambda = 47.2$; (b) als Funktion von λ, (1) $P(K) = .05$, (2) $P(K) = .15$.



> **$\mathbf{r}^+$:** E favorisiert H_1 relativ zu H_2 genau dann, wenn $r(H_1, E) > r(H_2, E)$ *und* das Ausmaß, in dem E die Hypothese H_1 relativ zurHypothese H_2 favorisiert, ist durch das Verhältnis $r(H_1, E)/r(H_2, E)$ gegeben.

Nach Fitelson ist das Kriterium **$\mathbf{r}^+$:** dem Kriterium **$\mathbf{LL}^+$:** logisch äquivalent (ein expliziter Beweis wird nichtg geliefert); womit dan gezeigt wäre, dass Royall sich gar nicht von den Bayesianern abgesetzt hat!

6 Bayes-basierte Dateninterpretationen

Nachdem die Arbeit von Bayes 1763 veröffentlicht worden war, erschien die Anwendung des Bayesschen Satzes zur Evaluation von Hypothesen der natürliche Ansatz zu sein. Pierre Simon Laplace elaborierte und popularisierte die Bayessche Idee, insbesondere für den Parameter einer Binomialverteilung in seiner Arbeit 1774, aber auch in späteren Arbeiten. Einen Überblick findet man in Joyce (2003), historische Betrachtungen insbesondere zum 'objektiven' Bayesianismus findet man in Fienberg (2006). Die folgenden Betrachtungen sind weniger eine Einführung in die Bayessche Statistik, sondern fokussieren mehr auf grundsätzliche Fragen.

6.1 Bayessche Evidenz: Parameter, Hypothesen, Entscheidungen

6.1.1 Evidenz: Schätzung von Parametern

Es ist nützlich, zunächst zu sehen, in welchem Sinne eine Bayessche Datenanalyse Evidenz über Hypothesen liefert. Es wird zunächst das Prinzip der Schätzung von Parametern und des Tests von Hypothesen betrachtet; dabei gehen a-priori-Verteilungen ein, die als subjektive Komponente des Bayesschen Ansatzes gelten. In welchem Sinne diese Kritik gilt, wird in den folgenden Abschnnitten diskutiert.

Hypothesen sind im Folgenden Hypothesen über Parameter. Es wird zunächst der einparametrige Fall betrachtet. θ kann der Parameter einer Binomial- oder

einer Poisson-Verteilung sein, oder es kann ein Parameter irgendeiner anderen Verteilung sein. Der Fokus liegt in jedem Fall auf der a-posteriori-Verteilung:

$$f(\theta|x) = \frac{f(x|\theta)g(\theta)}{\int_\Theta f(x|\theta)g(\theta)d\theta}, \quad \int_\Theta f(x|\theta)g(\theta)d\theta = f(x). \tag{102}$$

g die a-priori-Verteilung. $f(x|\theta)$ ist die Likelihood-Funktion, und $f(\theta|x)$ die a-posteriori-Verteilung.

Unter der Annahme, dass der Erwartungswert von θ existiert, ist er durch

$$\mathbb{E}(\theta|x) = \int_\Theta \theta f(\theta|x)d\theta \tag{103}$$

gegeben, gegeben die Daten x. In analoger Weise können andere Größen der a-posteriori-Verteilung wie der Modus Mod und der Median Med definiert werden:

$$\text{Mod}(\theta|x) = \max_\theta f(\theta|x), \quad \text{Med} = a = \int_{-\infty}^{a} f(\theta|x)d\theta = \int_a^\infty f(\theta|x)d\theta = .5. \tag{104}$$

Statt eines Konfidenzintervalles für θ kann nun ein Kredibilitätsintervall definiert werden:

Definition 6.1 *Es sei $\alpha \in (0,1)$; dann ist das $(1-\alpha)$-Kredibilitätsintervall durch dass Intervall (t_1, t_2) definiert, für das*

$$\int_{\theta_1}^{\theta_2} f(\theta|x)d\theta = 1 - \alpha \tag{105}$$

gilt.

Üblicherweise wird θ_1 durch das $\alpha/2$-Quantil und θ_2 durch das $(1-\alpha/2)$-Quantil definiert.

Definition 6.2 *Das Intervall (θ_1, θ_2) heißt HPD-Intervall (*highest posterior density interval*, wenn für alle $\theta \in (\theta_1, \theta_2)$ die Relation*

$$f(\theta|x) \geq f(\tilde{\theta}|), \quad \textit{für alle } \tilde{\theta} \notin (\theta_1, \theta_2) \tag{106}$$

gilt.

Das Kredibilitätsintervall hat, im Unterschied zum Konfidenzintervall, eine sehr direkte evidentielle Interpretation: es ist das Intervall, in dem mit der Wahrscheinlichkeit $1-\alpha$ der gesuchte Parameterwert liegt. Für das Konfidenzintervall gilt diese Interpretation ja bekanntlich nicht.

Man kann fragen, für welchen Parameterwert θ_0 die a-posteriori-Wahrscheinlichkeit $P(\theta|x)$ maximal wird. Man hat

$$\left.\frac{dP(\theta|x)}{d\theta}\right|_{\theta=\theta_0} = \left.\frac{dP(x|\theta)}{d\theta}\right|_{\theta=\theta_0} P(\theta) + P(x|\theta)\left.\frac{dP(\theta)}{d\theta}\right|_{\theta=\theta_0} = 0. \tag{107}$$

Die a-priori-Verteilung $P(\theta)$ sei die Gleichverteilung. Dann ist $dP(\theta)/d\theta = 0$ für alle $\theta \in \Theta$ und

$$\left.\frac{dP(x|\theta)}{d\theta}\right|_{\theta=\theta_0} P(\theta) = 0$$

dann, wenn

$$\left.\frac{dP(\theta|x)}{d\theta}\right|_{\theta=\theta_0} = 0.$$

In diesem Fall nimmt die a-posteriori-Verteilung ein Maximum an, wenn die Likelihood ein Maximum annimmt, d.h. im Falle einer gleichverteilten a-priori-Verteilung ist die Maximum-Likelihood-Schätzung gerade diejenige, die auch die a-posteriori-Verteilung maximiert. Natürlich ist das Vorliegen einer Gleichverteilung nur eine hinreichende, aber keine notwendige Bedingung für diesen Sachverhalt.

Beispiel 6.1 Die zufällige Veränderliche X sei binomialverteilt mit den Parametern θ und n, $\theta \in (0,1)$ unbekannt. Als a-priori-Verteilung soll die Beta-Verteilung $B(\alpha,\beta)$ gewählt werden, $\alpha,\beta > 0$. Dann hat man

$$f(x|\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x} \tag{108}$$

und die a-priori-Verteilung ist

$$g(\theta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)} \quad B(\alpha,\beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}dt. \tag{109}$$

Offenbar ist $B(1,1) = 1$ und $g(\theta) = 1$ ist die Gleichverteilung auf $(0,1)$. Für die a-posteriori-Verteilung erhält man allgemein

$$f(\theta|x) \propto f(x|\theta)g(\theta) = \theta^{\alpha+x-1}(1-\theta)^{\beta+n-x-1}. \tag{110}$$

Der Proportionalitätsfaktor ist wie in (102) definiert. Damit ist die zufällige Veränderliche $\theta|x$ Beta-verteilt mit den Parametern $\alpha+x$, $\beta+n-x$. Für den Erwartungswert von $\theta|x$ erhält man

$$\mathbb{E}(\theta|x) = \frac{\alpha+x}{\alpha+\beta+n}. \tag{111}$$

Man kann diesen Ausdruck umformen in

$$\mathbb{E}(\theta|x) = \frac{\alpha+\beta}{\alpha+\beta+n}\frac{\alpha}{\alpha+\beta} + \frac{n}{\alpha+\beta+n}\cdot\frac{x}{n}. \tag{112}$$

Damit wird deutlich, dass der Erwartungswert $\mathbb{E}(\theta|x)$ ein gewogenes Mittel des Erwartungswertes der a-priori-Verteilung und der Maximum-Likelihood-Schätzung x/n von θ ist. Man sieht leicht, dass $\mathbb{E}(\theta|x) \to x/n$ für $n \to \infty$, d.h. für wachsenden Wert von n strebt der Erwartungswert gegen die Maximum-Likelihood-Schätzung von θ.

Ist die a-priori-Verteilung eine Gleichverteilung ($\alpha = \beta = 1$), so findet man

$$\mathbb{E}(\theta|x) = \frac{x+1}{n+2} \tag{113}$$

und der a-posteriori-Modus ist durch

$$\mathrm{Mod}(\theta|x) = \frac{x}{n}, \tag{114}$$

gegeben, ist also gleich dem Maximum-Likelihood-Schätzer für θ. □

6.1.2 Hypothesentests

Die Frage ist weiter, wie Hypothesen über Parameter getestet werden können. Jeffreys (1961) beschreibt das allgemeine Vorgehen, eine gute Einführung findet sich in Edwards et al. (1963); in einer großen Anzahl von Artikeln werden spezielle Probleme bzw. Fragestellungen diskutiert.

Zunächst werde der Fall betrachtet, dass eine einfache, "scharfe" Nullhypothese H_0: $\theta = \theta_0$ gegen eine diffuse Alternativhypothese H_1: $\theta \neq \theta_0$ getestet werden soll.

Es sei, der Kürze wegen, $\pi_0 = P(H_0)$. Dann wird der Quotient

$$\frac{P(H_0|x)}{P(H_1|x)} = \frac{P(H_0|x)}{1 - P(H_0|x)} = \frac{P(x|\theta_0)}{m_g(x)} \cdot \frac{\pi_0}{1 - \pi_0} \tag{115}$$

betrachtet. Die Größen

$$\frac{P(H_0|x)}{1 - P(H_0|x)}, \quad \frac{\pi_0}{1 - \pi_0}$$

sind die a-posteriori-Chance (posterior odds) bzw. die a-priori-Chance (prior odds) für H_0. $m_g(x)$ ist wie üblich ergibt durch

$$m_g(x) = \int_\Theta P(x|\theta) g(\theta)\, d\theta \tag{116}$$

gegeben. Die Größe

$$B_g(x) = \frac{P(x|H_0)}{m_g(x)} \tag{117}$$

ist hier der Bayes-Faktor für H_0 versus H_1. $B_g(x)$ kann als der Likelihood-Quotient für die Daten unter H_0 versus H_1 betrachtet werden. Die unteren Grenzen für $P(H_0|x)$ hängen mit den unteren Grenzen für B_g zusammen.

Es werde zunächst eine einseitige Hypothese betrachtet:

$$H_0\text{: } \theta \leq \theta_0. \tag{118}$$

Dann ist

$$P(H_0|x) = P(\theta \leq \theta_0|y) = \int_{-\infty}^{\theta_0} P(\theta|x) d\theta, \tag{119}$$

und für H_1: $\theta > \theta_0$ erhält man sofort

$$P(H_1|x) = 1 - P(H_0|x). \tag{120}$$

Hypothesen der Form H_0: $\theta \in I$, I ein Intervall aus Θ liefern dann nach (119) die a-posteriori-Wahrscheinlichkeit

$$P(H_0|x) = \int_I P(\theta|x) d\theta. \tag{121}$$

Aus (121) ergibt sich auch sofort der Fall einer "scharfen" Hypothese

$$H_0\text{: } \theta = \theta_0, \quad \theta \subseteq \mathbb{R} \tag{122}$$

nämlich

$$P(H_0|x) = \int_{\theta_0}^{\theta_0} P(\theta|x) d\theta = 0, \tag{123}$$

entsprechend der Tatsache, dass die Wahrscheinlichkeit, dass eine stetige zufällige Veränderliche X einen bestimmten Wert x annimmt, gleich Null ist. Der Fall eines stetigen Parameterraums ist aber der allgemeine Fall (man denke an den Erwartungswert einer stetigen zufälligen Veränderlichen, oder den Parameter $\theta = p$ einer binomialverteilten Variablen). Eine wirklich "scharfe" H_0 ist deswegen im Allgemeinen nicht plausibel, – es sei denn, man wird aus grundsätzlichen Betrachtungen auf eine solche Hypothese geführt (etwa: es gibt Telepathie oder es gibt sie nicht). Plausibler ist eine Formulierung

$$H_0\text{: } |\theta - \theta_0| \leq b \tag{124}$$

für hinreichend kleinen Wert von b (Berger & Sellke (1987)). Äquivalent dazu kann eine a-priori-Verteilung mit entsprechend kleiner Varianz gewählt werden, etwa eine Normalverteilung $N(\theta_0, \sigma^2)$. Es sei angemerkt, dass $\lim_{\sigma\to 0} N(\theta, \sigma^2) = \delta(\theta - \theta_0)$, δ die Dirac-Delta-Funktion, die der scharfen Nullhypothese H_0: $\theta = \theta_0$ entspricht.

Beispiel 6.2 (Berger & Sellke, 1987) Die a-priori-Wahrscheinlichkeit sei π_0 und es sei $g = N(\theta_0, \sigma^2)$. Der Mittelwert $\bar{x}$ ist eine suffiziente Statistik für θ und bekanntlich gilt $\tilde{\sim} N(\theta, \sigma^2/n)$. Dann folgt

$$m_g(\bar{x}) \sim N(\theta_0, \sigma^2(1 + 1/n).$$

Daraus folgt

$$B_g(x) = \frac{P(x|\theta_0)}{m_g(\bar{x})};$$

explizit findet man

$$B_g(x) = \frac{(2\pi\sigma^2/n)^{-1/2} \exp\left(-\frac{n(\bar{x}-\theta_0)^2}{2\sigma^2}\right)}{(2\pi\sigma^2(1+1/n)^{-1/2} \exp\left(-\frac{(\bar{x}-\theta_0)^2}{2\sigma^2(1+1/n)}\right)},$$

woraus sich durch etwas Rechnung

$$B_g(x) = \sqrt{1+n} \exp\left(-\frac{t^2}{2(1+1/n)}\right), \quad t = \frac{\bar{x} - \theta_0}{\sigma} \tag{125}$$

ergibt. Damit erhält man

$$P(H_0|x) = \left[1 + \frac{1-\pi_0}{\pi_0\sqrt{1+n}} \exp\left(\frac{t^2}{2(1+1/n)}\right)\right]^{-1}. \tag{126}$$

(Spezial für $\pi_0 = 1/2$ ergibt sich (41) auf Seite 30. Man sieht, dass für $n \to \infty$ wieder das Lindley-Paradoxon resultiert, denn dann $P(H_0|) \to 1$. Die Annahme $\pi_0 = 1/2$ entspricht einer "objektiven" a-priori-Annahme. □

Natürlich können mehr als ein Parameter zur Diskussion stehen. Ein Beispiel ist eine zufällige Veränderliche X, von der angenommen wird, sie sei normalverteilt mit dem Erwrtungswert $\mathbb{E}(X) = \mu$ und der Varianz $\mathbb{V}(X) = \sigma^2$. $\mathbf{x} = (x_1, \ldots, x_n)$ sei eine Stichprobe für diese Variable. Dann soll also gelten

$$\begin{aligned} P(\mathbf{x}|\mu, \sigma^2) \quad &\propto \quad \frac{\exp\left(-\frac{1}{2\sigma^2}\sum_i (x_i - \mu)^2\right)}{\sigma^n} = \\ &\frac{1}{\sigma^n} \exp\left[-\frac{n(\bar{x}-\mu)^2 + (n-1)s^2}{2\sigma^2}\right] \end{aligned} \tag{127}$$

mit

$$\bar{x} = \frac{1}{n}\sum_i x_i, \quad s^2 = \frac{1}{n-1}\sum_i (x_i - \bar{x})^2.$$

Für die a-priori-Verteilung werden die Annahmen

1. μ und σ^2 sind unabhängig voneinander,
2. $P(\mu) \propto 1$, d.h. die a-priori-Verteilung ist uneigentlich (s. Abschnitt 6.4),
3. $P(\sigma) \propto 1/\sigma$ (s. Abschnitt 6.4)

Die gemeinsame a-posteriori-Dichte ist dann

$$P(\mu, \sigma^2|\mathbf{x}) \propto \frac{1}{\sigma^{n+2}} \exp\left[-\frac{(n-1)s^2 + n(\bar{x}-\mu)^2}{2\sigma^2}\right]. \tag{128}$$

Von dieser Verteilung kann man zu einer Verteilung etwa nur für σ^2 übergehen; dies ist die *marginale a-posteriori-Verteilung*. Man erhält sie durch Integration über μ:

$$\begin{aligned} P(\sigma^2|\mathbf{x}) &= \int_{-\infty}^{\infty} P(\mu, \sigma^2|\mathbf{x}) d\mu \\ &= \frac{1}{\sigma^{n+2}} \exp\left[-\frac{(n-1)s^2}{2\sigma^2}\right] \int_{-\infty}^{\infty} \exp\left[\frac{n(\mu-\bar{x})^2}{2\sigma^2}\right] d\mu \\ &= \frac{1}{\sigma^{n+2}} \exp\left[-\frac{(n-1)s^2}{2\sigma^2}\right], \end{aligned} \tag{129}$$

und es läßt sich zeigen, dass

$$\mathbb{E}(\sigma^2|\mathbf{x}) = \frac{n-1}{n-3}s^2, \quad \mathbb{V}(\sigma^2|\mathbf{x}) = \frac{2(n-1)^2 s^4}{(n-3)^2(n-5)}. \tag{130}$$

Die Bewertung der Hypothese geschieht nun nach Maßgabe des Wertes von

$$\Omega = \frac{P(H_0|x)}{P(H_1|x)} = \frac{P(H_0|x)}{1 - P(H_0|x)}.$$

Diese *odds* können als Maß der Evidenz von H_0, gegeben die Daten x, betrachtet werden. Für $P(H_0|x) = \alpha = .05$ erhält man

$$\Omega = \frac{P(H_0|x)}{1 - P(H_0|x)} = \frac{.05}{.95} = .0526.$$

Umgekehrt erhält man für $P(H_1|x) = .05$ die Odds $\Omega = \alpha/(1-\alpha) = 19$. Man kann sagen, dass für $\Omega \leq .053$ die Daten stark gegen H_0 und entsprechend stark für H_1 sprechen, und für $\Omega > 19$ sprechen sie stark für H_0 und gegen H_1.

6.1.3 Bayessche Entscheidungen

Die Bewertung von Hypothesen durch Inspektion von a-posteriori-Verteilungen kann unter bestimmten Bedingungen nicht hinreichend sein, z.B. dann,wenn Entscheidungen gefällt werden müssen. Um hier Bayesianisch vorgehen zu können,

wird eine Verlustfunktion L (loss) eingeführt, die auf der Menge der möglichen Handlungen (actions) $\mathcal{A} = \{a_1, \ldots, a_k\}$ definiert ist. Die Entscheidung für eine Hypothese wird dann nach Maßgabe des erwarteten minimalen Verlusts gefällt. Komplementär zur Verlust kann auch eine Nutzenfunktion (utility function) eingeführt werden, und die Entscheidung wird nach Maßgabe des maximalen erwarteten Nutzens gefällt.

Beispiel 6.3 Monty Hall Der Teilnehmer der Show kann zwischen drei Türen wählen; hinter einer steht ein Auto, hinter den übrigen eine Ziege (oder nichts). Nach der Wahl einer Tür öffnet der Showmaster eine Tür, hinter der eine Ziege steht, und die "Versuchsperson" (Vp) soll entscheiden, ob sie bei ihrer unrsprünglichen Wahl bleiben will oder ob sie die Tür wechseln will. Die Lösung ist bekannt (mit der Wahrscheinlichkeit 2/3 wählt man eine Tür, hinter der der Preis *nicht* ist, woraus sofort folgt, dass es besser ist, zu wechseln als nicht zu wechseln), aber es soll die Mechanik Verlustminimierung bzw. der Nutzenmaximierung illustriert werden.

Man hat drei Hypothesen θ_i, $i = 1, 2, 3$; θ_i bedeutet, dass das Preishinter der i-ten Tür T_i steht (gemäß einer vorher festgelegten Durchnummerierung der Türen). Die a-priori-Verteilung ist dann

$$P(\theta_i) = \frac{1}{3}, \quad i = 1, 2, 3 \tag{131}$$

Eine direkte Lösung findet man, wenn man eine Nutzenfunktion annimmt:

$$U(a, a_i) = \begin{cases} 1, & a = a_i \\ 0, & a \neq a_i \end{cases} \tag{132}$$

d.h. wählt man (a) die Tür a_i und befindet sich das Preis hinter dieser Tür, so ist der Nutzen gleich 1, und 0 sonst. Der erwartete Nutzen für a_i ist dann

$$\mathbb{E}[U(a_1, \theta)] = 1 \cdot \frac{1}{3} + 0 \cdot \frac{1}{3} + 0 \cdot \frac{1}{3} = \frac{1}{3}, \tag{133}$$

und analog für die übrigen Handlungen. Alle drei Möglichkeiten haben zunächst den gleichen erwarteten Nutzen.

Es werde nun die Tür T_1 gewählt. Ist der Preis hinter T_1, so wählt Monty zufällig zwischen den Türen T_2 und T_3; die Wahl von T_j entspricht einer Messung x_j, und es gilt

$$P(x_2|\theta_1) = P(x_3|\theta_1) = \frac{1}{2}.$$

Ist der Preis nicht hinter der Tür T_1, so muß Monty T_3 wählen, wenn der Preis hinter T_2 ist, und er muß T_2 wählen, wenn der Preis hinter T_3 ist, also

$$\begin{aligned} P(x_2|\theta_2) &= 0, \qquad P(x_3|\theta_2) = 1 \\ P(x_2|\theta_3) &= 1, \qquad P(x_3|\theta_3) = 0 \end{aligned}$$

Für die a-posteriori-Wahrscheinlichkeiten gilt dann

$$P(\theta_1|x_3) \quad \propto \quad P(x_3|\theta_1)P(\theta_1) = \frac{1}{2} \cdot \frac{1}{3} = \frac{1}{6} \tag{134}$$

$$P(\theta_2|x_3) \quad \propto \quad P(x_3|\theta_2)P(\theta_2) = 1 \cdot \frac{1}{3} = \frac{1}{3} \tag{135}$$

$$P(\theta_3|x_3) \quad \propto \quad P(x_3|\theta_3)P(\theta_3) = 0 \cdot \frac{1}{3} = \frac{1}{3} \tag{136}$$

Die erwarteten Nutzen für die a_i sind dann

$$\mathbb{E}[U(a_1,\theta)] \propto \frac{1}{6},\ \ \mathbb{E}[U(a_2,\theta)] \propto \frac{1}{3},\ \ \mathbb{E}[U(a_3,\theta)] = 0,$$

so dass die Vp zur Tür T_2 wechseln sollte. Da die Nummerierung beliebig ist, gilt der Befund für jede anfängliche Wahl, d.h. das Wechseln ist stets besser als das Nichtwechseln. □

Beispiel 6.4 Signalentdeckung In einem Experiment zur Psychophysik wird in einem einzelnen Versuch entweder ein Stimulus gezeigt ($\theta = 1$) oder nicht gezeigt ($\theta = 0$) mit einem so geringen Kontrast, dass der Stimulus nicht in allen Fällen gesehen wird. Die Vp richtet sich nach einem internen "Messwert" x, für den

$$x = \mu(\theta) + \varepsilon, \tag{137}$$

wobei $\mu(0) = \mu_0$ unabhängig von μ. Die Vp trifft die Entscheidung $a = 1$: es sei ein Stimulus präsentiert worden, oder $a = 0$: es sei kein Stimulus gezeigt worden. Die Vp gehe von a-priori-Wahrscheinlichkeiten $\pi_0 = P(\theta = 0)$, $\pi_1 = P(\theta = 1^{)} = 1-\pi_0$, und die Werte der Verlustfunktion sind in der folgenden Tabelle gegeben. L_{01} und

	state of nature	
	$\theta = 0$	$\theta = 1$
$a = 0$	L_{00}	L_{01}
$a = 1$	L_{10}	L_{11}

L_{10} sind Verluste für ein "miss", d.h. für das Nichtentdecken eines tatsächlich gezeigten Stimulus bzw. für einen falschen Alarm, während L_{00} und L_{11} negative Verluste, also Gewinne, für die Entdeckung eines tatsächlich gezeigten Stimulus bzw. für die Nichtentdeckung eines nichtgezeigten Stimulus sind. Die erwarteten Verluste für die beiden Handlungen sind

$$\mathbb{E}[L(0,\theta)] = L_{00}\pi_0 + L_{01}\pi_1 \tag{138}$$
$$\mathbb{E}[L(1,\theta)] = L_{10}\pi_0 + L_{11}\pi_1 \tag{139}$$

Wird nach Maßgabe des erwarteten Verlusts entschieden, so wählt die Vp $a = 1$, wenn $\mathbb{E}[L(1,\theta)] < \mathbb{E}[L(0,\theta)]$, und $a = 0$ sonst, d.h. für $a = 1$

$$L_{10}\pi_0 + L_{11}\pi_1 < L_{00}\pi_0 + L_{01}\pi_1,$$

d.h. wenn

$$\frac{(L_{11} - L_{01})\pi_1}{(L_{00} - L_{10})\pi_0} < 1. \tag{140}$$

Nun werde die Beobachtung x gemacht. Nach Bayes hat man

$$P(\theta = 0|x) \propto P(x|\theta = 0)\pi_0 \propto \pi_0 \exp\left[-\frac{(x-\mu_0)^2}{2\sigma^2}\right] \tag{141}$$
$$P(\theta = 1|x) \propto P(x|\theta = 1)\pi_1 \propto \pi_1 \exp\left[-\frac{(x-\mu_1)^2}{2\sigma^2}\right] \tag{142}$$

Jetzt können die erwarteten Verluste, gegeben x, berechnet werden:

$$\mathbb{E}[L(0,\theta|x)] = L_{00}P(\theta = 0|x) + L_{01}P(\theta = 1|x) \tag{143}$$
$$\mathbb{E}[L(1,\theta|x)] = L_{10}P(\theta = 0|x) + L_{11}P(\theta = 1|x) \tag{144}$$

Substituiert man hier die Ausdrücke (141) und (142), so erhält man nach einigen Umformungen die der Ungleichung $\mathbb{E}[L(0,\theta)|x] > \mathbb{E}[L(1,\theta|x)]$ entsprechende Ungleichung

$$x > \frac{\mu_0+\mu_1}{2} - \frac{\log\left(\frac{\pi_1(L_{01}-L_{11})}{\pi_0(L_{10}-L_{00})}\right)}{(\mu_1-\mu_0)/\sigma^2}, \tag{145}$$

d.h. die Vp minimiert den Verlust, wenn sie eine Entdeckensantwort gibt, wenn die Aktivierung größer als ein kritischer Wert ist, der durch die rechte Seite dieser Ungleichung bestimmt ist. □

Punktschätzungen und Verlustfunktionen Eine Punktschätzung ist die Schätzung $\hat{\theta} \in \mathbb{R}$ eines freien Parameters θ. Auch hier lassen sich Verlustfunktionen definieren:

Definition 6.3 *Die Funktion*

$$L(\hat{\theta},\theta) = (\hat{\theta}-\theta)^2 \tag{146}$$

heißt quadratische Verlustfunktion.

Der erwartete Verlust für eine quadratische Verlustfunktion ist

$$\mathbb{E}[L(\hat{\theta},\theta)] = \int (\hat{\theta}-\theta)^2 P(\theta)\,d\theta = \int (\hat{\theta}-\mathbb{E}(\theta)+\mathbb{E}(\theta)-\theta)^2 P(\theta)\,d\theta$$

Setzt man $u = \hat{\theta} - \mathbb{E}(\theta)$ und $v = \theta) - \mathbb{E}(\theta)$, so so erhält man wegen $(u-v)^2 = u^2+v^2-2uv$ und einigen Vereinfachungen

$$\mathbb{E}[L(\hat{\theta},\theta)] = \mathbb{V}(\theta) + (\mathbb{E}(\theta)-\hat{\theta})^2. \tag{147}$$

Offenbar ist dieser Erwartungswert minimal, wenn $\hat{\theta} = \theta$. Dies bedeutet, dass der Mittelwert eine Schätzung ist, der den quadratischen Verlust minimalisiert.

Eine Entscheidungsregel d liefert eine Zurordnung einer Beobachtung $\mathbf{x}$ zu einer Handlung a, $d : \mathbf{x} \mapsto a$. Zu dieser Regel korrespondiert ein *Bayes-Risoko*; dies ist der erwartete Verlust über θ und $\mathbf{x}$:

$$r(d) = \mathbb{E}[L(\theta,\mathbf{x})] = \iint L(d(\mathbf{x}),\theta)P(\mathbf{x},\theta)d\mathbf{x}d\theta. \tag{148}$$

Nun kann eine Bayessche Entscheidungsregel d^* erklärt werden:

$$d^*(\mathbf{x}) = \min_a \mathbb{E}[L(a,\theta|\mathbf{x})]. \tag{149}$$

Es kann gezeigt werden, dass unter den möglichen Entscheidungsregeln die Bayessche Entscheidungsregel das kleinste Risiko impliziert.

Modelle und Entscheidungen zwischen Modellen Die Diskussion von Daten bezieht sich im Allgemeinen auf Modelle, mit denen die Daten "erklärt" werden. Eines der Standardmodelle, die hierfür herangezogen werden, ist etwa das Modell M

$$M: \quad X = \mu + \varepsilon, \quad \mathbb{E}(X) = \mu$$

und ε ein "Fehler", der unabhängig von μ ist; häufig wird zusätzlich angenommen, dass $\varepsilon \sim N(0,\sigma^2)$. μ kann eine Funktion anderer Variablen sein, etwa der Zeit, so

dass $\mu = \mu(t)$, $\mu(t)$ nun eine Funktion der Zeit. Diese Funktion stellt ein weiteres Modell dar. Die Unabhängigkeit von ε und $\mu(t)$ ist keineswegs gewährleistet, und ebensowenig gilt notwendig die Annahme $\varepsilon \sim N(0, \sigma^2$, σ^2 eine Konstante.

Ein Modell enthält freie Parameter $\theta_1, \theta_2, \ldots$, die aus den Daten geschätzt werden müssen; im oben genannten Modell M sind es im einfachsten Fall die Parameter $\theta = \mu$ und σ^2. Ist μ eine Funktion der Zeit, etwa $\mu(t) = k \exp(-\lambda t)$, k und λ Konstante, so hat man ein zweites Modell M_2 mit den freien Parametern $\theta_1 = k$ und $theta_2 = \lambda$. Man kann nun die Modelle M_1 : μ =konstant gegen M_2 testen. Ein Modell mit einer größeren Anzahl freier Parameter kann oft besser an die Daten angepasst werden als ein Modell mit einer kleineren Anzahl freier Parameter. Die Verbesserung des Fits durch eine größere Anzahl freier Parameter muß nicht notwendig bedeutsam sein, d.h. der bessere Fit kann trivial sein. Dies ist ein Sachverhalt, der beim Test berücksichtigt werden muß. Der Effekt einer größeren Anzahl freier Parameter kann durch Anwendung bestimmter Kriterien bestimt werden:

1. **Akaike Information Criterion (AIC):** Es sei $L(M; \mathbf{x})$ die (Max-)Likelihood der Daten $\mathbf{x}$, gegeben das Modell M, und $\nu(M)$ sei die Anzahl der freien Parameter des Modells. Dann ist das Kriterium AIC durch

$$\mathrm{AIC}(M) = -2 \log L(M; \mathbf{x}) + 2\nu(M). \tag{150}$$

   (Akaike (1974)). Der Term $2\nu(M)$ ist eine Art Bestrafung. Denn $L(M; \mathbf{x})$ ist, wie ja schon angemerkt, i. A. um so größer, je mehr freie Parameter ein Modell hat. Durch Subtraktion von $2\nu(M)$ wird dieser Effekt wieder aus dem Fit herausgenommen.

2. **Baysian Information Criterion (BIC):** Nach diesem Kriterium gilt

$$\mathrm{BIC}(M) = -2 \log L(M; \mathbf{x}) + \nu(M) \log n, \tag{151}$$

   wobei n der Stichprobenumfang ist. Hier geht noch der Stichprobenumfang in die Bestrafung (Penalisierung) ein.

3. **Bayesian Reference Criterion (BRC):** Dieses Kriterium wurde von Bernardo & Rueda (2002) vorgeschlagen. Die Autoren argumentieren, dass bei die Überprüfung der Hypothese H_0: $\theta = \theta_0$ als formales Entscheidungsproblem (Handlung a_0 versus a_1) behandelt werden soll. Dabei soll ein einfaches Modell M_0 betrachtet werden. Man entscheidet sich für M_0, wenn die Differenz der Verluste $L(a_0, \theta) - L(a_1, \theta)$ proportional zum Verlust $\delta(\theta_0, \theta)$ an Information ist, derentsteht, wenn man M_0 als Approximation für das komplexere Modell M wählt. Für eine gegebene a-priori-Verteilung $P(\theta)$ erhält man eine Lösung für das Entscheidungsproblem, indem man die a-posteriori-Erwartung

$$\iint \delta(\theta_0, \theta) P(\theta) d\theta$$

   betrachtet; ist diese hinreichen groß, verwirft man M_0. Der Vorteil dieses Kriteriums ist, dass nicht angenommen wird, dass unter H_0 die Wahrscheinlichkeitsmasse auf θ_0 konzentriert ist. Damit wird das Lindley-Paradoxon (Konvergenz der a-posteriori-Wahrscheinlichkeit für H_0 strebt gegen 1) vermieden. Ebenso wird im multidimensionalen Fall das *Rao Paradoxon* vermieden: univariate und multivariate Schätzungen eines Parameters können inkonsistent sein.

Eine detailierte Darstellung dieses Kriterium geht über den gegebenen Rahmen hinaus; Bernardo & Ruedas Arbeit ist aber lesenswert, da die Herleitung des Kriteriums in eine allgemeine Diskussion des Hypothesentests eingebettet ist.

Der Kern dieser Kriterien ist jeweis die Penalisierung; einfache Modelle sind den komplexeren vorzuziehen. Wie Held (2008) zu Recht anmerkt, entspricht diese Logik dem berühmten *Ockhamschen Rasierer*, also eben dem wissenschaftstheoretischen Prinzip, dass von zwei Modellen, die einen Datensatz ungefähr gleich gut erklären, das einfachere vorzuziehen ist.

6.1.4 Die Stop-Regel

Das Resultat von Signifikanz- oder Hypothesentests hängt vom Stichprobenumfang ab. Die Annahme ist im Allgemeinen, dass der Stichprobenumfang vor Beginn der Studie festgelegt wird, von sequentiellen Erhebungsschemata einmal abgesehen. In der empirischen Praxis kommt es aber vor, dass die Datenerhebung durch nicht vorhergesehene und insofern zufällige Ereignisse abgebrochen oder aber auch verlängert wird. Dies kann bedeuten, dass der Ausgang des Tests vom Zufall abhängt.

Der Einfachheit halber soll ein Bernoulli-Experiment betrachtet werden. In einem Versuchsdurchgang tritt mit einer Wahrscheinlichkeit θ das Ereignis A ein oder, mit der Wahrscheinlichkeit $1-\theta$ das Ereignis $\neg A$. Es soll geprüft werden, ob die Hypothese $\theta = \theta_0$ gilt.

Es gibt zwei Möglichkeiten:

(i) Es wird festgelegt, dass n Versuche durchgeführt werden

(ii) Die Versuchsreihe endet, wenn r-mal A eingetreten ist.

Im Falle (i) enthält der Stichprobenraum 2^n mögliche Folgen von A und $\neg A$. Sieht man von der Reihenfolge von A und $\neg A$ ab, so kann man die Ergbnisse (Elemente des Stichprobenraums) in der Form

$$(20,0),(19,1),(18,2),(17,3),\ldots,(0,20)$$

anschreiben, wobei die erste Zahl die Anzahl der As und die zweite die Anzahl der $\neg A$-s angibt.

Im Fall (ii) sieht der Stichprobenraum anders aus:

$$(r,0),(r,1),(r,2),\ldots$$

denn es muß stets r-mal A eingetreten sein, bevor das Experiment abgebrochen wird.

Je nachdem, welches Erhebungsschema gewählt wird, ergibt sich eine andere Form des Signifikanztests. Für das Schema (i) ist die Wahrscheinlichkeit für r-mal A, wenn H_0: $\theta = \theta_0 = 1/2$

$$P(X = r|n) = \binom{n}{r}\theta_0^r(1-\theta_0)^{n-r} = \binom{20}{r}\left(\frac{1}{2}\right)^n, \quad 0 \leq r \leq 20 \tag{152}$$

Für das Schema (ii) endet die Versuchsreihe, wenn A zum r-ten Male aufgetreten ist. Auf die Verteilung der A-Ereignisse vor dem r-ten Auftreten kommt es

nicht an. Es sei (r, k) das Ereignis, das tatsächlich beobachtet wird. Dann hat es $k-1$ Versuche gegeben, bei denen insgesamt $(r-1)$-mal A eingetreten ist. Die Wahrscheinlichkeit hierfür ist unter H_0

$$\binom{k-1}{r-1}\theta_0^{r-1}(1-\theta_0)^{k-1-(r-1)} = \binom{k-1}{r-1}\left(\frac{1}{2}\right)^{k+r-1}.$$

Mit der Wahrscheinlichkeit $\theta_0 = 1/2$ tritt im k-ten Versuch A ein, so dass man

$$P(X=k) = \binom{k-1}{r-1}\left(\frac{1}{2}\right)^{k+r-1}\frac{1}{2} = \binom{k-1}{r-1}\left(\frac{1}{2}\right)^{k+r} \tag{153}$$

hat. Angenommen, es sei gerade $k+r = n = 20$, n der für das Schema (i) gewählte Wert, und $r = 6$. Die Frage ist nun, ob die Nullhypothese zurückgewiesen werden muß. Für das Schema (i) muß dafür die Wahrscheinlichkeit

$$\begin{aligned} P_i^* &= P(X \leq 6|20) \\ &= P(X=5|20) + P(X=4|20) + \cdots + P(X=0|20) = .115 \end{aligned} \tag{154}$$

berechnet werden. Da $P^* = .115 > \alpha = .05$, kann die Nullhypothese $\theta_0 = 1/2$ beibehalten werden.

Für das Schema (ii) hat man

$$P_{ii}^* = P(6,14) + P(6,15) + \cdots + P(6,20) = .0319, \tag{155}$$

und hier ist $P_{ii}^* = .0319 < \alpha = .05$, so dass H_0 zurückgewiesen werden muß, wenn die Fishersche Logik angewendet wird. Wendet man den Neyman & Pearson - Test auf (i) und (ii) an, kommt man ebenfalls zu diesen widersprüchlichen Ergebnissen (Howson & Urbach (1984), p. 170).

Die tatsächlich beobachtete Folge sei $(A\neg A\neg AA\cdots A)$, und sie werde einem Statistiker vorgelegt, ohne dass ihm gesagt wird, ob sie anhand des Schemas (i) oder des Schemas (ii) erzeugt wurde. Die tatsächliche Evidenz, die ihm vorliegt, ist aber nur die Folge $(A\neg A\neg AA\cdots A)$. Hat sich der Statistiker der Orthodoxie – sei es die Fishersche oder die Neyman-Pearsonsche – verschrieben, so hat er ein Problem, denn die Annahme entweder des Schemas (i) oder des Schemas (ii) führt zu verschiedenen Bewertungen von H_0.

Howson & Urbach verweisen auf weitere Erhebungsschema. Man stelle sich einen Wissenschaftler vor, der erkunden will, ob eine Münze "fair" ist, deswegen beginnt er, die Münze zu werfen und jeweils zu notieren, ob Kopf oder Zahl oben liegen. Das tut er, bis seine Frau ihn zum Abendessen ruft. Dieser Zeitpunkt liege nicht genau fest, so dass die Anzahl der Würfe notwendig zufällit ist. Ist $A =$ Kopf und $\neg A =$ Zahl, liegt am Ende eine Folge der Form $(A\neg A\neg AA\cdots A)$ oder $(A\neg A\neg AA\cdots \neg A)$ vor; wie soll der Wissenschaftler nun entscheiden, ob die Münze fair ist, wenn er etwa den Fisherschen Signifikanztest anwenden will? Er könnte sich helfen, indem er eine Poisson-Verteilung für die Anzahl der Würfe annimmt, so dass $n = 0, 1, 2, \ldots$, aber dann muß er den Intensitätsparameter dieser Verteilung schätzen, der wiederum seine Frau, nicht aber die Münze charakterisiert. In eine ähnliche Lage kommt er, wenn er eine Annahme über die Verteilung der Dauer der Wartezeit bis zum Ruf "Dinner is ready!" macht. Vermutlich kümmert er sich nicht um diese Problematik, betrachtet n nicht als zufällige Veränderliche und berechnet einfach den Anteil r/n der Würfe, bei denen der Kopf oben gelegen hat.

Das Beispiel mag artifiziell erscheinen, spiegelt aber doch wichtige Aspekte der wissenschaftlichen Wirklichkeit wider. Statt des Münzwurfs denke man an Experimente, in denen die Häufigkeit von Reaktionen A und $\neg A$ bestimmt wird, etwa um die Wirkung eines Medikaments abzuschätzen. Für eine gegebene Dosis kann, unter der vereinfachenden Annahme, dass die Dosis entweder zu einem "Erfolg" oder zu einem "Misserfolg" führt, die Folge der Einzelversuche als Bernoulli-Folge betrachtet werden, und die Interpretation der Ergebnisfolge hängt nun von der Stop-Regel ab. Es ist ja keineswegs sicher, dass der Wert n der Anzahl der Versuche stets wirklich von vornherein festgelegt wurde, oft wird so lange experimentiert, bis man den Eindruck hat, nun eine "hinreichende" Anzahl von Versuchen beieinander zu haben, oder die Versuchsreihe wird, wie beim Wissenschaftler, der auf den Ruf "Dinner is ready!" wartet, nach einem zufällig eintretenden Ereignis (der Versuchsleiter fängt sich eine Erkältung ein, oder die Laborleitung beschließt, die Untersuchung aus nicht vorhergesehenen Kostengründen zu beenden) abgebrochen. Der orthodoxe Statistiker müßte genau wissen, nach welcher Stop-Regel vorgegangen wurde, schon um Vergleiche mit den Ergebnissen anderer Untersuchungen zhur glechen Fragestellung ziehen zu können. Da die Stop-Regel oft gar nicht genau spezifiziert ist, wird oft einfach angenommen, der Wert von n sei vorher festgelegt worden. Dass sich diese Entscheidung auf die Interpretation der Daten auswirken kann, ist oben illustriert worden.

Beim Bayesschen Ansatz wird die a-posteriori-Wahrscheinlichkeit der Hypothese(n) betrachtet,

$$P(H|E) \propto P(E|H)P(H),$$

deren Berechnung wiederum die Wahl einer a-priori-Wahrscheinlichkeit erfordert, die ja in der "Orthodoxen" Statistik gar nicht berücksichtigt wird. Es zeigt sich etwa bei einem Bernoulli-Experiment, dass sich der Faktor $K = \binom{n}{r}$ oder $K = \binom{k-1}{r-1}$ beim Bayesschen Ansatz herauskürzt. So seien n Versuche durchgeführt worden. Die Wahrscheinlichkeit für r "Erfolge" ist dann $K\theta^r(1-\theta)^{n-r}$. Unter H_0 sei $\theta = \theta_0$, so dass unter H_0 die Wahrscheinlichkeit für r Erfolge

$$K\theta_0^r(1-\theta_0)^{n-r}$$

ist. Unter der Alternativhypothese H_1 hat man für r Erfolge

$$\int_0^1 K\theta^r(1-\theta)^{n-r} f(\theta|H_1)d\theta,$$

wobei f für die a-priori-Verteilung $P(H)$ steht. Der Likelihood-Quotient ist dann

$$L(\theta_0; r, n) = \frac{K\theta_0^r(1-\theta_0)^{n-r}}{\int_0^1 K\theta^r(1-\theta)^{n-r} f(\theta|H_1)d\theta} = \frac{\theta_0^r(1-\theta_0)^{n-r}}{\int_0^1 \theta^r(1-\theta)^{n-r} f(\theta|H_1)d\theta}, \quad (156)$$

K kürzt sich heraus. Es kommt also nur auf die tatsächlich vorliegende Folge von Erfolgen und Mißerfolgen an, nicht davon, ob die Anzahl der Versuche von vornherein festgelegt wurde, oder ob so lange experimentiert wurde, bis r Erfolge eingetreten sind, etc. Wichtig ist nur, dass der Faktor nicht vom Parameter θ oder θ_0 abhängt. Man kann nun den Quotienten

$$\frac{P(H_0|E)}{P(H_1|E)} = \frac{P(E|H_0}{P(E|H_1)}\frac{P(H_0)}{P(H_1)} = L(\theta_0; r, n)\frac{P(H_0)}{P(H_1)} \quad (157)$$

betrachten. Wenn $P(H_0|E)/P(E|H_1) > 1$ kann man sich für H_0 entscheiden, wenn $P(H_0|E)/P(E|H_1) < 1$, für H_1. Diese Entscheidungsregel wäre eine einfache Form,

sich für oder gegen H_0 zu unterscheiden. Im Allgemeinen wird man komplexere Regeln betrachten; insbesondere muß für die a-priori-Verteilung $K\theta_0^r(1-\theta_0)^{n-r}$ eine Alternative gefunden werden. Hier kommt es nur darauf an, zu zeigen, dass die Stop-Regel nicht mehr von Bedeutung ist.

6.2 Die Interpretation von Wahrscheinlichkeiten

Wahrscheinlichkeiten werden gewöhnlich (i) über den Begriff des zufälligen Ereignisses, und (ii) über die Kolmogorovschen Axiome eingeführt. In der Bayesschen Statistik wird aber auch die Wahrscheinlichkeit einer Hypothese, also einer Aussage, betrachtet. Deshalb werden die Axiome auch für Aussagen bzw. für Mengen von Aussagen spezifiziert.

Es sei $S = \{a, b, c, \ldots\}$ eine Menge von Sätzen (Propositionen). Die Aussage "a enthält b" (a entails b:), in Zeichen $a \vdash b$, soll heißen, dass es nicht möglich ist, dass a wahr ist, aber b nicht. Es folgt demnach: $a \vdash a$, $a \vdash \neg\neg a$, $a \,\&\, b \vdash a$, $a \vdash a \vee a$. $a \Leftrightarrow b$ heißt a ist äquivalent b.

S sei gegenüber den Verknüpfungen $\wedge$, $\vee$, $\neg$ abgeschlossen. Dann werden die folgenden Axiome eingeführt:

1. $P(a) \geq 0$ für alle $a \in S$.
2. $P(t) = 1$, wenn t eine Tautologie ist.
3. $P(a \vee b) = P(a) + P(b)$, $a, b \in S$ und a und b sind wechselseitig inkonsistent, d.h. $a \vdash \neg b$, $b \vdash \neg a$.

Tabelle 12: Axiome der Wahrscheinlichkeit; die Bedingung $a \vdash \neg b$, $b \vdash \neg a$ für 3'. bedeutet, dass a und b sich wechselseitig ausschließen.

Kolmogorov		logische Wahrs.	
1.	$P(A) \geq 0$, $A \in \Sigma$	1'.	$P(a) \geq 0$, $a \in S$
2.	$P(\Omega) = 1$	2'.	$P(t) = 1$, t Tautologie
3.	$P(A \cup B) = P(A) + P(B)$ für $A \cap B = \emptyset$	3'.	$P(a \vee b) = P(a) + P(b)$, für $a \vdash \neg b$, $b \vdash \neg a$
4.	$P(A\|B) = P(A \cap B)/P(B)$	4'.	$P(a\|b) = P(a \wedge b)/P(b)$

Die Tabelle 12 zeigt eine Gegenüberstellung der Kolmogorovschen Axiome und der Axiome bezüglich eines Systems S von Aussagen. Jaynes (2003, p. 651) diskutiert die beiden Systeme; hier muß nicht weiter darauf eingegangen werden.

Unmittelbare Folgerungen:

$$P(\neg a) = 1 - P(a) \tag{158}$$

$$P(f) = 0, \quad f \text{ ein Widerspruch} \tag{159}$$

$$P(a) = P(b), \quad \text{wenn } a \text{ und } b \text{ logisch äquivalent, d.h. wenn } a \Leftrightarrow b \tag{160}$$

$$P(a) \leq P(b), \quad \text{wenn } a \vdash b \tag{161}$$

Satz von Bayes Man kann dann den Bayesschen Satz anschreiben:

$$P(a|b) = \frac{P(b|a)P(a)}{P(b)}, \quad p(a), P(b) > 0. \tag{162}$$

Der Satz folgt sofort aus der Definition der bedingten Wahrscheinlichkeit 4'. Üblicherweise wird $a = H$, wobei H eine Hypothese ist, und $b = E$ gesetzt, E die Daten bzw. die Evidenz.

Laplace hat den Satz in verallgemeinerter Form geschrieben: es seinen Aussagen $a, b_1, \ldots, b_n$ gegeben und es gelte $b_i \vdash \neg b_j$ für $i \neq j$, und $P(a), P(b_i) > 0$ für alle i; dann gilt

$$P(b_k|a) = \frac{P(a|b_k)P(b_k)}{\sum_{i=1}^{n} P(a|b_i)P(b_i)}. \tag{163}$$

Hier ist $\sum_{i=1}^{n} P(a|b_i)P(b_i) = P(a)$ einfach der Satz der Totalen Wahrscheinlichkeit.

Zum Wahrscheinlichkeitsbegriff: Man unterscheidet zwischen einer Definition von Wahrscheinlichkeiten als einer objekten Größe und als einer Definition von Wahrscheinlichkeit einer epistemischen und insofern subjektiven Größe. Die Interpretation der Wahrscheinlichkeit als einer epistemischen Größe geht schon auf Bayes zurück (s.a. Earman (1992)), der eine Definition von Wahrscheinlichkeiten über Wettquotienten nahelegte, die wiederum ihren Vorläufer bei Jakob Bernoulli[14] hat, der in seiner *Ars Conjectandi* (1713) Wahrscheinlichkeit als einen "Grad von Ungewißheit" – eben auch für Aussagen, d.h. Hypothesen – bezeichnete. Ungewißheit ist ein subjektives Merkmal, verschiedene Personen können in Bezug auf ein Ereignis oder auf eine Hypothese verschieden ausgeprägte Ungewißheitserlebnisse haben. Die Zuordnung von Wahrscheinlichkeiten zu Aussagen impliziert die epistemische Interpretation. Laplace (1774) charakterisierte die Wahrscheinlichkeitsrechnung als 'Kalkül für die induktive Argumentation'. Demnach sind Wahrscheinlichkeiten kognitive Größen, oder, in einem anderen Sprachgebrauch, subjektive Größen. Diese Interpretation wurde von den meisten Mathematikern und insbesondere auch Physikern geteilt (die Bernoullis, außer von Laplace von Poisson,Legendre, Gauß, Boltzmann, Maxwell, Gibbs). Wie in Abschnitt 1.2, 8, erwähnt wird, waren es Nichtphysiker – Mathematiker und Biologen – die eine Uminterpretation bewirkten.

Da Wissenschaft sich um Objektivität bemüht, ergibt sich das Bestreben, Wahrscheinlichkeiten als Repräsentation von vom Individuum unabhängiger Eigenschaften der betrachteten Ereignisse zu sehen. von Mises (1928/1957) verhalf dann dem 'objektiven' Wahrscheinlichkeitsbegriff zum Durchbruch, indem er die Wahrscheinlichkeit eines Ereignisses A als den Grenzwert $p(A) = \lim_{n\to\infty} n_A/n$ definierte, wobei n_A die Häufigtkeit des Auftretens von A bei n Versuchen ist. Diese Definition scheint sich – der Tradition des Wiener Kreises folgend – nur auf das Gegebene auszurichten, verschweigt aber, unter welchen Bedingungen ein solcher Grenzwert denn existiert. Reichenbach (1935a) hat die von Misesche Definition aufgenommen und versucht, sie durch Einführung des Begriffs des *limes partialis* retten zu können. Wegen des Verweises auf relative Häufigkeiten in der Definition ergibt sich die Redeweise von der *frequentistischen* Auffassung der Wahrscheinlichkeit. Grundannahme bei alles frequentistischen Konzeptionen der Wahrscheinlichkeit ist, dass der Grund der mangelnden Determiniertheit, die ja durch eine Wahrscheinlichkeit ausgedrückt wird, in nicht weiter bestimmten – in welchem Sinne auch immer – physikalischen Prozessen liegt. Jaynes (2003) zitiert Cramér (1946), für den es Axiom ist, dass jeder zufälligen Veränderlichen eindeutig eine Wahrscheinlichkeitsverteilung zugeordnet ist. Man denke an die Münze und den

[14] 1654 - 1705

Würfel. Wahrscheinlichkeiten sollten dann nach Cramér als physikalische Konstanten betrachtet werden, die Frage nach ihrem numerischen Wert folge nicht aus der Wahrscheinlichkeitstheorie, sondern eben aus physikalischen Betrachtungen. Popper (1957, 1959, 2001) hat in diesem Zusammenhang eine Interpretation von Wahrscheinlichkeiten als *Propensitäten* vorgeschlagen, – eine Interpretation, die intensiv (und mit negativem Ausgang für die Poppersche Theorie) diskutiert worden ist. Jaynes (2003, 318) gibt eine ausführliche, kritische Diskussion dieser Interpretation (nicht explizit der Popperschen Propensitätsinterpretation), indem er auf physikalische Aspekte der Mechanik des Münz- und Würfelwurfs eingeht, und kommt zu dem Schluß, dass es bei diesem (u.a. Cramérs) Ansatz um einen Zirkelschluß handelt. Jaynes unterscheidet dabei noch einmal zwischen der Biologie und der Quantenmechanik. In der Biologie würde i.A. so lange gesucht, bis kausale Faktoren für Prozesse, die anfangs probabilisitsch erscheinen, gefunden werden. In der Quantenmechanik (QM) allerdings würde anders gedacht, so Jaynes (p. 328). Jaynes betrachtet den photoelektischen Effekt, demzufolge Elektronen aus einer Metalloberfläche gestossen werden, wenn Licht auf die Oberfläche gelenkt wird, – aber nicht immer. Licht ist demnach *ein* Faktor, das den Effekt erzeugt, aber wohl nicht stets. Würde in der QM wie in der Biologie argumentiert, müßte man nun nach weiteren kausalen Faktoren suchen. Aber:

> "What is done in quantum theory today is just the opposite; when no cause is apparent one simply postulates that no cause exists – ergo, the laws of physics are indeterministic an can expressed only in probability form. The central dogma is that the light determines not whether a photoelectron will appear,but only the probability that it wil appear." (Jaynes (2003, p. 328))

Jaynes fährt fort mit der Behauptung, dass Biologen deshalb ein mechanistisches Bild von der Welt hätten, weil man sie dazu erzogen habe, in deterministischen Bezügen zu Denken: "they continue to use the full power of their brains to search for them" – gemeint sind die kausalen Faktoren der Prozesse, die sie studieren. Aber die (Quanten-)Physiker haben eine andere Sozialisierung erlebt:

> "Quantum physicists have only probability laws because for two generations we have been indoctrinated not to believe in causes – and so we hve stopped looking for them. Indeed, any attempt to search for the causes of microphenomena is met with scorn and a charge of professional incompetence and 'obsolete mechanistic materialism'. Therefore, in order to explain the indeterminacy in curent quantum theory we need not suppose there is any indeterminacy in Nature; the mental attitude of quantum physicists is already sufficient to guarantee it. " (Jaynes (2003, p. 328))

Jaynes sieht Niels Bohr als Begründer dieser Theorie (Wissenschaftshistoriker sehen allerdings eher Heisenberg als Hauptprotagonisten). Nach Bohr darf gar nicht gefragt werden, was wirklich geschieht, man könne eben nur nach den Wahrscheinlichkeiten fragen. Jaynes vergleicht diese Redeweisen mit Orwellschem Neusprech, die Quantentheorie sei in einer zirkulären Argumentation gefangen: die Wahrscheinlichkeiten in der Quantentheorie drückten Unvollständigkeit des Wissens aus. Man muß hier bedenken,dass hbier kein Mathematike oder Philosoph spricht, sondern dass Jaynes Physiker ist. Generell seien Wahrscheinlichkeiten also als *epistemische Größen* aufgefasst. In Mortensen, Wissenschaftstheorie III, Kapitel 2,

findet man eine etwas ausführlichere Diskussion der verschiedenen Wahrscheinlichkeitsbegriffe, und in Wissenschaftstheorie IV, Abschnitt 3.3 findet man eine kurze Darstellung der Kritik an der Kopenhagener Deutung, s. a. Passon (2006).

6.3 Induktion

Francis Bacon (1561 – 1626), einer der Begründer des modernen Empirismus, hat bekanntlich die Induktion als ein zentrales Mittel zur Gewinnung von Kenntnissen gehalten, wobei er aber nicht einem reinen Verifikationismus das Wort redete, sondern der Falsifikation durch Gegenbeispiele einen großen Wert zuordnete. Insbesondere seit David Hume kämpfen aber die Philosophen mit der Idee der Induktion. Howson (2000) zitiert Broad (1952, 143):

> "May we venture to hope that when Bacon's next centenary is celebrated thegreat work which he set going will be completed; and that Inductive Reasoning, which has long been the glory of science, will have ceased to be the scandal of philosophy?"

6.3.1 Das Zirkularitätsargument – aussichtslos

Fisher spricht im Zusammenhang mit dem Signifikanztest von induktivem Schließen, Neyman & Pearson ziehen es vor, von induktivem Verhalten zu sprechen. Jeffreys (1939/1961/2003) beginnt das erste Kapitel seines Klassikers *Theory of Probability* mit

> "The fundamental problem of scientific progress, and a fundamental one of everyday life, is that of learning from experience. Knowledge obtained in this way is partly merely description of what we have already observed, but part consists of making inferences from past experiences to predict future experience. This part may be called generalization or induction."

Seine Beispiele sind schlagend genug: die Vorhersagen des *Nautical Almanach* von Planetenpositionen, die Schätzungen eines Ingenieurs der Leistungen eines Dynamos oder eines Statistikers der Wirksamkeit eines Düngemittels; alle diese Vorhersagen beruhen auf den Erfahrungen der Vergangenheit. Wie Jeffreys weiter ausführt, werden diese Vorhersagen nicht durch die übliche – deduktive – Logik gedeckt, die nur drei Einstellungen gegenüber einer Aussage erlaube: sie könne entweder definitiv bewiesen werden, oder ihre Negation könne bewiesen werden, oder man könne gar nichts über sie aussagen. Die der Wissenschaft zugrundeliegende Logik sei induktiv und einem reinen Mathematiker in seiner offiziellen Rolle als reiner Mathematiker gar nicht verständlich, – in seiner inoffiziellen Rolle als Mitmensch könne er die induktive Logik natürlich sehr wohl verstehen. Die Versuche, den deduktiven Ansatz der Mathematik und die induktiven Methoden der Wissenschaft zusammenzubringen enthielten oft die Tendenz, so Jeffreys, die induktiven Ansätze der Wissenschaft auf die deduktive Logik zu reduzieren, "which is the most fundamental fallacy of all: it can be done only by rejecting its chief feature, induction." (Jeffreys (2003), p. 2) Jeffreys führt weiter aus, dass man üblicherweise das jeweils einfachste Gesetz wähle, um Sachverhalte zu erklären bzw. vorauszusagen (etwa das tatsächliche Funktionieren der Straßenbahn anhand der

Gesetze der Elektrodynamik), dass die Wahl eines Gesetzes auf einem 'reasonable degree of belief' basiere, und

> "the fact that deductive logic provides no explanation of the choice of the simplest law is an absolute proof that deductive logic is grossly inadequate to cover scientific and practical requirements" (Jeffreys (2003), p. 5).

Jeffreys hat einen objektiven Bayesianismus vertreten, was nach den vorangegangenen Zitaten nicht verwundert. Er war auch kein Philosoph, sondern Geophysiker, und dass sein Buch über Wahrscheinlichkeitstheorie ein Klassiker wurde, hängt vermutlich damit zusammen, dass er eben empirisch arbeitender Wissenschaftler war. Karl Popper dagegen war Philosoph, der seine These, alles wissenschaftliche Denken sei Deduktion, aus der Humeschen Skepsis gegenüber der Induktion entwickelte und zu seinem philosophischen Mantra machte (s. Abschnitt 6.3.2). In der neueren Philosophie gilt David Hume als Begründer der Ablehnung der Induktion, allerdings wird seine Originalität hier bestritten; Weintraub (1995) findet die wesentlichen Argumente bereits bei Sextus Empiricus (ca 200 nC), der sich wiederum auf den Skeptiker Pyrrhon von Elis (360 - 270) bezog. Es genügt hier, die Ansichten David Humes, wie in seinem *Enquiries Concerning Human Understanding* 1739 formuliert, zu betrachten:

> Converning matter of fact and existence [] there are no demonstrative arguments [] since it implies no contradiction that the course of nature may change [] and the trees will flourish in in December [] All arguments concerning existence are founded on the relation of cause and effect [] Our knowledge of that relation is derived entirely from experience [] All our experimental conclusions proceed upon the supposition that the future will be conformable to the past. To endeavour, therefore, the proof of this last supposition by probable arguments, or arguments regarding existence, must be evidently going in a circle, and taking that for granted, which is the very point in question.
>
> (vergl. Weintraub 1995, p. 461)

Dies ist die *Zirkularitätsthese*: Der Schluß von der Vergangenheit auf die Zukunft kann durch nichts aus den Erfahrungen der Vergangenheit gerechtfertigt werden, denn eine Veränderung der Bedingungen in der Zukunft ist immer denkbar, führt auf keinen Widerspruch und kann mit a-priori-Argumenten nicht widerlegt werden, und Folgerungen aus der bisherigen Erfahrung basieren eben auf dem, was in der Vergangenheit geschehen ist und sind insofern zirkulär. Auch die Aussage, dass *wahrscheinlich* die Bedingungen der Zukunft denen der Vergangenheit ähneln, kann nicht gefolgert werden.

Nach Hume sind alle Argumentationen entweder 'demonstrative' Argumentationen, was für 'deduktive' Argumentionen steht, und in probabilistische Argumentationen, die für Verallgemeinerungen kausaler Argumente stehen.

Humes Argument zufolge muß eine induktive Schlußfolgerung aus der Erfahrung abgeleitet werden ('derived entirely from experience'), denn eine deduktive Begründung kann ja gerade nicht geliefert werden. Aber die Erfahrung ist in der Vergangenheit erworben worden. Damit man von der Erfahrung, die in der Vergangenheit gewonnen wurde, induktiv auf die Zukunft schließen kann, muß notwendig angenommen werden, dass die Zukunft der Vergangenheit ähnelt, dh dass die Bedingungen, die die Daten (die Erfahrung) in der Vergangenheit erzeugt haben,

auch in der Zukunft in der gleichen Weise wirken. Damit die Sonne morgen wieder aufgeht, müssen morgen die Bedingungen, unter denen die Sonne überhaupt aufgehen kann, wieder gegeben sein. Diese Annahme kann nicht aus den Erfahrungen abgeleitet werden, da diese ja eben nur in der Vergangenheit und den Bedingungen der Vergangenheit gewonnen wurden. So kommt Hume zu der Charakterisierung von induktiven Schlußfolgerungen als kontingenten Folgerungen, und deduktiven Schlüssen als notwendigen Folgerungen. So kommt man zu dem Schluß, dass deduktive Schlüsse niemals kontingente Folgerungen, etwa meteorologische Vorhersagen, stützen können, oder dass man aufgrund deduktiver Schlüsse allein das Versagen des Motors bei einer Autofahrt vorhersagen kann. Induktive Schlüsse vermögen dies, weil sie *ampliativ* (Peirce) sind: sie vergrößern gewissermaßen unsere Erfahrung, während Deduktion unser Wissen nur ordne und rearrangiere, ohne etwas zum Inhalt hinzuzufügen (Vickers (2009)).

Howson (2000) diskutiert acht der gängigsten Argumente gegen die Zirkularitätsthese und kommt zu dem Schluß, dass keines von ihnen das Humesche Argument entkräftet. Der Schluß von Erfahrungen der Vergangenheit auf zu erwartende Erfahrungen in der Zukunft erfordert eine von den Erfahrungen der Vergangenheit unabhängige Annahme über die Gleichartigkeit der Zukunft, – diese Annahme läßt sich nicht durch Erfahrung begründen, auch wenn sie sich in der Wissenschaft durchaus bewährt hat; dies begründet Broad's oben zitierte Aussage über die Induktion als der 'glory of science' und dem 'scandal of philosophy'.

6.3.2 Poppers Argumente – nicht zwingend

Poppers Falsifikationstheorie kann etwas plakativ als Elaborat der Humeschen These, dass Induktion nicht deduktiv begründet werden kann, gesehen werden. Die Bayessche Statistik ist wesentlich induktiv: Wegen

$$P(H|E) \propto P(E|H)P(H)$$

wird ein Schluß von der empirischen Evidenz E auf die Wahrscheinlichkeit der Hypothese H vorgenommen. Der Schluß ist induktiv, weil die Aussage H über die konkret gegebene Evidenz E hinausgeht. Wenn aber Induktion in Wirklichkeit unmöglich ist, so kann dieser Schluß nicht gerechtfertigt werden, – etwas ist falsch an der Bayesschen Statistik.

Die Unhaltbarkeit (Popper) der induktiven Logik: So liefert Popper (2002) im Neuen Anhang $^*VII$ seiner *Logik der Forschung* einen formalen Beweis, dass "die Idee einer induktiven Wahrscheinlichkeit unhaltbar ist". Insbesondere versucht Popper zu zeigen, dass in einem unendlichen Universum die Wahrscheinlichkeit jedes nichttautologischen allgemeinen Gesetzes gleich Null ist, – weshalb Popper Wert darauf legt, dass der Bestätigungsgrad einer Hypothese für ihn keine Wahrscheinlichkeit sein kann. Unter Wahrscheinlichkeit versteht er

> "entweder die *absolute* logische Wahrscheinlichkeit des allgemeinen Gesetzes oder seine *relative* Wahrscheinlichkeit *in bezug auf irgendwelche als gegeben angenommenen Sätze über Ereignisse (Tatsachenfeststellungen)*, d.h. in bezug auf einen *besonderen Satz (singulären Satz)* oder eine endliche Konjunktion von besonderen Sätzen. Wenn also a unser Gesetz ist und b irgendeine Tatsachenfeststellung, dann behaupte ich
>
> $$p(a) = 0 \qquad (1)$$

und auch

$$p(a|b) = 0. \qquad (2)$$

" (Popper (2002), p. 313 - 314, Kursivsetzungen von Popper)

Begründung: über $\lim_{n\to\infty} a^n = 0$ im einfachen Fall. Poppers allgemeine Begründung ist zu länglich, um sie hier wiedergeben zu können. Insbesondere hat Howson (1987) eine detaillierte Kritik des Popperschen Arguments geliefert: während die Poppersche Argumentation zunächst einmal formal korrekt ist, sind es implizite Annahmen, die der Argumentation die Wucht nehmen. So hat, wie Howson ausführt, Popper übersehen, dass jede Wahl einer a-priori-Verteilung die Annahme einer Hypothese bedeutet (Popper hat sich im Wesentlichen auf die Klassische Statistik bezogen), die sich der empirischen Überprüfung entzieht. Ein Beispiel sei etwa die a-priori-Wahrscheinlichkeit für die Allgemeine Relativitätstheorie einerseits und die Newtonsche Theorie andererseits. Die Menge der Möglichkeiten ist hier eine Klasse von Universen. Howson zeigt dann, dass Poppers Wahl von a-priori-Wahrscheinlichkeiten ihn in Widerspruch zu seiner eigenen Behauptung, Induktion sei unmöglich, bringt. Eine vollständigere Darstellung der Howsonschen Argumente ist hier allerdings nicht möglich.

Gegen die Bayessche Statistik: Im "Neuen Anhang $^*xvii$" *Argumente gegen die Bayessche induktive Wahrscheinlichkeit* in Popper (2002) liefert Popper einen Beweis für die Unzulänglichkeit der Bayesschen Statistik. Nach Popper ist das Ziel des Bayesschen Ansatzes, für eine Theorie t möglichst hohe Wahrscheinlichkeiten zu erreichen, so dass $1/2 \ll p(t,b) < 1$, – der Fall $p(t,b) = 1$ sei grundsätzlich nicht zu erreichen. b sind 'empirische Prüfsätze'. Dieses Ziel sei, so Popper nicht zu erreichen.

Zum Beweis geht Popper von n Theorien $t_1, \ldots, t_n$ aus, denen nach dem Indifferenzprinzip gleiche Wahrscheinlichkeiten zugeordnet werden:

$$p(t_1) = \cdots = p(t_n) \leq \frac{1}{n}. \qquad (164)$$

(Warum Popper "$\leq$" statt "$=$" schreibt, wird nicht begründet; er will wohl zum Ausdruck bringen, dass die Wahrscheinlichkeit nicht größer als $1/n$ sein kann). Dann wird angenommen, dass wegen empirischer Prüfsätze b einige, etwa f, der Theorien widerlegt werden. Für die $m = n - f$ verbleibenden Theorien $p(b|t_i) = 1$ gilt (die b sind dann Implikationen der t_i) und da weiter

$$p(t_i \wedge b) = p(t_i)p(b|t_i) = p(t_i)$$

gelte (weil $p(b|t_i) = 1$ angenommen wird, d.h. die b *als Prüfsätze* werden von t_i impliziert), folge

$$p(t_i|b) = p(t_i \wedge b)/p(b) = p(t_i)/p(b), \qquad (165)$$

woraus, zusammen mit (164),

$$p(t_1|b) = p(t_2|b) = \cdots = p(t_m|b) \leq \frac{1}{m} \qquad (166)$$

folge. Die m *nicht-falsifizierten* Theorien seien also gleichwahrscheinlich, und diese Wahrscheinlichkeit sei höchstens gleich $1/m$. Daraus folge

$$\text{Wenn } 0 < p(t_i|b) < 1, \text{ dann } p(t_i|b) \leq 1/2. \qquad (167)$$

Diese Folgerung kann nur gelten, wenn $m \geq 2$; Popper führt aus, dass der Fall $m = 2$ höchst unwahrscheinlich sei, es blieben normalerweise mehr als 2 Theorien als nicht falsifiziert übrig. Weiter sei die Annahme der Gleichverteilung nicht notwendig, sondern nur hinreichend, und er schließt mit dem Resumé: die Bayessche Theorie, angewendet auf "empirisch-wissenschaftliche universelle Aussagen ist völlig irrelevant". Nur die Falsifikation als Bewertung wissenschaftlicher Aussagen besage etwas.

Beweis der Unmöglichkeit induktiver Wahrscheinlichkeit: Popper & Miller (1983) liefern ein formales Argument, demzufolge Induktion unmöglich sein sollte. Es werde zunächst angenommen[15], dass (i) $\{H, K\} \models E$, d.h. dass die Hypothese H zusammen mit dem Hintergrundswissen K den Beobachtungssatz E impliziert. Weiter gelte (ii) $0 < P(H, K) < 1$, (iii) $0 < P(E, K) < 1$. Der Satz von Bayes impliziert dann

$$P(H|E\&K) = P(H|K)/P(E|K). \tag{168}$$

Wendet man darauf (ii) und (iii) an, so folgt

$$P(H|E\&K) > P(H|K). \tag{169}$$

Diese Aussage wird üblicherweise so gedeutet, dass die Evidenz E die Hypothese H *inkrementell bestätigt*. Popper & Miller argumentieren aber, dass die Aussage (169) gerade nicht bedeutet, dass E die Hypothese inkrementell bestätigt.

Um ihre Behauptung zu beweisen, machen sie von dem Sachverhalt Gebrauch, dass für irgend eine Hypothese H und Evidenz E,

$$H \equiv (H \vee E)\&(H \vee \neg E) \tag{170}$$

ist, d.h. H ist äquivalent der Aussage, dass H oder E und H oder $\neg E$ gilt. Andererseits impliziert E die Disjunktion $H \vee E$; es muß nun, nach Popper & Miller, gefragt werden, was E für $H \vee \neg E$ bedeutet. Nach Popper & Miller repräsentiert $H \vee \neg E$ den Teil von H, "der über E hinausgeht". Sie beweisen nun zuerst das Lemma

$$P(\neg H|E\&K)P(\neg E|K) = P(H \vee \neg E|K) - P(H \vee \neg E|E\&K). \tag{171}$$

Beweis:

$$\begin{aligned} P(\neg H|E\&K)P(\neg E|K) &= (1 - P(H|E\&K))(1 - P(E|K) \\ &= 1 - P(E|K) - P(H|E\&K) \\ &= +P(H|E\&K)P(E|K) \\ &= [1 - P(E|K) - P(H\&E|K)] - P(H|E\&k) \end{aligned} \tag{172}$$

Nun gilt

$$P(E|K) = P(H\&E|K) + P(\neg H\&E|K),$$

Dann hat man

$$P(E|K) - P(H\&E|K) = P(\neg(H \vee \neg E)|K) = 1 - P(H \vee \neg E),$$

so dass in (172)

$$[1 - P(E|K) - P(H\&E|K)] = P(H \vee \neg E|K).$$

[15] In der Notation von Earman (1992); Poppers & Millers Notation ist ein wenig idiosynkratisch.

Da $P(H|E\&K) = P(H \vee \neg E|E\&K)$ folgt (171). □

Eine direkte Folgerung von (171) ist nun

$$P(H|E\&K) \neq 1 \neq P(E|K) \Rightarrow P(H \vee \neg E|E\&K) < P(H \vee \neg E|K). \qquad (173)$$

Wenn also die harmlosen Bedingungen auf der linken Seite erfüllt sind, so zeigt die rechte Seite, ist der konfirmatorische Effekt von E negativ. Diese Beziehung – formal korrekt hergeleitet – sei verheerend für jede induktive Interpretation von Wahrscheinlichkeiten, – es gebe nur Deduktion, aber keine Induktion.

Das Argument hat eine größere Pro- und Kontra-Diskussion ausgelöst, die hier nicht *in extenso* wiedergegeben werden kann. Der Kern des Arguments liegt allerdings in der Interpretation von $A_1 = H \vee \neg E$; dieser Ausdruck enthalte alles, was über E hinausgeht. Um die Bedeutung dieser Aussage klarer zu fassen, kann man mit Redhead (1985) die Menge $Cn(H \vee E)$ einführen: dies ist die Menge aller Aussagen, die logisch von $H \vee E$ impliziert werden. Da $A_2 = H \vdash E$ vorausgesetzt wurde, hat man nun

$$Cn(A_2) = Cn(E) \subset Cn(H).$$

Redhead fragt, was die Behauptung, $H \vee \neg E$ enthalte alles, was über E hinausgeht, bedeuten könne. Eine Interpretation wäre, es bedeute, dass jede Folgerung, die nicht aus A_2 ableitbar ist, aus A_1 ableitbar ist. Das ist nicht möglich, denn

$$Cn(H) \neq Cn(A_1) \cup Cn(A_2); \qquad (174)$$

d.h. $Cn(H)$ sind alle Implikationen von H, die weder aus A_1 noch aus A_2, jeweils für sich genommen, ableitbar sind. So sei E' eine Implikation von H derart, dass ihre Wahrheit oder Falschheit nicht von der Wahrheit oder Falschheit von E abhängt, so dass $E \nvdash E'$ und $\neg E' \nvdash E$. Dann auch $A_1 \nvdash E'$ und $A_2 \nvdash E'$. E' kann nur aus A_1 *und* A_2 abgeleitet werden. Die Existenz von E' bedeutet aber, das die Behauptung Poppers & Millers, A_1 nicht in dieser Allgemeinheit gelten kann; A_1 ist eine schwache Aussage in dem Sinne, dass sie nur einen kleinen Teil der Implikationen enthält, die nicht deduktiv aus E folgt. Was Popper & Miller gezeigt haben, ist, dass nur der Teil von A_1 und A_2, die zusammen H implizieren, für E gegenindikativ sind, und dies reicht nicht, um die induktive Stützung von H durch E zu negieren. Popper & Millers Beweis ist interessant, – aber nicht wasserdicht Howson & Urbach (1989; 265)), vergl. auch Salmon (1981), Earman (1992, Kap. 4), Howson (2000, Kap. 5), Levi (1984), Jeffrey (1984).

6.3.3 Die Laplacesche Sukzessionsregel

(Rule of Succession) Diese Regel wurde von Laplace (1812) im Zusammenhang mit Fragen nach der Möglichkeit der Induktion hergeleitet und gibt seit dem Anlaß zu vielen, insbesondere auch philosophischen Diskussionen (Zabell (1989)). Laplace betrachtete insbesondere – wohl aus Gründen der Illustration – die Wahrscheinlichkeit, dass morgen wieder die Sonne aufgeht unter der Bedingung, dass sie dies in den letzten 5000 Jahren getan hat. Bei wörtlicher Auslegung der Bibel existiert die Erde seit 5000 Jahren. Man kann diese Wahrscheinlichkeit über den diskreten Fall, aber auch über den kontinuierlichen Fall herleiten.

Es geht darum, dass man eine Folge voneinander unabhängiger Ereignisse betrachtet, die jeweils mit der Wahrscheinlichkeit p eintreten. X_i, $i = 1, \ldots, n$ sei eine Indikatorvariable: $X_i = 1$ zeigt an, dass im i-ten Versuch das Ereignis eingetreten

ist, $X_i = 0$ zeigt an, dass es nicht eingetreten ist. Nachdem $X_1, \ldots, X_n$ beobachtet worden sind, wird nach der Wahrscheinlichkeit $P(X_{n+1}|X_n, \ldots, X_1)$ gefragt, mit der das Ereignis im $n+1$)-ten Versuch eintritt. p sei unbekannt, und alle Werte aus $(0, 1)$ seien gleichwahrscheinlich.

Die Gleichwahrscheinlichkeit von p in $(0, 1)$ läßt sofort an den kontinuierlichen Fall denken. Der von Laplace verfolgte Ansatz geht aber von einer diskretisierten Version aus. Fellers (1968, p. 123) Ansatz bietet eine interessante Weise, das Problem über ein Urnenmodell zu lösen. Dazu werden $N+1$ Urnen $U_0, \ldots, U_N$ angenommen. Jede Urne enthält N Kugeln. Die Urne U_k enthält k rote und $N-k$ weiße Kugeln, $k = 0, 1, \ldots, N$. Nun wird eine Urne zufällig gewählt; "zufällig" soll heißen, dass jede Urne mit gleicher Wahrscheinlichkeit gewählt wird. Da das Verhältnis roter zu weißer Kugeln in der k-ten Urne gleich $p_k = k/N$ ist, bedeutet die zufällige Wahl einer Urne die zufällige Wahl des Parameters p_k; alle p_k-Werte werden mit gleicher Wahrscheinlichkeit gewählt. Nach Wahl einer Urne U_k werden n Kugeln mit Zurücklegen gezogen. Die Wahrscheinlichkeit, dass alle aus U_k gezogenen n Kugeln rot sind, ist

$$P(k|U_k) = \left(\frac{k}{N}\right)^n.$$

A sei das zufällige Ereignis, dass überhaupt alle n gezogenen Kugeln rot sind; nach dem Satz der Totalen Wahrscheinlichkeit ist

$$P(A) = \sum_{k=0}^{N} P(k|U_k)P(U_k). \tag{175}$$

Da alle U_k mit gleicher Wahrscheinlichkeit gezogen werden, ist $P(U_k) = 1/(N+1)$, so dass

$$P(A) = \sum_{k=0}^{N} \left(\frac{k}{N}\right)^n \frac{1}{N+1} = \frac{1^n + 2^n + \cdots + N^n}{N^n(N+1)}. \tag{176}$$

Es sei B das zufällige Ereignis, dass alle $n+1$ Züge eine rote Kugel geliefert haben. Es sei $\mathbf{X} = (X_1, \ldots, X_n)$. Die bedingte Wahrscheinlichkeit, dass der $(n+1)$-te Zug eine rote Kugel liefert, gegeben $\mathbf{x}$ ist

$$P(X_{n+1} = 1|\mathbf{X}) = \frac{P(B \wedge A)}{P(A)} = \frac{P(B)}{P(A)},$$

wobei, analog zu $P(A)$,

$$P(B) = \frac{1^{n+1} + 2^{n+1} + \cdots + N^{n+1}}{N^{n+1}(N+1)},$$

Es ist

$$\sum_{k=1}^{N} k^n \approx \int_0^N x^n \, dx = \frac{N^{n+1}}{n+1}.$$

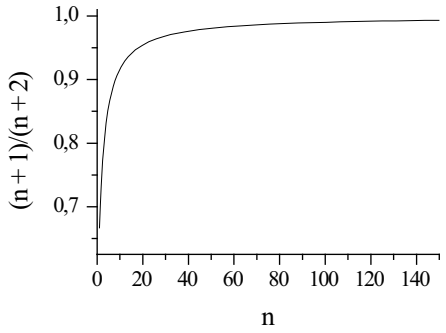
Damit hat man

$$P(X_{n+1} = 1|\mathbf{X}) = P(B|A) = \frac{P(B)}{P(A)} \approx \frac{n+1}{n+2}. \tag{177}$$

Offenbar ist

$$\lim_{n\to\infty} \frac{n+1}{n+2} = \lim_{n\to\infty} \frac{1+1/n}{1+2/n} = 1.$$

Abbildung 12: Induktion nach der Laplaceschen Folgeregel



Ist n die Anzahl der Tage, an denen bisher die Sonne aufgegangen ist, so ist die Wahrscheinlichkeit, dass sie auch morgen wieder aufgeht, demnach so gut wie 1. Der kontinuierliche Fall ist eleganter. Es ist

$$P(X_{n+1}|\mathbf{x}) = \frac{P(X_{n+1} \wedge \mathbf{X})}{P(\mathbf{X})}$$

Da Unabhängigkeit angenommen wird und die Wahrscheinlichkeit, dass $X_{n+1} = p$ ist, hat man unter der Bedingung p,

$$P(X_{n+1} \wedge \mathbf{X}|p) = pp^r(1-p)^{n-r} = p^{r+1}(1-p)^{n-r},$$

wobei der Faktor $\binom{n}{r}$ der Übersichtlichkeit wegen fortgelassen wurde, – er kürzt sich, da er in $P(\mathbf{X})$ ebenfalls auftaucht, heraus. Nach dem Satz der Totalen Wahrscheinlichkeit ist dann[16] (wegen $P(p) = 1$ für alle $p \in (0,1)$)

$$P(X_{n+1} \wedge \mathbf{X}) = \int_0^1 p^{r+1}(1-p)^{n-r}\,dp = \frac{(r+1)!(n-r)!}{(n+2)!}$$

Analog dazu erhält man für $P(\mathbf{X})$

$$P(\mathbf{X}) = \int_0^1 p^r(1-p)^{n-r}\,dp = \frac{r!(n-r)!}{(n+1)!},$$

so dass

$$P(X_{n+1}|\mathbf{X}) = \frac{r+1}{n+2}. \tag{178}$$

Für den Spezialfall $r = n$ (die Sonne ist bisher jeden Tag aufgegangen) ist dies gerade das Resultat (177).

Die 'Rule of Succession' hat eigenartige Implikationen. Es werde eine Münze geworfen, und die Zahl liegt oben (Z_1). Wir groß ist die Wahrscheinlichkeit, dass beim folgenden Wurf ebenfalls die Zahl oben liegt (Z_2)?

$$P(Z_2|Z_1) = \frac{1+1}{1+2} = \frac{2}{3}.$$

[16]Die Integrale dieser Art sind Spezialfälle von $\int_0^1 x^m(1-x)^n\,dx = m!n!/(m+n+1)!$, vergl. Gröbner, W., Hofreiter, N.: Integraltafel - Zweiter Teil, Bestimmte Integrale. Wien 1973, Seite 11, Formel 1a.

Üblicherweise nimmt man an, dass Münzwürfe voneinander unabhängig sind, und man erhält

$$P(Z_2|Z_1) = \frac{P(Z_1 \wedge Z_2)}{P(Z_1)} = \frac{\frac{1}{2} \cdot \frac{1}{2}}{\frac{1}{2}} = \frac{1}{2}.$$

Nach fünf Würfen, bei denen jedesmal die Zahl oben lag, hat man schon eine Wahrscheinlichkeit von .857, dass beim sechsten Wurf ebenfalls eine Zahl oben liegt, und nach dem zehnten Wurf beträgt die Wahrscheinlichkeit, dass beim elften Wurf wieder eine Zahl oben liegt, bereits .92. Die Unabhängigkeitsannahme sagt aber konstant eine Wahrscheinlichkeit von 1/2 voraus. Also stimmt entweder die Unabhängigkeitsannahme nicht, oder die Rule of Succession hat einen Haken. Keynes (1921) kommentierte die Folgeregel bereits so:

> No other formula in the alchemy of logic has exerted more astonishing powers. For it has established the existence of God from total ignorance, and has measured with numerical precision the probability that the sun will rise tomorrow. (p. 89)

Die Alchemie der Logik ...

6.4 A-priori-Wahrscheinlichkeiten

6.4.1 Übersicht

Der zunächst großen Plausibilität des Bayes-Ansatzes

$$P(H|E) \propto P(E|H)P(H)$$

steht die Frage, wie die a-priori-Wahrschlichkeiten $P(H)$ gewählt werden sollen gegenüber. Im Allgemeinen sind die $P(H)$ epistemische Größen. Gleichwohl existieren Versuche, die a-priori-Verteilungen als objektive Verteilungen zu definieren; dies sind die Ansätze von Jeffreys (1966/2003) und Jaynes (2003), und die Rede ist vom *objektiven Bayesschen Ansatz*, im Unterschied zum *subjektiven Ansatz*, bei dem die Wahrscheinlichkeitsbegriff von de Finetti bzw. Savage zugrundegelegt werden.

Der Übersicht halber sollen einige Klassen von a-priori-Verteilungen aufgeführt werden:

1. **Gleichverteilung:** Bayes und Laplace gingen von dem Fall aus, dass vor der Untersuchung nichts über das Ergebnis, also etwa über den Wert eines zu schätzenden Parameters bekannt sei. Man sprach vom Prinzip des Unzureichenden Grundes, Keynes (1922) führte dann dann den Ausdruck *Principle of Indifference* ein, der sich durchgesetzt hat. Die Rede ist nun vom Indifferenzprinzip. Wird ein fairer Würfel geworfen, so bedeutet diese Indifferenz, dass man i.a. keine Information darüber hat, welche Seite oben liegt, also nimmt man an, dass alle Seiten die gleiche Wahrscheinlichkeit haben, oben zu liegen. Bei einer fairen Münze gilt das gleiche. Generell kann man mit dem Indifferenzprinzip die Gleichverteilung assoziieren. So plausibel allerdings im Falle völliger Indifferenz die Annahme einer Gleichverteilung ist, so schnell geht man inParadoxien verloren, wenn es sich nicht gerade um Würfel handelt, oder um den Parameter $\theta = p$ einer Binomialverteilung, für den die

Gleichverteilung auf $[0, 1]$ als a-priori-Verteilung angenommen werden kann. In Abschnitt 6.4.2 wird deshalb ausführlich auf das Indifferenzprinzip eingegangen.

Die a-priori-Verteilung sei eine Gleichverteilung auf dem Intervall $[a, b]$, so dass $f(\theta) = 1/(b - a) = k$ eine Konstante auf $[a, b]$. Für die a-posteriori-Verteilung erhält man dann

$$P(\theta|\mathbf{x}) = \frac{P(\mathbf{x}|\theta)k}{P(\mathbf{x})} \tag{179}$$

und

$$P(\mathbf{x}) = k \int_a^b P(\mathbf{x}|\theta)\, d\theta.$$

Damit gilt

$$P(\theta|\mathbf{x}) \propto P(\mathbf{x}|\theta), \tag{180}$$

d.h. die Evidenz in den Daten $\mathbf{x}$ bezüglich θ ist bereits in der Likelihood enthalten.

2. **Nichtinformative Verteilungen:** Da es sich zeigt, dass das Problem, Indifferenz auszudrücken, nicht stets über die Annahme der Gleichverteilung gelöst werden kann, versucht man, andere Verteilungen zu finden, die möglichst wenig Wissen über die Parameterwerte ausdrücken. Verteilungen, die hierfür in Frage kommen, heißen auch *nichtinformative Verteilungen* (Jeffreys (1961/2003)). Im Falle der Binomialverteilung ist die a-priori-Verteilung

$$f(\theta) = \frac{d\theta}{\theta(1 - \theta)} \tag{181}$$

vorgeschlagen worden; diese Verteilung ist als *Haldane-Verteilung* bekannt. Jeffreys (1961/2003) ist aufgrund von Überlegungen, bestimmte, mit der Gleichverteilung verbundene Fragen zu umgehen, auf die Verteilung

$$f(\theta) = \frac{d\theta}{\sqrt{\theta(1 - \theta)}} \tag{182}$$

gekommen. Für die unbekannte Varianz $\sigma^2 > 0$ kommt er aufgrund der gleichen Überlegungen auf die Verteilung

$$f(\sigma^2) = \frac{d\sigma}{\sigma}. \tag{183}$$

Diese Verteilung sowie die Verteilung (182) sind Spezialfälle der *Jeffreys' Prior*; sie wird weiter unten näher erläutert.

3. **Uneigentliche a-priori-Verteilungen:** Für die Dichte f einer zufälligen Veränderlichen X gilt

$$f(x) = \int_{-\infty}^{\infty} f(x)\, dx = 1.$$

Für die Dichte (183) etwa ist diese Bedingung nicht erfüllt. Hat man keine Information über den Erwartungwert einer zufälligen Veränderlichen, so wird gelegentlich

$$f(\mu) = d\mu, \quad -\infty < \mu < \infty$$

angenommen; offenbar ist das Integral von f über $(-\infty, \infty)$ nicht gleich 1. Die Rede ist dann von *Uneigentlichen a-priori-Verteilungen.*

4. **Konjugierte Verteilungen:** Gelegentlich hat man Grund zu der Annahme, dass die a-posteriori-Verteilung zu einer bestimmten Klasse gehören sollte, dass sie etwa eine Normalverteilung sein sollte, oder es ist aus mathematischen Gründen bequem, eine bestimmte a-posteriori-Verteilung zu haben. Dies kann erreicht werden durch Wahl einer entsprechenden a-priori-Verteilung. Das Paar von a-priori- und a-posteriori-Verteilungen heißt dann *konjugiert*, und insbesondere die a-priori-Verteilung heißt *konjugiert bezüglich der Likelihood-Funktion*. So ist die Gauß-Verteilung konjugiert bezüglich einer Gaußschen Likelihood-Funktion. Ist die Likelihood-Funktion eine Gauß-Dichte, so ergibt sich für die a-posteriori-Verteilung eine Gauß-Dichte, wenn die a-priori-Verteilung eine Gauß-Dichte ist. Ist X binomialverteilt, so ist die Beta-Verteilung konjugiert zur Likelihood-Funktion und die a-posterior-Verteilung ist ebenfalls eine Beta-Verteilung.

Spezielle Probleme treten auf, wenn sich Hypothesen auf einen Punkt des Kontinuums beziehen: H_0: $\mu = \mu_0$ etwa. Hier Abschnitt über Berger & Delampadys Arbeit ankündigen.

6.4.2 Das Indifferenzprinzip

Dieses Prinzip[17] ist von Bedeutung im Zusammenhang mit der Frage, wie die a-priori-Wahrscheinlichkeiten $P(H)$ bzw. $P(H_i)$ definiert werden sollen. Hat man keinerlei Information über die Korrektheit der verschiedenen Hypothesen, so hatte schon Bayes, und nach ihm Laplace (1820), argumentiert, dass allen Hypothesen die gleiche Wahrscheinlichkeit zugeordnet werden müsse, woraus

$$P(H_i) = \frac{1}{n}, \quad i = 1, 2, \ldots, n \tag{184}$$

folgt. Im kontinuierlichen Fall kann man u.U. Gleichverteilung annehmen, – wenn man annehmen kann, dass der abzuschätzende Parameter in einem endlichen Intervall $[a, b]$ liegt. Hier ergibt sich sofort eine Schwierigkeit, wenn a oder b oder sowohl a und b keine endlichen Größen sind. In jedem Fall wird man versuchen, eine *nichtinformative* a-priori-Verteilung zu finden, also eine Verteilung, die keine Präferenz für bestimmte Werte des zu schätzenden Parameters ausdrückt.

Earman (1992, p. 17) verweist auf Wittgenstein, der in seinem *Tractatus logico-philosophicus* das Indifferenzprinzip auf seine Weise einführt:

> 5.15 Ist W_r die Anzahl der Wahrheitsgründe des Satzes r, W_{rs} die Anzahl derjenigen Wahrheitsgründe des Satzes s, die zugleich Wahrheitsgründe von r sind, dann nennen wir das Verhältnis: $W_{rs} : W_r$ das Maß der *Wahrscheinlichkeit*, welche der Satz r dem Satz s gibt.
>
> ⋮
>
> 5.1511 Es gibt keinen besonderen Gegenstand, der den Wahrscheinlichkeitssätzen eigen wäre.
>
> 5.152 Sätze, welche keine Wahrheitsargumente miteinander gemein haben, nennen wir voneinander unabhängig.
>
> Zwei Elementarsätze geben einander die Wahrscheinlichkeit 1/2.

[17] Jakob Bernoulli und später von Kries (1886) sprachen vom *Prinzip des Unzureichenden Grundes*; Keynes (1921) führte den Ausdruck *Indifferenzprinzip* ein.

> Folgt p aus q, so gibt der Satz q dem Satz p die Wahrscheinlichkeit 1. Die Gewissheit des logischen Schlusses ist ein Grenzfall der Wahrscheinlichkeit. (Anwendung auf Tautologie und Kontradiktion.)
>
> 5.153 Ein Satz ist an sich weder wahrscheinlich noch unwahrscheinlich. Ein Ereignis trifft ein, oder es trifft nicht ein, ein Mittelding gibt es nicht.
>
> ⋮

Das Interessante an der Textstelle ist, dass Wittgenstein Wahrscheinlichkeiten auf Aussagen, nicht auf Ereignisse bezieht. Earman weist darauf hin, dass Wittgenstein eine dogmatische a-priori-Verteilung annimmt, die das Lernen aus der Erfahrung verhindere: jemand, der nach Wittgensteins Regel vorgehe, handele in einem Versuch (trial) unabhängig von dem, was er in den vorangegangenen Versuchen gelernt habe. Ob man aber so vorgehen müsse, hinge von der Einstellung zur Frage nach der Induktion ab: Anti-Induktivisten von Hume bis Popper würden Wittgensteins Argument als Ausdruck ihrer Doktrin sehen, dass die Erfahrungen der Vergangenheit keine Vorhersagen für die Zukunft rechtfertigen.

Problematische Aspekte Das Indifferenzprinzip ist von trügerischer Einfachheit; schon der oben angedeutete Fall nicht-endlicher Werte für das Intervall $[a, b]$ bereitet bereits Schwierigkeiten. Keynes (1921/2008) verwendet ein 20-seitiges Kapitel auf die Frage, wie das Prinzip zu begründen sei. Fisher (1922, p. 232) beschreibt seine Maximum-Likelihood Methode und kontrastiert sie mit dem Bayesschen Ansatz anhand der Schätzung der Parameters p der Binomialverteilung

$$P(X = x|n, p) = \binom{n}{x} p^x (1-p)^{n-x}.$$

Die Maximum-Likelihood Schätzung liefert die Schätzung $\hat{p} = x/n$, und die Frage ist, wie gut diese Schätzung ist. Dies sei die Frage nach der Häufigkeitsverteilung der Schätzungen, wenn der "wahre" Wert gleich p sei, jeweils n Beobachtungen gemacht würden und $X = x$ sei. Nach Bayes sei die Wahrscheinlichkeit, dass p in einem Bereich dp liege, für alle $dp \in [0, 1]$ gleich, und die a-posterior-Wahrscheinlichkeit ist dann

$$p^x (1-p)^{n-x} dp$$

Fisher argumentiert nun, dass diese Betrachtung zwar sehr nützlich wäre, wäre sie denn eindeutig, aber tatsächlich sei sie "extremely arbitrary":

> "For we might never have happened to direct our attention to the particular quantity p: we might equally have measured probability upon an entirely different scale. If, for instance,
>
> $$\sin\theta = 2p - 1,$$
>
> the quantity, θ, measures the degree of probability, just as well as p, and is even, for some purposes, the more suitable variable. The chance of obtaining a sample of x successes and y failures is now
>
> $$\frac{n!}{2^n x!(n-x)!}(1+\sin\theta)^x(1-\sin\theta)^{n-x}.$$
>
> Die Anwendung der Maximum-Likelihood-Methode liefert
>
> $$\sin\theta = \frac{2x-n}{2n},$$

"an exactly equivalent solution to that obtained using the variable p. But what a priori assumption are we to make as to the distribution of θ? Are we to assume that θ is equally likely to lie in all equal ranges $d\theta$? In this case the a priori probability will be $d\theta/\pi$, and that after making the observations will be proportional to

$$(1+\sin\theta)^x(1-\sin\theta)^{n-x}.$$

But if we interpret this in terms of p, we obtain

$$p^x(1-p)^{n-x}\frac{dp}{\sqrt{1-p}} = p^{x-1/2}(1-p)^{n-x-1/2}dp,$$

a result inconsistent with that obtained previously. In fact, the distribution previously assumed for p was equivalent to assuming the special distribution for θ,

$$df = \frac{\cos\theta}{2}d\theta,$$

the arbitrariness of which is fully apparent when we use any variable other than p.

Das Fishersche Beispiel mag auf den ersten Blick nicht sehr plausibel erscheinen: wenn man den Parameter einer Binomialverteilung schätzen will, wird man bei vollständiger Unwissenheit über seinen Wert eben die Gleichverteilung *dieses* Parameters und nicht die eines anderen als a-priori-Verteilung annehmen. Andererseits ist es richtig, dass θ eine deterministische Funktion von p ist, und wenn man nichts über p weiß, dann weiß man auch nichts über θ, und warum soll man nun für θ *keine* Gleichverteilung annehmen? Tatsächlich ist die hier angesprochene Frage der Transformation eines Parameters nur ein Anfang; wenn die Verteilung von Verhältnissen betrachtet wird, stellt sich die Frage erneut und wohl auch auf noch unangenehmere Weise. Bevor darauf näher eingegangen wird, soll noch eine Verteilung betrachtet werden, bei der der Parameter auf $[0,\infty)$ verteilt ist.

Häufigkeiten können Poisson-verteilt sein, so dass

$$P(X=k|\lambda) = e^{-\lambda t}\frac{(\lambda t)^k}{k!}, \quad \lambda>0, \;\; k=0,1,2,\ldots \tag{185}$$

gilt. Dann macht die Gleichverteilung für λ auf $(0,\infty)$ zunächst einmal keinen Sinn, denn

$$f(\lambda) = \frac{1}{\infty - 0} = 0, \quad \text{für alle}\,\lambda,$$

so dass

$$P(\lambda|X) = \frac{P(X|\lambda)f(\lambda)}{P(X)} = 0 \text{ für alle } \lambda.$$

Ein analoges Problem stellt sich, wenn die Daten mit unbekanntem σ $N(\mu,\sigma^2)$-verteilt sind; eine direkte Anwendung des Indifferenzprinzips legt eine Gleichverteilung für σ über $(0,\infty)$ nahe. Bayes und Laplace haben, wie es scheint, nur an endliche Parameterräume gedacht. Jeffreys führte hierfür die bereits genannten uneigentlichen a-priori-Verteilungen (improper priors) ein, deren Integral nicht gleich 1 ist; vielmehr kann

$$\int f(\theta)d\theta = \infty$$

gelten. Da die a-posterior-Verteilung eine Dichte ist, muß

$$\int_\Theta P(\theta|\mathbf{x}) = 1$$

gelten. Ist $f(\theta)$ eine uneigentliche Verteilung, so wird aber zunächst

$$\int_\Theta P(\theta|\mathbf{x}) = c \int_\Theta P(\mathbf{x}|\theta) f(\theta)\, d\theta \neq 1$$

gelten, c eine Konstante. Sofern das Integral auf der rechten Seite endlich ist, etwa den Wert K annimmt, kann man $P(\theta|\mathbf{x})$ durch $P(\theta|\mathbf{x})/K$ ersetzen und erhält damit eine Dichte, deren Integral nun gleich 1 ist.

Ein Problem ganz anderer Art ergibt sich, wenn sich die zu schätzenden Parameter auf Verhältnisse beziehen. Ein bekanntes Paradoxon ist das Folgende.

Das Wein/Wasser-Paradoxon Dieses Paradoxon geht auf R. von Mises zurück (v. Mises (1981), p. 77). Es entsteht, wenn das Indifferenzprinzip auf Verhältnisse angewendet wird. So sei ein Krug mit einer Mischung von Wasser und Wein gegeben. Das genaue Verhältnis von Wein und Wasser ist nicht bekannt, aber man weiß, dass der Anteil einer der beiden Substanzen höchstens dreimal so groß wie der der anderen Substanz ist. Ist also X das Verhältnis von Wein zu Wasser, so muß $X \geq 1/3$ sein, andernfalls wäre der Anteil von Wasser mehr als dreimal so groß wie der des Weins. Ebenso muß $X \leq 3$ gelten, sonst wäre der Anteil des Weins mehr als dreimal so groß wie der des Wassers. Also muß gelten

$$\frac{1}{3} \leq X \leq 3, \quad \frac{1}{3} \leq Y \leq 3. \tag{186}$$

mit $Y = 1/X$, und der rechte Ausdruck ergibt sich durch eine analoge Argumentation. Weiß man nichts über das tatsächliche Verhältnis von Wein und Wasser, außer den Bedingungen (186), so führt das Prinzip der Indifferenz auf eine Gleichverteilung für X auf $[1/3, 3]$. Aber dann ist Y nach (300) nicht gleichverteilt. Andererseits kann man ebenso gut annehmen, Y sei auf $[1/3, 3]$ gleichverteilt. Aber dann kann X nicht mehr gleichverteilt sein. In der üblichen Formulierung des Paradoxones wird gezeigt, dass die Annahme der Gleichverteilung sowohl für X als auch für Y auf widersprüchliche Ergebnisse führt, was nach den vorangegangenen Überlegungen nicht verwunderlich ist: so werde etwa nach der Wahrscheinlichkeit $P(X \leq 2)$ gefragt. Es ist

$$P(X \leq 2) = P(1/Y \leq 2) = P(1/2 \leq Y). \tag{187}$$

Nimmt man nun sowohl für X als auch für Y eine Gleichverteilung an, so erhält man einerseits

$$P(X \leq 2) = \frac{2 - 1/3}{3 - 1/3} = \frac{5}{8},$$

und andererseits, wegen (299), wenn man X durch Y ersetzt,

$$P(Y \geq 1/2) = \frac{3 - 1/2}{3 - 1/3} = \frac{15}{16},$$

also $P(X \leq 2) \neq P(Y \geq 1/2)$, in Widerspruch zu (187). Eine mögliche Auflösung des Paradoxons wird im Anhang gegeben.

Der Widerspruch zwischen $P(X \leq 2) = P(Y \geq 1/2)$ einerseits und $P(X \leq 2) \neq P(Y \geq 1/2)$ andererseits wird im Allgemeinen dem Indifferenzprinzip angelastet. Keynes versuchte, den Widerspruch zu überwinden, indem er forderte, es dürfe nur endlich viele, nicht weiter teilbare Alternativen geben; es läßt sich

aber zeigen, dass dieses Postulat nicht aufrechtzuerhalten ist. Van Fraassen (1989) hält das Wein-Wasser-Paradoxon für "the ultimate defeat" des Indifferenzprinzips, Gillies (2000a) spricht von einem "tödlichen" Argument gegen dieses Prinzip, und Oakes (1986) folgert aus dem Paradoxon, dass dieses die klassische Konzeption der Wahrscheinlichkeit überhaupt ins Wanken bringe. □

Geschwindigkeiten Der Ausweg, der sich biem Wein-Wasser-Paradoxon finden läßt, scheint für andere Verhältnisse, etwa dem von Weg zu Zeit bzw. Zeit zu Weg nicht zu existieren. Man möchte etwa die Geschwindigkeit, mit der bestimmte Bewegungen durchgeführt werden, bestimmen. Man hat zwei Möglichkeiten: (i) man mißt die Zeit, die benötigt wird, um eine Bewegung einer bestimmten Länge durchzuführen, oder (ii) man mißt die Strecke, die in einer vorgegebenen Zeit zurückgelegt wird. Man könnte als a-priori-Verteilung etwa eine Gleichverteilung für die Zeit annehmen, – aber ebenso gut auch für die Strecke. Nimmt man die Gleichverteilung für die Zeit an, so kann die Verteilung für die Strecken nicht gleichverteilt sein, und umgekehrt. Die zufälligen Veränderlichen X für die Zeit und Y für die Strecke sind reziprok zueinander.

Das Problem ist allerdings nicht spezifisch für die Gleichverteilung, die Problematik bleibt für jede andere a-priori-Verteilung erhalten. Auf der einen Seite ist es beliebig, ob man die Zeit oder die Strecke mißt, auf der anderen Seite kann man nicht die gleiche a-priori-Verteilung für X und für $Y = 1/X$ annehmen. Die Frage ist ähnlich der nach der Gleichverteilung des Verhältnisses von Wasser und Wein. Auch dort findet sich eine Art ad-hoc-Ausweg, allerdings wird auch dort die Frage nach der Beliebigkeit des Parameters, den man als gleichverteilt annimmt, eher umgangen als eindeutig beantwortet. Dies wirft die Frage auf, ob sich nicht a-priori-Verteilungen finden lassen, die gegenüber Transformationen der Art $X \to Y$ invariant sind. Damit gibt man allerdings die Gleichsetzung von Indifferenz und Gleichverteilung auf. □

Das Paradoxon von Bertrand:[18] Man betrachte einen Kreis und das in ihm eingeschriebene gleichseitige Dreieck. Dann wird zufällig eine Sehne des Kreises ausgewählt. Gesucht ist die Wahrscheinlichkeit, dass diese Sehne länger als die Seite des Dreiecks ist.

Bertrand , 3 verschiedene Methoden, die Sehne auszuwählen, und jede Methode führt auf der Basis des Indifferenzprinzips zu einer anderen Wahrscheinlichkeit. Székely (1990) weist darauf hin, dass noch weitere Möglichkeiten existieren, mit jeweils charakteristischen Wahrscheinlichkeiten. Nach Székely ist die Forderung, dass das Indifferenzprinzip auf nur eine Wahrscheinlichkeit führen müsse, eine Fehlinterpretation des Prinzips: Wenn die Methode nach einem bestimmten Gesichtspunkt gewählt wird, so liefere das Indifferenzprinzip die zur Methode korrespondierende Wahrscheinlichkeit. □

Die Farbe von Büchern: Man bekommt ein Buch geschenkt. Es ist noch eingepackt und man soll die Farbe raten; alles, was man weiß, sei, dass die Farbe eine Primärfarbe ist. Dann ist es entweder grün oder nicht grün, und nach dem Indifferenzprinzip ist die Wahrscheinlichkeit, dass es grün ist, gleich 1/2. Andererseits ist es entweder rot oder nicht rot, und drückt man die Indifferenz in Bezug auf diese Alterntiven durch eine Gleichverteilung aus, so ist die Wahr-

[18] Joseph Louis François Bertrand (1822 – 1900), Mathematiker. Das Paradoxon wurde von ihm in seinem Buch Calcul de Probabilités (1888) veröffentlicht.

scheinlichkeit, dass das Buch rot ist, gleich 1/2. Desgleichen kann es blau oder nicht blau sein, und die Wahrscheinlichkeit für diese beiden Alternativen ist wieder 1/2. Nun ist es entweder rot, oder blau, oder grün, und gemäßt der Regel $P(A_1 \vee A_2 \vee A_3) = P(A_1) + P(A_2) + P(A_3)$ für sich ausschließende Ereignisse erhält man eine Gesamtwahrscheinlichkeit von 3/2, obwohl diese nur maximal gleich 1 sein darf.

Dieses 'Paradoxon' geht auf Keynes (1973) zurück und ist, als Argument gegen das Indifferenzprinzip, wenig überzeugend. Denn wenn man weiß, dass das Buch eine der *drei* Primärfarben hat, bekommt jede Farbe nach dem Indifferenzprinzip die Wahrscheinlichkeit 1/3 zugeordnet und von einem Paradoxon kann keine Rede mehr sein. Die Alternativen grün versus nicht grün, rot versus nicht rot und blau versus nicht blau haben jeweils ihren eigenen *sample space*, und diese Stichprobenräume kann man am Ende nicht einfach zu einem einzigen zusammenwerfen. Gleichwohl, Howson (2000), p. 82, gibt eine längere Diskussion des Paradoxons, die er in eine Problematik eines Kontinuums von Alternativen münden läßt, wie sie beim Wein/Wasser- oder beim GeschwindigkeitsParadoxon vorliegt.

Es ist also kein Wunder, dass das Indifferenzprinzip Gegenstand länglicher Debatten wurde. Jaynes (2003, p. 343) konstatiert, dass die "Orthodoxie" – also v. Mises, Fisher, Neyman & Pearson und die Anhänger dieser Richtungsweiser – dem Problem einfach nur auswichen.

Parametertransformationen und Indifferenz Wie schon oben beim Fisher-Zitat in Bezug auf den Binomialparameter angemerkt, können Transformationen der Parameter betrachtet werden. Wird der ursprüngliche Parameter als gleichverteilt angenommen, so ist der transformierte Parameter dann nicht mehr gleichverteilt, wenn die Transformation nichtlinear ist. Die Betrachtung von Transformationen ist keineswegs nur eine akademische Spielerei. So können verschiedene Einheiten für t in (191) verschiedene λ-Werte bedeuten, die aber durch eine Transformation ineinander überführbar sind; dieser Sachverhalt muß bei der Definition der der a-priori-Verteilung berücksichtigt werden. Dies führt zum Begriff der *invarianten a-priori-Verteilung.* Solche a-priori-Verteilungen werden in Abschnitt **??** behandelt.

Zusammenfassung Hacking (1965) bezeichnet des Indifferenzprinzip abwechselnd als 'notorious' und 'noxious' (p. 147), und schließlich als 'tedious' (p. 201). In der Tat führt das Prinzip nicht nur zu einer Reihe vonParadoxien (Bertrand, Farben, etc), sondern auch zu einer Reihe von "ernsthaften" Schwierigkeiten, etwa im Zusammenhang mit Signifikanztests bezüglich der Schätzungen von Parametern (Jeffreys (1961/2003), Jaynes (2003)), auf die hier nicht eingegangen werden kann. Gilt das Indifferenzprinzip, so folgt für $n \to \infty$, , dass jede Hypothese die Wahrscheinlichkeit $\lim_{n\to\infty} 1/n = 0$ hat. Für Popper (2002, Neuer Anhang VII) folgt daraus, dass die Wahrscheinlichkeit eines jeden, nicht-tautologischen Gesetzes gleich Null ist; wegen

$$P(H|E) \propto P(E|H)P(H) = 0 \text{ für } P(H) = 0$$

sei jede induktive Logik unmöglich.

6.4.3 Jeffreys Priors

Der Bayes-Laplace-Ansatz fürht auf Probleme mit der Annahme der Gleichverteilung:

1. Was passiert, wenn die Verteilung des parameters auf $[0, \infty)$ definiert ist?
2. Es sei v gleichverteilt, – wie ist die Größe v^p verteilt?
3. Wie entscheidet man sich für verschiedene Klassenbildungen?

Fisher Information: Jeffreys zeigt, dass diese Größe invariant gegenüber Reparametrisierungen ist. Ist α ein Parameter und α' eine differenzierbare Transformation von α:

$$I(\alpha') = \frac{d\alpha}{d\alpha'} I(\alpha) \frac{d\alpha^T}{d\alpha'}$$

dann definiert Jeffreys

$$f(\alpha) = \sqrt{|I(\alpha)|}. \tag{188}$$

Die einzige Motivation für die Definition dieser Klasse von a-priori-Verteilungen ist eben die Invarianz gegenüber von Transformationen.

Binomialverteilung: Hier ergibt sich die a-priori-Verteilung

$$f(p) = \frac{1}{\pi} \frac{1}{\sqrt{p(1-p)}}, \tag{189}$$

d.h. die Arcus-Sinus-Transformation, im Unterschied zur *Haldane-Verteilung.*

$$f_H(p) = \frac{1}{p(1-p)}. \tag{190}$$

Poisson-Verteilung: Hier ergibt sich die a-priori-Verteilung

$$f(\lambda) \propto \frac{1}{\sqrt{\lambda}}. \tag{191}$$

Korrelationskoeffizient Die zufälligen Veränderlichen X, Y seien gemeinsam normalverteilt; gesucht ist die a-priori-Verteilung für den Korrelationskoeffizienten ρ. Jeffreys findet

$$f(\rho, \tau, \sigma) \propto \frac{1}{\tau\sigma(1-\rho^2)^{3/2}}; \tag{192}$$

für τ und σ fix ergibt sich die bedingte a-priori-Verteilung

$$f(\rho) = \frac{1}{\pi} \frac{1}{\sqrt{1-\rho^2}} \tag{193}$$

Es ist eine offene Frage (Robert et al. (2009)) ob all diese Verteilungen wirklich besser als die Gleichverteilung des entsprechenden Parameters sind.

Exponentialfamilien: Eine Verteilung gehöre zur Exponentialfamilie mit dem Parameter β,

$$f(x|\beta) = \psi(x)\phi(\beta)\exp(\beta v(x)). \tag{194}$$

Die Fisher Information ist dann

$$I(\beta) = \frac{\partial^2 \log \phi(\beta)}{\partial \beta^2} \tag{195}$$

(Weitere Betrachtungen findet man in Robert et al 2009, s. a. Kass & Wasserman (1996)).

6.4.4 Jaynes' Maximale Entropie und Transformationsgruppen

Diskrete Verteilungen Jaynes (1968) betrachtet a-priori-Verteilungen vom Standpunkt der Informationstheorie aus. Für eine Verteilung über n Zuständen mit den Wahrscheinlichkeiten p_i läßt sich dann die Entropie

$$H(p_1, \ldots, p_n) = -\sum_{i=1}^{n} p_i \log p_i \tag{196}$$

definieren; diese soll dann unter eventuell gegebenen Nebenbedingungen – außer der, dass $\sum_i p_i = 1$ – maximiert werden.

Es werde angenommen, dass $\mathbf{x} = (x_1, \ldots, x_n)$ gegeben sei, wobei n endlich oder zumindest abzählbar unendlich sei. Die x_i dürfen beliebig sein. Die verfügbare Information I möge eine Reihe von Randbedingungen für die Wahrscheinlichkeitsverteilung $p(x_i|I)$ spezifizieren. Eine mögliche Spezifikation ist, dass diese Randbedingungen die Mittelwerte der Funktionen $\{f_1(\mathbf{x}), f_2(\mathbf{x}), \ldots, f_m(\mathbf{x})\}$ festlegen, $m < n$. Dann soll

$$H = -\sum_{i=1}^{n} p(x_i|I) \log p(x_i|I)$$

unter den Bedingungen

$$\sum_i p(x_i|I) = 1, \quad \sum_i p(x_i|I) f_k(x_i) = F_k, \;\; k = 1, \ldots, m \tag{197}$$

maximiert werden. Die F_k sind festgelegte Mittelwerte. Eine allgemeine Lösung ist (Jaynes (1968, p. 7), (2003, p. 355))

$$p(x_i|I) = \frac{1}{Z(\lambda_1, \ldots, \lambda_m)} \exp(\lambda_1 f_1(x_i) + \cdots + \lambda_m f_m(x_i)). \tag{198}$$

$Z(\lambda_1, \ldots, \lambda_m)$ ist eine *partition function*

$$Z(\lambda_1, \ldots, \lambda_m) = \sum_i \exp(\lambda_1 f_1(x_i) + \cdots + \lambda_m f_m(x_i)), \quad \lambda_k \in \mathbb{R},$$

wobei die λ_k gemäß den Randbedingungen (197) bestimmt werden, die sich in der Form

$$F_k = \frac{\partial}{\partial \lambda_k} \log Z(\lambda_1, \ldots, \lambda_m) \tag{199}$$

darstellen lassen. Wie Jaynes es formuliert: die Verteilung (198) ist diejenige, die unter den gegebenen Randbedingungen so weit es irgend geht "ausgespreizt" ist. Damit erfüllt sie eine Bedingung für a-priori-Verteilungen: maximale Ungewißheit

unter gegebenen Randbedingungen zu repräsentieren und damit maximale Freiheit für die Entscheidungen über die Parameter zu lassen.

x werde nun durch ein Zufallsexperiment bestimmt: Das Experiment werde M-mal wiederholt. Die Frage ist, was über die x_i ausgesagt werden kann, wenn $M \to \infty$; dies ist, was in der frequentistischen Interpretation der Wahrscheinlichkeit angenommen wird. Es gibt n^M mögliche Resultate. Die M Wiederholungen liefern m_1 x_1-Werte, m_2 x_2-Werte, etc, mit $\sum_i m_i = M$. Die festgelegten Mittelwerte mögen gefunden worden sein, so dass es die Bedingungen

$$\sum_{i=1}^{n} m_i f_k(x_i) = MF_k, \quad k = 1, 2, \ldots, m \tag{200}$$

gibt. Die Frage ist, wie viele der n^M Resultate nun mit den gefundenen Zahlen $\{m_1, m_2, \ldots, m_n\}$ kompatibel sind. Die Anzahl ist durch den Multinomialkoeffizienten

$$W = \frac{M!}{m_1! m_2! \cdots m_n!} = \frac{M!}{(Mf_1)! \cdots (MF_m)!} \tag{201}$$

gegeben. Damit ist die Menge der Häufigkeiten $\{f_1, f_2, \ldots, _n\}$, die in der größtmöglichen Variation erzeugt werden kann, diejenige, die (201) maximiert relativ zu den Bedingungen (200). Gleichwertig dazu kann man jede monoton wachsende Funktion von W maximieren, etwa $M^{-1} \log W$, und über die Stirling-Approximation erhält man für $M \to \infty$

$$M^{-1} \log W \to -\sum_{i=1}^{n} f_i \log f_i = H_f. \tag{202}$$

Hat man also *testbare* Information, so ist die Wahrscheinlichkeitsverteilung, die die Entropie maximiert, identisch mit der Häufigkeitsverteilung, die in der größmöglichen Anzahl realisiert werden kann.

Beispiel 6.5 Herleitung der Binomialverteilung (Jaynes (1968, p. 11)) Es wird ein Experiment durchgeführt, das dem Experimentator eine "Mitteilung" liefert: das Alphabet bestehe aus dem möglichen Resultat eines Versuchsdurchganges, und in jedem Versuchsdurchgang wird ein "Buchstabe" der Mitteilung geliefert. Insbesondere handele es sich um Bernoulli-Versuche, mit den zufälligen Veränderlichen

$$y_i = \begin{cases} 1, & \text{"Erfolg"} \\ 0, & \text{"kein Erfolg"} \end{cases}$$

Nach n Wiederholungen bzw. Versuchsdurchgängen hat man die Mitteilung

$$M \equiv \{y_1, y_2, \ldots, y_n\}.$$

Die Gesamtzahl der "Erfolge" ist dann

$$r(M) = \sum_i y_i.$$

Es werde nun $\mathbb{E}(r) = np$ angenommen. Gesucht ist nun Wahrscheinlichkeit, r Erfolge in n Versuchen zu erhalten. Dazu wird das Maximum-Entropie-Prinzip angewendet. Gesucht ist die Wahrscheinlichkeit

$$P_M \equiv p\{y_0 y_1 \cdots y_n\},$$

auf dem 2^n-Stichprobenraum aller möglichen Mitteilungen; gesucht ist also die Verteilung P_M, die die Entropie

$$H = -\sum_M P_M \log P_M$$

maximiert unter der Nebenbedingung $\mathbb{E}(r) = np$. (198) liefert

$$P_M = \frac{1}{Z(\lambda)} \exp(\lambda r(M)), \tag{203}$$

mit

$$Z(\lambda) = \sum_M \exp(\lambda r(M)) = (e^\lambda + 1)^n.$$

Die Lösung ergibt sich durch Anwendung von (199):

$$\mathbb{E}(r) = \frac{\partial}{\partial \lambda} \frac{n}{\exp(-\lambda) + 1},$$

woraus sich

$$\lambda = \log \frac{\mathbb{E}(r)}{n - \mathbb{E}(r)} = \log \frac{p}{1-p}$$

ergibt. Hieraus und aus (203) folgt dann

$$P_M = p^r (1-p)^{n-r}. \tag{204}$$

P_M ist die Wahrscheinlichkeit, eine bestimmte Mitteilung zu erhalten, mit "Erfolgen" in bestimmten Versuchen. Kommt es auf die Positionen der Erfolge nicht an, muß noch mit $\binom{n}{r}$ multpliziert werden, und man erhält die Binomialverteilung

$$p(r|n) = \binom{n}{r} p^r (1-p)^{n-r}. \tag{205}$$

□

Stetige Verteilungen Die Verallgemeinerung auf den Fall stetiger Verteilungen ist schwierig, weil die Verallgemeinerung der Definition der Entropie nicht einfach in der Form

$$H = -\int p(x) \log p(x) dx$$

angeschrieben werden kann; dieser Ausdruck ist nicht invariant unter Variablentransformationen $x \to y(x)$. Die gleiche Aussage gilt für das Bayessche Theorem mit der Konsequenz, dass man nicht sagen kann, welche Parametrisierung in Bezug auf das Indifferenzprinzip gewählt werden muß[19]. Eine ausführliche Darstellung findet man in Jaynes (2003, Kap. 11).

Jaynes schlägt vor, das Problem über geeignete Transformationsgruppen zu lösen, und illustriert den Gedanken zunächst an einem Beispiel.

Es wird eine Stichprobe $\mathbf{x}$ gebildet, wobei die Population durch eine 2-Parameterverteilung

$$p(dx|\mu, \sigma) = h\left(\frac{x-\mu}{\sigma}\right) \frac{dx}{\sigma} \tag{206}$$

[19] Wie Jaynes (1968, p. 16) anmerkt, ist das hier entstehende Problem nicht typisch für die Bayes-Statistik; es existiert auch für erwartungstreue und effiziente Schätzer, kleinste Konfidenzinervalle, etc, der "orthodoxen" Statistik.

definiert sei. Gegeben $\mathbf{x} = (x_1, \ldots, x_n)$ ist die Aufgabe, μ und σ zu schätzen. So lange keine a-priori-Verteilung $f(\mu, \sigma)d\mu d\sigma$ erklärt ist, ist das Schätzproblem nicht definiert. Wenn man nun nach dem Indifferenzprinzip vollständige Unwissenheit postuliert, ist nicht klar, welche Funktion f gewählt werden muß.

Andererseits ist μ ein Lokationsparameter und σ ein Skalenparameter, und die Funktion h in (206) ist bekannt. Nimmt man nun komplettes Unwissen über μ und σ an, so heißt das, dass ein Wechsel des Skalenparameters und des Lokationsparameters den Zustand vollständigen Unwissens nicht ändert. Man betrachte die Transformationen

$$\mu' = \mu + b, \quad \sigma' = a\sigma, \quad x' - u' = a(x - \mu), \tag{207}$$

mit $0 < a < \infty$, $(-\infty < b < \infty)$. Die Verteilung (206) geht über in

$$p(dx'|\mu', \sigma') = h\left(\frac{x' - \mu'}{\sigma'}\right)\frac{dx'}{\sigma'}, \tag{208}$$

d.h. d.h. h bleibt unverändert. Die a-priori-Verteilung geht aber über in

$$g(\mu', \sigma') = \frac{1}{a}f(\mu, \sigma). \tag{209}$$

Man habe nun eine zweite Stichprobe $\mathbf{x}' = (x'_1, \ldots, x'_n)$ und soll μ' und σ' schätzen. Man sei wieder vollständig unwissend bezüglich der Werte dieser Parameter. Die Fragestellung ist vollständig symmetrisch zu der gerade behandelten, so dass aus Konsistenzgründen die a-priori-Vereilungen identisch sein müssen, d.h. es muß

$$f(\mu, \sigma) = g(\mu, \sigma) \tag{210}$$

gelten, unabhängig von den Werten von a und b in (207). Allerdings ist nun die Form von f bzw. g festgelegt, denn wegen (209) muß nun die Funktionalgleichung

$$f(\mu, \sigma) = af(\mu + b, a\sigma) \tag{211}$$

gelten. Diese Gleichung hat die Lösung

$$f(\mu, \sigma) = \frac{\text{Konstante}}{\sigma}. \tag{212}$$

Diese a-priori-Verteilung wurde bereits von Jeffreys betrachtet und hat deshalb den Namen *Jeffreys-Regel*.

Beispiel 6.6 Poisson-Parameter Die Wahrscheinlichkeit, dass genau n Ereignisse im Zeitintervall t beobachtet werden, sei durch die Poisson-Verteilung

$$p(n|\lambda, t) = e^{-\lambda t}\frac{(\lambda t)^n}{n!} \tag{213}$$

gegeben. Aus der Anzahl der beobachteten Ereignisse soll nun der Parameter λ geschätzt werden. Man stelle sich zwei Beobachter vor, deren Uhren auf verschiedenen Zeitskalen laufen. Die Messungen eines bestimmten Intervalls sind dann durch die Beziehung $t = qt'$ aufeinander bezogen. Da beide das gleichen physikalische Experiment beobachten, sind die Raten λ und λ' durch $\lambda t = \lambda' t'$, also durch $\lambda' = q\lambda$ aufeinander bezogen. Die Beobachter wählen nun die a-priori-Verteilungen

$$p(d\lambda|\mathbf{X}) = f(\lambda)d\lambda, \quad p(d\lambda'|\mathbf{X}') = g(\lambda')d\lambda'. \tag{214}$$

Die Verteilungen müssen wechselseitig konsistent sein, so dass $f(\lambda)d\lambda = g(\lambda')d\lambda'$ gelten muß. Nun seien die beiden Beobachter vollständig unwissend bezüglich λ bzw. λ'. Also muß $f = g$ gelten, d.h. es muß gelten

$$f(\lambda) = qg(q\lambda), \text{ d.h. } p(d\lambda'|X') = \frac{d\lambda}{\lambda}. \tag{215}$$

Jede andere Wahl der a-priori-Verteilung bedeutet, dass eine Änderung der Zeitskala die Form der Verteilung änderte und damit ein anderes Maß an Wissen über λ ausdrücken würde. Aber die Annahme vollständiger Unwissenheit, entsprechend dem Indifferenzprinzip, legt die Form der a-priori-Verteilung durch (215) fest, wenn sie invariant gegenüber Skalentransformationen sein soll. □

Die Transformationen, die hier betrachtet wurden, waren linear. Linearität ist aber eine weder notwendige noch hinreichende Bedingung, wie das folgende Beispiel zeigt.

Beispiel 6.7 Unbekannter Bernoulli-Parameter Es werden n Bernoulli-Versuche mit unbekannter Erfolgswahrscheinlichkeit θ betrachtet; die Wahrscheinlichkeit von r Erfolgen sei also durch

$$P(r|n,\theta) = \binom{n}{r}\theta^r(1-\theta)^{n-r}. \tag{216}$$

Gesucht ist nun die a-priori-Verteilung, die vollständiges Uniwssen über θ ausdrückt.

Die "natürliche" Annahme ist, jedem Wert für θ zwischen 0 und 1 die gleiche Wahrscheinlichkeit zuzuordnen, so dass man die a-priori-Verteilung $f(\theta) = 1$ erhält. Aber die *rule of succession* zeigt merkwürdige Implikationen dieser Annahme. Der Ansatz, sich der Frage nach der a-prriori-Verteilung über Transformationsgruppen zu nähern, kann hier aber nicht so einfach wie im Falle der Parameter μ und σ durchgeführt werden, denn eine lineare Transformation von θ führt leicht aus dem Intervall $[0, 1]$ heraus. Gesucht ist eine Transformation, die θ wieder auf das Intervall $[0, 1]$ abbildet.

Jaynes (2003, p. 384) findet die Transformation über einen interessanten Umweg. $f(\theta)$ beschreibe nicht das Wissen einer Person, sondern die Verteilung der θ-Werte in einer Population von Individuen, bei der jede Person durch einen zunächst festen θ-Wert gekennzeichnet ist. Der Begriff der vollständigen Unwissenheit über θ muß nun auf die Population angewendet werden: f soll den Zustand vollständiger Konfusion bezüglich θ in der Population beschreiben. Jedes Mitglied der Population verändere aber sein Wissen nach Maßgabe des Bayesschen Satzes. Vor Beginn des eigentlichen Experimentes werde nun jedes Mitglied der Population mit der gleichen Evidenz E bezüglich θ versehen. Herr A habe davor die Ansicht gehabt, die Wahrscheinlichkeit eines Erfolges sei durch $\theta = p(S|A)$ gegeben. Diese Wert wird durch E in

$$\theta' = P(S|EA) = \frac{P(E|SA)P(S|A)}{P(E|SA)P(S|A) + P(E|FA)P(F|A)}$$

verwandelt. Dabei ist $P(F|X) = 1 - P(S|A)$, S steht für Erfolg (Success), und F für Mißerfolg (Failure). Der neue Wert θ' und der alte Wert θ sind demnach durch die Beziehung

$$\theta' = \frac{a\theta}{1 - \theta + a\theta} \tag{217}$$

aufeinander bezogen, mit $a = P(E|SA)/P(E|FA)$.

Die Population als Ganzes habe durch die neue Evidenz E nichts gelernt, durch konfligierende Propaganda sei sie in einen Zustand totaler Konfusion bzw. vollständiger Unwissenheit versetzt worden. Dies soll bedeuten, dass nach der Transformation (217) der Anteil der Personen mit $\theta_1 < \theta < \theta_2$ der gleiche ist wie vor der Gabe von E. Ist die a-priori-Verteilung vor der Transformation f und nach der Transformation g, so soll demnach

$$f(\theta)d\theta = g(\theta')d\theta', \tag{218}$$

und wenn die Population durch E nichts gelernt hat, muß darüber hinaus gelten

$$f(\theta) = g(\theta). \tag{219}$$

Kombiniert man nun (217), (218) und (219), so ergibt sich die Funktionalgleichung

$$af\left(\frac{a\theta}{1-\theta-a\theta}\right) = (1-\theta-a\theta)^2 f(\theta). \tag{220}$$

Die Gleichung läßt sich lösen durch Elimination von a über (217) und (220)), oder durch Differentiation nach a und anschließender Setzung $a = 1$. Es ergibt sich die Differentialgleichung

$$\theta(1\theta)f'(\theta) = (2\theta - 1)f(\theta) \tag{221}$$

mit der Lösung

$$f(\theta) = \frac{\text{Konstante}}{\theta(1-\theta)}. \tag{222}$$

Die vielen Leute aus der Population können nun wieder zu einer Person zusammengefasst werden. Hat man r Erfolge in n Versuchen beobachtet, so erhält man aus (216) und (222) die a-posteriori-Verteilung

$$P(d\theta|r,n) = \frac{(n-1)!}{(r-1)!(n-r-1)!}\theta^{r-1}(1-\theta)^{n-r-1}d\theta, \tag{223}$$

und diese Verteilung hat den Ewartungwert und die Varianz

$$\mathbb{E}(\theta) = \frac{r}{n}, \quad \mathbb{V}(\theta) = \frac{r(n-r)}{n^2(n+1)}. \tag{224}$$

Also ist die beste Schätzung für θ durch r/n gegeben, und r/n ist auch die Wahrscheinlichkeit des Erfolgs im folgenden Versuch(sdurchgang), wie es sich für eine Folge von Bernoulli-Versuchen gehört, im Unterschied zu der Vorhersage durch die Laplacesche Folgeregel, die $(r+1)/(n+2)$ behaupten würde und die sich aus der Annahme der Gleichverteilung für θ auf $[0,1]$ als a-priori-Verteilung ergibt. □

Curiosita: Keynes (1921/2008) verweist auf einige Arbeiten, die die Mühsal reflektieren, die das Verständnis von Folgen von Ereignissen machte, bevor sich Kolmogoroffs Axiomatik durchsetzte, die ebenfalls das Verständnis von 'Bernoullis Theorem' erleichterte. Dies besagt, nach Keynes (p. 338)

> "... in its simplest form ... If the probability of an event's occurrence under certain conditions is p, then, if these conditions are present on m occasions, the most probable number of the event's occurrences is mp

(or the nearest integer to this, i.e. the most probable *proportion* of its occurrences to the total number of occasions is p; further, the propability that the proportion p by less than a given amount b, increases as m increases, the value of this probability being calculable by a process of approximation."

Das ist also die Aussage, dass bei einer binomialverteilten Variablen X bei m Beobachtungen der Erwartungswert durch $\mathbb{E}(X) = mp$ und die Varianz durch $\mathbb{V}(X) = mp(1-p)$ gegeben ist; der Anteil der 'Erfolge' ist dann $\mathbb{E}(X)/m = p$; ist k die Anzahl der Erfolge und $\hat{p} = k/m$, so gilt bekanntlich

$$\lim_{m\to\infty} P(|\hat{p} - p|) = 0,$$

d.h. kleinere Abweichungen haben bei wachsendem m eine größere Wahrscheinlichkeit als größere Abweichungen; dies sieht man leicht ein, wegen (i) $\mathbb{E}(\hat{p}) = p$, also $\hat{p} \to p$, und $\mathbb{V}(\hat{p}) = p(1-p)/m$, also $\lim_{m\to\infty} \mathbb{V}(\hat{p}) \to 0$. Keynes berichtet, dass Simon Laplace der Ansicht war, die hier genannten Sachverhalte seien Ausdruck eines allgemeinen Naturgesetzes. Sein berühmtes Werk *Essai philosophique sur la probabilité* aus dem Jahre 1812 war ursprünglich Napoleon (*A Napoléon-le-Grand*) gewidmet. In der Neuauflage aus dem Jahr 1814 hat er diese Widmung ersetzt durch eine Deutung des Bernoullischen Theorems. Es bringe zum Ausdruck, dass jede große Kraft, die, trunken an ihrer Liebe zur Eroberung und universeller Herrschaft, am Ende zum Niedergang gezwungen werde[20]. Allgemein nahm man an, dass Bernoullis Theorem auf alle "Korrekt" berechneten Wahrscheinlichkeiten anzuwenden sei. So hat man längliche Auszählungen von Folgen von Ereignissen beim Roulette vorgenommen (Beispiele bei Keynes, Seite 363). Besonderes Interesse erregten die Untersuchungen von Dr. Karl Marbe (der später Professor und Begründer des Würzburger Instituts für Psychologie wurde). Marbe betrachtete die Folgen von 80 000 Würfen beim Roulette in Monaco und ähnlichen Anstalten und untersuchte insbesondere das Auftreten bestimmter Folgen von Ereignissen. Seine Ergebnisse bestätigten ihn in seiner Ansicht, dass die Welt so strukturiert sei, dass lange "runs" nicht nur unwahrscheinlich seien, sondern überhaupt nicht vorkämen. Er versuchte also, eine metaphysische Aussage über das Universum anhand der Eigenschaften des Roulettes zu bestätigen (Marbe (1899, 1916)). Keynes führt aus, dass Marbes (1899) Buch vor allem in Deutschland diskutiert worden sei, – aber nicht wegen der grotesken (preposterous) Art und Weise, ein allgemeines Gesetz konstituieren zu wollen, sondern weil seine Interpretation der Daten nicht wirklich aus den Daten folgt und wegen eines kleinen Fehlers, der Marbe dazu brachte, die Wahrscheinlichkeiten der Folgen korrekt zu berechnen. Man mag darüber spekulieren, welch philosophischer Zeitgeist in Deutschland dazu beigetragen hat, dass die Marbeschen Ideen hier mehr verfingen als in anderen Ländern; schließlich geht es nur um die richtige Verwendung des Begriffs unabhängiger Ereignisse.

6.4.5 Bernardos Referenz-Priors

Lindley (1951) wandte den von Shannon 1948 definierten Begriff der Information auf die Frage, wie viel Information durch ein Experiment geliefert wird, an. Es soll $p(x)$ die Wahrscheinlichkeitsverteilung für Beobachtungen (Messungen) x

[20] "C'est encore un résultat du calcul des probabilités, confirmé par des nombreuses et funestes expériences."

aus der Menge X der möglichen Messungen bedeuten, und $p(\theta)$ die Wahrscheinlichkeitsverteilung für den Parameter θ; diese (von Lindley eingeführte) Notation soll nicht bedeuten, dass die Verteilungen für x und θ gleich sind. $p(x|\theta)$ ist dann die bedingte Verteilung von x, gegeben θ (es wird also auch nicht zwischen der allgemeinen Bezeichung einer zufälligen Veränderlichen und ihrem speziellen Wert unterschieden). Nach Bayes hat man dann

$$p(\theta|x) = p(x|\theta)p(\theta)/p(x), \quad p(x) = \int p(x|\theta)p(\theta)d\theta. \tag{225}$$

Für die Verteilung $p(\theta)$ ist dann die (Shannon-) Information und $d\theta$ durch

$$I_p = \int_\Theta p(\theta) \log p(\theta) d\theta \tag{226}$$

gegeben. , wobei $I = 0$ gesetzt wird für den Fall $p(\theta) = 0$. Formal entspricht I_p dem Erwartungswert von $\log p(\theta)$. Ist das Experiment dann durchgeführt worden und hat man x beobachtet, so ist die a-posteriori-Information durch

$$I_p(x) = \int_\Theta p(\theta|x) \log p(\theta|x)\, d\theta \tag{227}$$

gegeben. Die durch das Experiment übermittelte Information wird dann von Lindley durch die Differenz

$$I[x, p(\theta)] = I_p - I_p(x) \tag{228}$$

definiert.

Bernardo (1979) hat diesen Ansatz als Ausgangspunkt gewählt, um eine Klasse von *reference priors* zu definieren. Der Begriff der Referenz-Prior wird bereits von Box & Tiao (1973) verwendet, allerdings in einem etwas anderen Sinn: Referenz-Priors sind a-priori-Verteilungen, die als "neutral" in Bezug auf die möglichen Parameterwerte gelten können (Box & Tiao, p. 23). Bernardo (1978) geht davon aus, dass eine a-priori-Verteilung einen Zustand relativ geringen Wissens über den Wert eines Parameters θ reflektieren sollte. Die resultierende a-posteriori-Verteilung ist dann eine *Referenzverteilung* in Bezug auf andere Verteilungen, die eine Art Standard definiert, in Bezug auf den die relative Wichtigkeit diskutiert werden kann, die das anfängliche Wissen zur Bewertung des experimentellen Ergebnisses (d.h. Schätzung des Parameterwerts) beurteilt werden kann. Historisch ist das von Bayes und Laplace eingeführte *principle of insufficient reason* der erste Ansatz dieser Art.

Bernardo definiert ein Experiment durch $\varepsilon = \{X, \Theta, p(x|\theta)\}$. Dabei ist X die zufällige Veränderliche, deren Wert im Experiment bestimmt wird, Θ der Parameterraum, und $p(x|\theta)$ die Likelihood von x, gegeben θ. Mit $p(\theta|x) = p(x|\theta)/p(x)$ definiert Bernardo die durch ε gelieferte erwartete Information

$$I^\theta[\varepsilon, p(\theta)] = \int p(x) \int p(\theta|x) \log \frac{p(\theta|x)}{p(\theta)}\, d\theta\, dx. \tag{229}$$

Es wird nun eine *Referenz-Posterior* konstruiert. Hierin ist

$$\int p(\theta|x) \log \frac{p(\theta|x)}{p(\theta)}\, d\theta$$

die *Kullback-Leibler-Distanz* (KL-Distanz) zwischen der a-posteriori-Verteilung $P(\theta|x)$ und der a-priori-Verteilung $P(\theta)$ (Kullback & Leibler (1951)) (eine ausführlichere Darstellung der KL-Distanz findet sich im Anhang, Abschnitt 7.3), Seite 138.

Es sei $I^\theta[\varepsilon(k), p(\theta)]$ die erwartete Information, die bezüglich θ erwartet wird, wenn k voneinander unabhängige Wiederholungen des Experiments ε durchgeführt werden, und C sei die Klasse der a-priori-Verteilungen, die mit einer objektiv gegebenen Anfangsinformation über θ verträglich sind. Für $k \to \infty$ würde man den präzisen Wert von θ erfahren. Demnach mißt $I^\theta[\varepsilon(\infty), p(\theta)]$ den Betrag fehlender Information, wenn die a-priori-Verteilung durch $p(\theta)$ gegeben ist. Mit "vager Anfangsinformation" wird nun diejenige a-priori-Verteilung π_0 bezeichnet, die die fehlende Information relativ zur Klasse C maximiert. Die Referenz-Posteriori-Verteilung $\pi(\theta|x)$ erhält man nun aus Bayes' Theorem

$$\pi(\theta|x) \propto p(x|\theta)\pi(\theta).$$

Ist $\Theta \subseteq \mathbb{R}$ ein Kontinuum, so ist

$$I^\theta[\varepsilon(\infty), p(\theta)] = \infty,$$

da die Bestimmung einer reellen Zahl $\theta \in \Theta$ unendlich viel Information erfordert. Deswegen betrachtet Bernardo den Grenzwert $\lim_{k\to\infty} p_k(\theta) = p(\theta)$. Die Definition der Referenz-Prior ist dann

Definition 6.4 *Es sei x das Ergebnis eines Experiments $\varepsilon = \{X, \Theta, p(x|\theta)\}$, C die Klasse der zulässigen a-priori-Verteilungen. Die Referenz-Posteriori nach Beobachtung von x ist* $\lim_k \pi_k(\theta)$*, wobei $\pi_k(\theta|x) \propto p(x|\theta)\pi_k(\theta)$ die a-posteriori-Dichte für die a-priori-Dichte π_k ist, die $I^\theta[\varepsilon(k), p(\theta)]$ in C maximiert. Eine* Referenz-Prior *ist eine Funktion $\pi(\theta) > 0$, die $\pi(\theta|x) \propto p(x|\theta)\pi(x)$ erfüllt.*

Es liegt nahe, zu fragen, warum nicht gleich $\lim_k \pi(\theta)$ betrachtet wird. Dazu betrachte man die a-priori-Verteilung

$$\pi_k(\theta) = B(\theta|1/k, 1/k) = \frac{\Gamma(1/k + 1/k)}{\Gamma(1/k)\Gamma(1/k)}\theta^{\frac{1}{k}-1}(1-\theta)^{\frac{1}{k}-1}.$$

Für $k \to \infty$ erhält man

$$\lim_k \pi_k(\theta) \propto \theta^{-1}(1-\theta)^{-1},$$

d.h. die Haldane-Prior (der Faktor $\Gamma(2/k)/\Gamma^2(1/k)$ strebt gegen Null für $k \to \infty$; damit die Grenzverteilung nicht identisch 0 wird, muß eine andere Proportionalität eingeführt werden). Die Definition 6.4 liefert aber eine andere a-priori-Verteilung:

$$\lim_k \pi_k(\theta) = \pi(\theta = 0) = \pi(\theta = 1) = \frac{1}{2}.$$

Es ist diese a-priori-Verteilung, die $\pi(\theta|x)$ im geforderten Sinn maximiert.

Beispiele: Zunächst sei der Fall betrachtet, dass θ nur endlich viele Werte annehmen kann. Die Referenz-Prior für die nicht beschränkte Klasse C aller Wahrscheinlichkeitsverteilungen für θ ist dann die Gleichverteilung $\pi(\theta) = \{1/m, \ldots, 1/m\}$. Der fehlende Betrag an Information $I^\theta[\varepsilon(\infty), p(\theta)]$ ist dann gleich der Entropie $H(p(\theta))$, und die Referenz-Prior ist gerade diejenige Verteilung, die die Entropie maximiert. Das ist aber gerade die Annahme, von der Jaynes (1968/2003) ausgeht.

In dem speziellen kontinuierlichen Fall, dass die a-posteriori-Verteilung asymptotisch normal ist erhält man Jeffreys' Prior; dies gilt auch für den multivariaten Fall. Kass & Wasserman (1996) liefern einen kritischen Überblick über Methoden, a-priori-Verteilungen durch formale Regeln zu bestimmen, wobei insbesondere auch auf den Bernardoschen Ansatz eingegangen wird.

6.4.6 Asymptotik und der Washing-out-Effekt

Die Notwendigkeit, eine a-priori-Verteilung wählen zu müssen, steht im Zentrum der Kritik am Bayesschen Ansatz. Hierdurch würde der Subjektivität der Interpretation von Daten Tür und Tor geöffnet. Die ausgiebigen Versuche, a-priori-Verteilungen durch formale Regeln gewissermaßen algorithmisch zu bestimmen, sind Versuche, diese Subjektivität zu überwinden, – daher die Rede vom 'Objektiven Bayesianismus'. Die Frage ist jedenfalls, in welchem Ausmaß sich eine u. U.falsch gewählte a-priori-Verteilung auf die a-posteriori-Verteilung auswirkt. Es könnte doch sein, dass sich zumindest für große Stichproben eine spezielle Wahl einer a-priori-Verteilung sich nicht weiter auswirkt, die Daten also letztlich die a-priori-Verteilung dominieren. Das ist in der Tat vielfach der Fall, wie die folgenden Beispiele nahelegen.

Beispiel 6.8 Gauß-a-prioris Box & Tiao (1973) Es soll ein bestimmter Parameter θ geschätzt werden; es kann sich dabei um eine physikalische Konstante, um eine psychophysische Größe oder Ähnliches handeln. Generell gilt

$$P(\theta|x) \propto P(x|\theta)P(\theta),$$

$P(\theta)$ die a-priori-Verteilung für θ, die das Vorwissen über θ wiedergibt, $P(x|\theta)$ die Likelihood-Funktion, und $P(\theta|x)$ die a-posteriori-Verteilung. Zwei Wissenschaftler A und B haben allerdings unterschiedliche Einschätzungen über den θ-Wert, die sich in unterschiedlichen a-priori-Verteilungen ausdrücken lassen. diese Verteilungen seien durch

$$P_A(\theta) = \frac{1}{20\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{\theta - 900}{20}\right)^2\right] \tag{230}$$

$$P_B(\theta) = \frac{1}{80\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{\theta - 800}{80}\right)^2\right] \tag{231}$$

gegeben. A nimmt an, dass θ einen Wert bei $\hat{\theta} = 900$ hat, und seine Unsicherheit wird durch $\sigma = 20$ reflektiert. B nimmt an, der Wert liege bei $\hat{\theta} = 800$, und seine Unsicherheit wird durch $\sigma = 80$ angegeben. Die experimentelle Methode zur Bestimmung von θ sei frei von systematischen Fehlern ('bias-frei'). Die Likelihood-Funktion sei wieder durch eine Gauß-Funktion gegeben. Für n voneinander unabhängige Messungen $\mathbf{x} = (x_1, \ldots, x_n)$ ergibt sich die Likelihood-Funktion

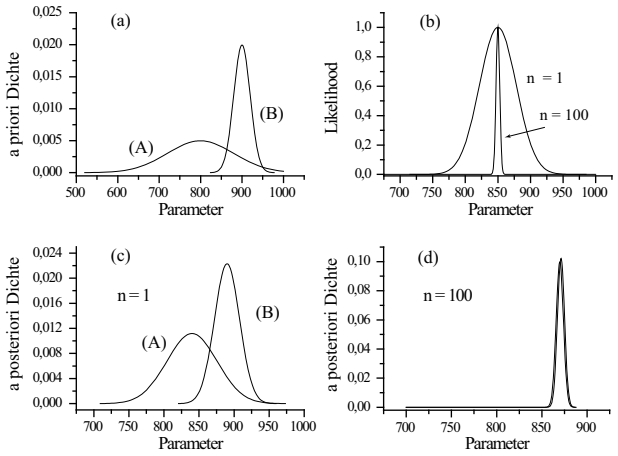
$$P(\mathbf{x}|\theta) \propto \exp\left[-\frac{1}{2}\left(\frac{\theta - \bar{x}}{\sigma/\sqrt{n}}\right)^2\right], \quad \bar{x} = \frac{1}{n}\sum_i x_i. \tag{232}$$

Es läßt sich zeigen, dass die a-posteriori-Verteilung wieder eine Gauß-Dichte ist mit dem Erwartungswert $\bar{\theta}$ und der Varianz $1/\bar{\sigma}^2$, wobei

$$\bar{\theta} = \frac{w_0\theta_0 + w_1 x}{w_0 + w_1}, \quad \frac{1}{\bar{\sigma}^2} = w_0 + w_1, \tag{233}$$

mit $1/\sigma_0^2$, $w_1 = 1/\sigma^2$ (die Gauß-Prior ist eine konjugierte a-priori-Verteilung). Abbildung 13 zeigt die Ergebnisse. Ein wichtiger Punkt dieses Beispiels ist, dass verschiedene a-priori-Verteilungen für kleine Werte von n auch verschiedene a-posteriori-Verteilungen bedeuten. Mit größer werdendem Wert von n konvergieren

Abbildung 13: A-priori- und a-posteriori-Verteilungen: (a) zwei verschiedene a-priori-Verteilungen für den Wert eines zu bestimmenden Parameters, (b) standardisierte Likelihoodfunktionen für $n = 1$ und $n = 100$, (c) a-posteriori-Verteilungen für die jeweiligen a-priori-Verteilungen, $n = 1$ Messung (d) a-posteriori-Verteilungen, $n = 100$ Messungen; die a-posteriori-Verteilungen für (A) und B sind kaum noch zu unterscheiden.



die a-posteriori-Verteilungen aber oft gegen *eine* Verteilung; die Daten dominieren die a-priori-Verteilungen, deren Varianz überdies immer kleiner wird.

In Bezug auf das Evidenzproblem ist klar, dass die a-posteriori-Verteilung die Evidenz der Daten für die jeweilige Hypothese direkt angibt. In Termen der a-posteriori-Verteilung kann ein Konfidenzintervall für den Parameter θ angegeben werden; man spricht auch von einem HPD-Intervall ("highest posterior density interval") oder *Kredibilitätsintervall* $1-\alpha \in (0,1)$. Es ist ein Intervall $I = [t_0, t_1] \subset \Theta$, wobei $f(\theta|\mathbf{x}) \geq f(\tilde{\theta}|\mathbf{x})$ gilt mit $\theta \in I$, $\tilde{\theta} \notin I$. □

Beispiel 6.9 Bernoulli-Versuche: Gelegentlich hat man statt einer endlichen Menge von Hypothesen eine unendliche Menge. Ein einfaches Beispiel ist der Fall einer Binomialverteilung mit unbekanntem Parameter θ. Man habe n Bernoulli-Versuche mit r "Erfolgen", so dass

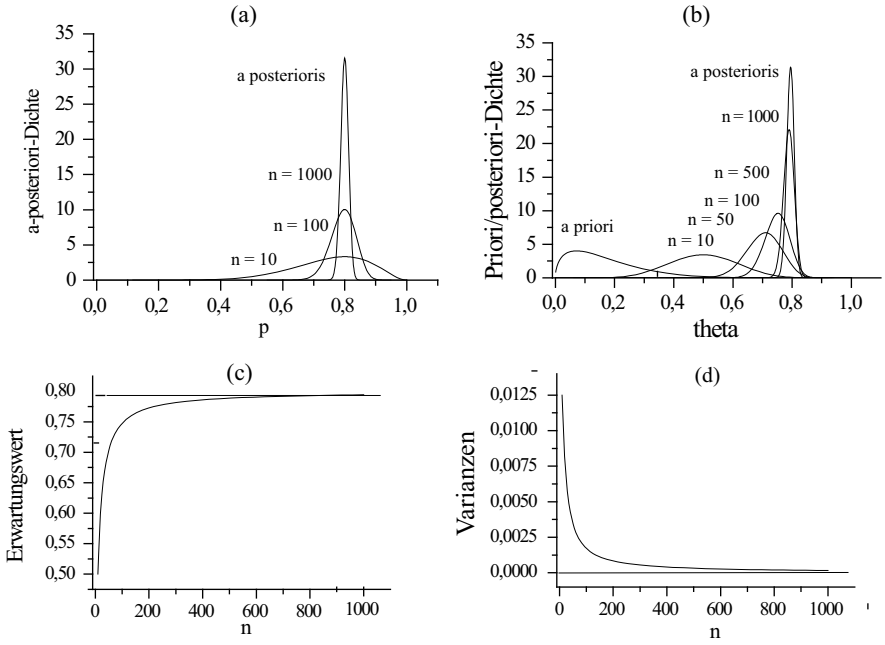
$$P(r|n,\theta) = \binom{n}{r}\theta^r(1-\theta)^{n-r}.$$

Dann ist nach Bayes

$$P(\theta|n,r) = P(r|n,\theta)\frac{f(\theta)}{P(r,n)} \tag{234}$$

Ist nichts über den Wert von θ bekannt, kann man nach dem Indifferenzprinzip die Gleichverteilung für die θ-Werte annehmen. Die Dichte für eine über dem (offenen) Intervall (a,b) gleichverteilte Variable ist $1/(b-a)$, so dass speziell für $a = 0$ und $b = 1$ die Dichte $f(\theta) = 1$ folgt. Für $P(r,n)$ hat man dann nach dem

Abbildung 14: (a) A-posteriori-Verteilungen für den Parameter $\theta_0 = .8$ einer Binomialverteilung bei uniformer a-priori-Verteilung. Das Maximum der a-posteriori-Verteilung liegt bei $r/n = \theta_0 = .8$, in Übereinstimmung mit der Maximum-Likelihood-Schätzung. (b) a-priori-Verteilung (Beta-Verteilung $B(a, b)$ mit $a = 1.5$, $b = 7.5$) und a-posteriori-Verteilungen für den Parameter θ einer Binomialverteilung mit dem Parameter $\theta_0 = .8$; (b) Erwartungswerte der a-posteriori-Verteilungen: schnelle Konvergenz gegen θ_0 für Werte bis $n \approx 200$ (d) Varianzen der a-posteriori-Verteilungen in Abhängigkeit von n, schnelle Konvergenz gegen Null bis $n \approx 150$.



Satz der Totalen Wahrscheinlichkeit

$$p(r, n) = \int_0^1 \binom{n}{r} \theta^r (1-\theta)^{n-r} f(\theta) d\theta = \binom{n}{r} \int_0^1 \theta^r (1 - \theta^{n-r} d\theta. \tag{235}$$

Bekanntlich gilt

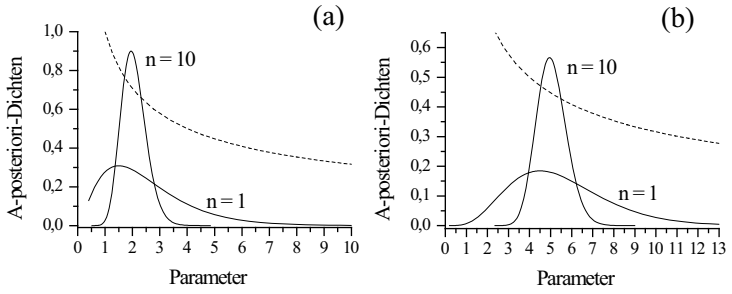
$$\int_0^1 \theta^r (1-\theta)^{n-r} d\theta = \frac{r!(n-r)!}{(n+1)!}, \tag{236}$$

und (234) liefert

$$P(p|n, r) = \frac{\theta^r (1-\theta)^{n-r}}{\int_0^1 \theta^r (1-\theta)^{n-r} d\theta} = \theta^r (1-\theta)^{n-r} \frac{(n+1)!}{r!(n-r)!} \tag{237}$$

für die a-posteriori-Verteilung von θ. Abb. 14 zeigt vier a-posteriori-Dichten für verschiedene n- und r-Werte. □

Abbildung 15: Poisson-a-posteriori-Verteilungenfür den Poisson-Parameter λ: (a) $\bar{k} = 2$, $n = 1$ und $n = 10$, (b) $\bar{k} = 5$, $n = 1$ und $n = 10$. Gestrichelte Kurve: A-priori-Verteilung $1/\sqrt{\lambda}$.



Beispiel 6.10 Poisson-Parameter Gegeben seien Häufigkeiten $\mathbf{x} = (k_1, k_2, \ldots, k_n)$, von denen angenommen werden kann, dass sie aus einer Poisson-verteilten Population stammen, d.h. es soll

$$P(X = k|\lambda) = e^{-\lambda}\frac{\lambda^k}{k!}, \quad k = 0, 1, 2, \ldots \tag{238}$$

also $\theta = \lambda$. Die Likelihood-Funktion für die Daten ist

$$\mathcal{L}(\lambda|\mathbf{x}) = \prod_{i=1}^{n} P(X = k_i|\lambda) = \frac{e^{-n\lambda}\lambda^{\sum_i k_i}}{\prod_i k_i!} = \frac{e^{-n\lambda}\lambda^{n\bar{k}}}{\prod_i k_i!}, \quad \bar{k} = \frac{1}{n}\sum_i k_i$$

Der Übergang zu $L = \log \mathcal{L}$ liefert

$$L(\lambda|\mathbf{x}) = -n\lambda + n\bar{k}\log\lambda + c, \quad c = -\log\sum_i k_i! \tag{239}$$

Hieraus ergibt sich aus $dL/d\lambda|_{\lambda=\hat{\lambda}} = 0$ sofort die Maximum-Likelihood-Schätzung $\hat{\lambda}$ für λ:

$$\hat{\lambda} = \bar{k}. \tag{240}$$

Wählt man für die a-priori-Verteilung

$$f(\lambda) = \frac{1}{\sqrt{\lambda}} \tag{241}$$

(diese Wahl wird später noch begründet), so erhält man für die a-posteriori-Verteilung

$$P(\lambda|\mathbf{x}) = C\lambda^{n\bar{k}}e^{n\lambda}, \quad \lambda > 0, \tag{242}$$

wobei C eine Normalisierungskonstante ist, für die man

$$\frac{1}{C} = \int_0^\infty \lambda^{n\bar{k}}e^{n\lambda}d\lambda = n^{-(n\bar{k}+1/2}\Gamma(n\bar{k} + 1/2),$$

also

$$C = \frac{n^{n\bar{k}+1/2}}{\Gamma(n\bar{k} + 1/2)} \tag{243}$$

findet. Die rechte Seite ist aber die Dichte einer zufälligen Veränderlichen, die wie $\chi^2/2$ mit df $= 2n\bar{k} + 1$ Freiheitsgraden verteilt ist; dies ist also die a-posteriori-Verteilung für λ. Abb. 15 zeigt verschiedene a-posteriori-Verteilungen für den Parameter λ. Für $n = 1$ liegt das Maximum der Verteilung links von $\bar{k}$, während für $n = 10$ die Verteilung nicht nur angenähert symmetrisch ist, sondern ein Maximum in der Nähe von $\bar{k}$ hat. □

Bereits Edwards et al. (1963) haben darauf hingewiesen, dass, wenn die Daten hinreichend präzise erhoben wurden, die Form und andere Eigenschaften der a-priori-Verteilung einen vernachlässigbaren Effekt auf die a-posteriori-Verteilung haben (p. 201). Die Subjektivität wird von den Daten überrollt. Blackwell & Dubins (1962) haben die Voraussetzungen hierfür in einem streng formalen Rahmen diskutiert.

Earman (1992) liefert ein weiteres Argument für die 'Konvergenz der Meinungen' (merger of opinion), das auf dem Doobschen Martingal-Konvergenzsatz beruht, und diskutiert einen Ansatz von Gaifman & Snir (1982), diese Betrachtungen in den philosophischen Rahmen der Konfirmationstheorie einzubetten, wobei Wahrscheinlichkeiten Aussagen einer formalen Sprache zugeordnet werden und die experimentellen Ergebnisse in wahrheitsfunktionalen 'atomaren Aussagen' formuliert werden. Details können hier nicht gegeben werden, aber Earmans Anmerkung, dass der positive Eindruck, der durch den Gaifman-Snir Ansatz zunächst hervorgerufen wird, "disappears in the light of their narcistic character" (wobei sich "their" auf die Aussagen von Gaifman & Snir bezieht). Gleichwohl, Earman kommt insgesamt, was die Objektivität der Dateninterpretation angeht, letztlich zu einer optimistischen Aussage.

6.4.7 Kritik und weitere Entwicklungen

Seidenfeld (1979) gibt eine klare Darstellung des objektiven Ansatzes. Er beginnt mit drei Prinzipien, die jeder Form des Bayesianismus zugrunde liegen:

(i) die Ansichten (beliefs) einer Person lassen sich durch eine Wahrscheinlichkeitsfunktion $P(\cdot)$ abbilden; diese Annahme repräsentiert das *Kohärenzprinzip*. In der Notation von Seidenfeld, die dieses Prinzip verdeutlichen, hat man dementsprechend

(ii) Änderungen in diesem System von Ansichten werden durch den Bayesschen Satz beschrieben: ist K die deduktiv geschlossene Wissensbasis, so beschreibt $P_K(\cdot)$ den Zustand der Person. d sei neue Evidenz; $P_{K^*}(\cdot)$ bvezeichnet die hypothetischen Ansichten, die durch d erzeugt werden, indem d zu K hinzugefügt wird, d.h. K^* bezeichnet die deduktiven Konsequenzen aus $K\&d$, und es gilt

$$P_K(\cdot|d) = P_k(d|\cdot)P_K(\cdot).$$

Es ist $P_{K^*} = P_K(\cdot|d)$, und diese Gleichsetzung wiederum nennt Seidenfeld das *Prinzip der Konditionalisierung* (wegen der bedingten Wahrscheinlichkeiten in Bayes' Theorem).

(iii) Die Wahrscheinlichkeitsfunktion $P_k(\cdot)$ hat als Basis das gesamte Wissen der Person; dies ist das *Prinzip der totalen Evidenz*.

Weiter unerscheidet Seidenfeld zwischen dem subjektiven Bayesschen Ansatz, wie er von Savage und de Finetti vertreten wird und denen zufolge die induktive Logik keine Begrenzung der akzeptablen Wissens- bzw. Glaubenszustände (belief states) hat; jede Wahrscheinlichkeitsfunktion, die (i) – (iii) genügt ist zulässig. Der Gegenpol zu diesem subjektiven Bayesianisums ist der Objektivismus, wie er insbesondere von Jeffreys und Jaynes vertreten wird. Deren Ansatz ist durch die Annahme charakterisiert, für einen Wissenszustand K existiere eine eindeutige, zulässige Wahrscheinlichkeitsfunktion P_K. P_K heißt *objektiv*, weil P_K nur durch K, nicht aber durch eine spezielle Person definiert ist. Damit erfordert der objektive Ansatz bestimmte Postulate, um P_K zu bestimmen.

Seidenfelds Arbeit liefert eine Kritik des objektiven Ansatzes. Grundlage dieses Ansatzes seien zwei Annahmen. Um für eine statistische Inferenz zu taugen, müßten bestmmte 'Ignoranzwahrscheinlichkeiten' existieren. So sei H eine Hypothese, die sich auf ein Intervall beziehe, in dem ein zu schätzender Parameter liege. Es soll ein 95 %-Intervall für H gefunden werden, gegeben die Daten d. Damit also $P(H|d) = .95$, sei eine spezielle a-priori-Wahrscheinlichkeit – eben die Ignoranzwahrscheinlichkeit – $P(H)$ notwendig. $P(H)$ sei 'informationslos' und repräsentiere die Ignoranz in Bezug auf H, *bevor* die Evidenz d gewonnen sei. Dieser Ansatz geht primär auf Jeffreys zurück und wird im Prinzip von Jaynes übernommen. Man sei etwa an einer Größe m interessiert. Aufgrund des gegenwärtigen Wissens wird angenommen, dass m auf einen Bereich $M \subset \mathbb{R}$ beschränkt ist. Es wird eine Funktion $f(m)$ definiert, f: $M \to \Theta$, also $f(m) = \theta \in \Theta$ für $m \in M$. θ ist der 'parameter of interest', Θ ist der Parameterraum. Die Frage ist dann, wie die anfängliche Ignoranz bezüglich m durch eine Wahrscheinlichkeitsfunktion P_K repräsentiert werden kann, wenn alles, was bekannt ist, die Aussage $\theta \in \Theta$ ist.

Nach Jeffreys wird Θ betrachtet und relativ dazu die 'Ignoranzverteilung' definiert. Für den Fall $\Theta = (-\infty, \infty)$ wird eine Gleichverteilung über Θ angenommen, – die natürlich uneigentlich (improper) ist. Für $\Theta = (0, \infty)$ wird die ebenfalls uneigentliche a-priori- Verteilung $P_k = 1/\theta$ angenommen. Man findet bei Jeffreys (zunächst) keine weitere Begründung für diese Annahme, – was die Lektüre des Buches erschwert, da man nicht weiß, ob man etwas nicht verstanden oder etwas überlesen hat (die Arbeit von Robert et al. erweist sich hier als nützlich). Seidenfeld bezeichnet denn auch Jeffreys Argument, die Wahl dieser a-prioi-Verteilungen sei vorteilhaft, weil sie konsistent mit anderen Parametrisierungen für m sei, als 'aprioristisch'. Betrachte man etwa die Parametrisierung θ^r, so werde man auf die gleichen a-priori-Verteilungen geführt. Ebenso führe eine lineare Transformation der Parameter auf die gleiche a-priori-Transformation. Überführe man einen auf $(0, \infty)$ definierten Parameter θ durch Logarithmierung in einen auf $(-\infty, \infty)$ definierten Parameter, so seien die beiden a-priori-Verteilungen konsistent mit dieser Transformation, d.h. eine uneigentliche Dichte $\propto 1/\theta$ sei äquivalent einer Gleichverteilung von $\log \theta$ $(\theta > 0)$ über dem Bereich von $\log \theta$.

Seidenfeld weist darauf hin, dass die beiden a-priori-Verteilungen nicht eigentlich informationslos seien. Ist etwa M begrenzt, z.B. $M = (0, 1)$, so müsse man entweder die Gleichverteilung auf $(0, 1)$ wählen oder $\theta = m/(1 - m)$. Eine Rechtfertigung für Jeffreys P_K-Funktionen ergab sich, als Jeffrey bemerkte, dass klassische statistische Tests nur dann Bayesiansische Modelle hatten, wenn er 'seine' P_k-Funktionen annahm. Damit wollte er die klassischen Tests auf 'philosophisch vernünftige' Basis stellen. Seine Methode ist besonders erfolgreich dann, wenn es etwa darum geht, den Erwartungswert und die Varianz von Normalverteilungen zu schätzen. Jeffreys Theorie der Invarianz lieferte dann eine systematische Basis für

seine Schätztheorie. Sie geht davon aus, dass die statistische Verteilung der Größe m in Rechnung gestellt werden muß. So sei man daran interessiert, das Massezentrum einer verbogenen Münze zu bestimmen. Das Experiment bestehe darin, die Münze zu werfen und jesweils zu registrieren, ob 'Kopf' oder 'Zahl' oben liegt. Das Modell für dieses Experiment ist das übliche Binomialmodell, und es wird eine Beziehung zwischen m und dem Binomialparameter θ angenommen. Die Vermutung, dass (i) das Massezentrum auch das Zentrum der Münze ist (ii) dass die Münze fair ist, führt zur Hypothese, dass $\theta = .5$ ist. Die von Jeffreys betrachteten Invarianten bezieen sich auf Verteilungen; sein Invarianzprinzip liefert eine Regel, nach der die 'Ignoranzverteilung' (also die a-priori-Verteilung) bestimmt wird. Die Invarianten sind 1-1 differenzierbare Transformationen der Parameter oder der zufälligen Veränderlichen, die in einer statistischen Verteilung enthalten sind. So erhält er

- für den Binomialparameter die 'informationslose' a-priori-Verteilung

$$f(\theta) \propto \frac{d\theta}{\pi\sqrt{\theta(1-\theta)}}$$

- für den Lokationsparameter μ aus $(-\infty, \infty)$ die Dichte $f(\mu) \propto d\mu$,
- für den Skalenparameter σ aus $(0, \infty)$ ergibt sich die Verteilung $d(\sigma) \propto d\sigma/\sigma$.

Aus Jeffreys Ansatz ergibt sich allerdings ein Problem. Man betrachte den Fall, dass zwei voneinander unabhängige Experimente durchgeführt werden, die die Evidenz d_1 und d_2 liefern, und beide Ergebnisse mögen einen Einfluß auf die Wahrscheinlichkeit, mit der H bewertet wird, haben. Dieser Einfluß wird durch die oben genannte Konditionalisierung repräsentiert:

$$P(H|d_i) \propto P(d_i|H)P(H), \quad i = 1, 2$$

Diese Beziehung gilt, wenn d_i die zuerst eintreffende Evidenz ist. Ist d_1 die Evidenz aus dem ersten Experiment, so muß für die Konditionalisierung durch d_2 für $P(H)$ die Größe $P(H') = P(H|d_1)$ eingesetzt werden. Ist d_2 die zuerst eintreffende Evidenz, so muß für $P(H)$ die Größe $P(H'') = P(H|d_2)$ eingesetzt werden. Es ist aber keineswegs gesagt, dass die Invarianzregel zur Gleichung

$$P(H|d_1) = P(H'|d_2)$$

führt ($P(H|d_1)$ und $P(H'|d_2)$ sind Verteilungen für einen Parameter!). Diese Situation ist keine akademische Haarspalterei, sondern spiegelt den Wissenschaftsbetrieb wieder: in verschiedenen Labors werden zu verschiedenen Zeitpunkten Untersuchungen zu den gleichen Hypothesen gemacht, deren Resultate in den allgemeinen Wissenshintergrund K_H eingehen. Seidenfeld illustriert das Argument mit einem einfachen Beispiel:

Beispiel 6.11 (Kubusvolumen) Es soll das Volumen eines Kubus bestimmt werden. Es werden zwei Methoden betrachtet: (i) der Kubus wird mit einer Flüssigkeit bekannter Dichte gefüllt, und dann wird der Inhalt gewogen, wobei vorausgesetzt werden, dass hierbei kein systematischer Fehler (Bias) auftritt. Die Messungen seien normalverteilt mit der Varianz 1, d.h. man hat die Verteilung $N(v, 1)$, v der Erwartungswert der Messungen. Die Invarianzregel erfordert dann, dass eine uneigentliche 'Ignoranzfunktion' – also a-priori-Verteilung – gewählt wird, also

eine Gleichverteilung für die Gewichte der Flüssigkeit. (ii) Es wird ein Stab von der Länge einer Kubusseite geschnitten, wobei das Material des Stabes wiederum eine bekannte Dichte haben möge. Dann wird der Stab gewogen. Der Invarianzregel entsprechend wird eine 'informationslose' a-priori-Verteilung für die Gewichte und damit für die Länge gewählt, da Länge und Gewicht fest gekoppelt sind. Das Volumen des Kubus und die Länge des Stabes stehen in der Beziehung $v = l^3$ zueinander. Damit führt die Invarianzregel zu einer Gleichverteilung von $l = v^{1/3}$. Die a-priori-Verteilungen für das Volumen hängen nun davon ab, welches Ergebnis zuerst eintrifft – das aus der Untersuchung (i) oder das aus der Untersuchung (ii) – und damit hat man auch unterschiedliche a-posteriori-Verteilungen. □

Seidenfeld argumentiert, dass damit gezeigt sei, dass die Problematik des Indifferenzprinzips nicht durch Jeffreyschen Ansatz aufgelöst werde. Der Fehler liege in der Forderung, dass 'Ignoranz' durch eine präzise angegebene Wahrscheinlichkeitsverteilung repräsentiert werden soll. Andererseits ist dies die Forderung, die den Jeffreysschen Ansatz zu einem *objektiven* Bayesschen Ansatz mache. Dieser Sachverhalt stelle das Jeffreyssche Programm insgesamt in Frage.

Auch Jaynes' Prinzip der maximalen Entropie wird von Seidenfeld kritisiert. Denn einerseits soll das Prinzip Indifferenz zum Ausdruck bringen. Andererseits zeigt Seidenfeld, dass für bestimmte Fälle keine Verteilung impliziert wird, die dem Indifferenzprinzip entspricht. So sei der Parameter θ auf $(-\infty, \infty)$ definiert und habe die a-priori-Verteilung $f(\theta)$ mit $\int_{-\infty}^{\infty} f(\theta)d\theta = 1$, dem Erwartungswert μ und der Varianz σ^2. Das MaxEnt-Prinzip liefert dann die a-priori-Verteilung $N(\mu, \sigma^2)$. Aber diese Verteilung repräsentiert nicht völlige Ignoranz bezüglich θ! Dementsprechend suggeriert die MaxEnt-Regel, dass mehr Information zur Verfügung steht als tatsächlich vorhanden ist.

6.5 Schlußbetrachtungen

Daten werden erhoben, um sie auf mögliche Strukturen hin zu explorieren, die bei der Formulierung von Hypothesen hilfreich sein können, oder aber um Hypothesen zu testen. In jedem Fall wird nach der Evidenz gefragt, die in den Daten steckt oder vermutlich steckt. Bei wissenschaftlichen Untersuchungen steht die Frage nach Entscheidungen, wie sie sowohl beim Fisherschen wie auch beim Neyman-Pearsonschen Ansatz gefordert werden, eher nicht im Vordergrund, wenn man einmal davon absieht, dass gelegentlich aus Zeit- und Kostengründen, manchmal auch aus Gründen der wissenschaftlichen Reputation die Fortsetzung einer bestimmten theoretischen Linie aufrechterhalten soll oder nicht. Auch wenn die Vorhersagen eines theoretischen Modells nicht signifikant von den Daten abweicht, kann es als letzlich unplausibel verworfen werden, oder umgekehrt kann es beibehalten werden, obwohl die Daten signifikant von den Vorhersagen abweichen, weil man Grund für die Vermutung hat, dass bei der Datenerhebung Fehler gemacht wurden, oder dass eine Modifikation von Auxiliarannahmen besser ist als die Aufgabe des Kerns des Modells. Vorgänge dieser Art sind insbesondere von Kuhn und Lakatos detailliert beschrieben worden. Signifikanz- und Hypothesentests haben eher die Funktion, Markierungen für eine intuitive Abschätzung der a posteriori-Wahrscheinlichkeit für das Modell zu setzen, also für eine letzlich subjektive Größe, die im Rahmen der orthodoxen statistischen Theorie eigentlich gar keine Rolle spielen sollte. Insofern kann man denjenigen Vertretern des Bayesschen Ansatzes Recht geben, die sagen, dass eben Bayessche Ansatz der eigentlich

objektive Ansatz sei, da er doch die vorgenommene Bewertung der Daten durch explizite Angabe der a priori-Wahrscheinlichkeiten transparent macht. Dies gilt auch für die Versuche, den Ansatz von Neyman & Pearson über explizite Angaben von erwarteten Effektgrößen fruchtbar zu machen. Cohens (1994) Rat: "..., don't look for a magic alternative to NHST[21], some other objective mechanical ritual to replace it. It doesn't exist." wird man kaum in dieser radikalen Form akzeptieren können. Letzlich sind die erwarteten Effektgrößen subjektive Größen, es liegt also nahe, sie in Form von a priori-Wahrscheinlichkeiten in die Diskussion der Daten einzubringen. Aber das hieße dann, von einem angeblich objektiven Wahrscheinlichkeitsbegriff abzulassen.

Von Wissenschaftstheoretikern wird bei der Diskussion der verschiedenen statistischen Ansätze oft die Frage nach der Möglichkeit der Induktion in den Vordergrund gestellt. Fisher und ebenfalls Neyman & Pearson sprechen gern von 'inductive behaviour', während etwa Karl Popper die Möglichkeit der Induktion nicht nur einfach negiert, sondern formal ihre Unmöglichkeit zu beweisen sucht. Autoren wie Howson stellen fest, dass an Humes Diktum von der Unmöglichkeit der Induktion, das ja auch hinter Poppers Philosophie steht, nicht zu rütteln sei, dass aber das in der Wissenschaft zu beobachtende unverkrampfte Verhältnis zur Induktion gleichwohl über den Bayesschen Ansatz zu rechtfertigen sei. Jaynes (2003; 310) zitiert Stove (1982), der von Philosophen wie Popper als von den Irrationalisten spricht, wenn er eine Antwort auf die Fragen (i) "how could such an absurd view ever have arisen?" und (ii) "by what linguistic practices do the irrationalists succeed in gaining an audience?" eine Antwort zu finden versucht.

In der Tat scheint der Bayessche Ansatz die Frage nach der Evidenz in den Daten am direktesten zu beantworten, die Beziehung

$$p(H|E) \propto P(E|H)P(H)$$

macht dies unmittelbar deutlich: die rechte Seite liefert eine Abschätzung der Wahrscheinlichkeit der Hypothese. Das Problem, das mit diesem Ansatz verknüpft wird, ist das der Wahr einer a priori-Wahrscheinlichkeit $P(H)$. Der Versuch, diese so "objektiv" wie möglich zu bestimmen, führt auf die Ansätze von Jeffreys, Jaynes und Bernardo und ihren Varianten; diese bestimmen den *Objektiven Bayes-Ansatz*; Berger (2006) hat die Vorteile dieses Ansatzes gegenüber dem *Subjektiven Bayesschen Ansatz* und die Art und Weise, in der Subjektivisten vom Objektiven Ansatz Gebrauch machen, dargestellt; der Subjektive Ansatz geht auf auf Rubin, de Finetti, Savage und Lindley zurück, s. a. Kaas & Wasserman (1996). Allerdings ist es nicht so, dass man nur zwischen Objektivem und Subjektivem Bayesschen Ansatz unterscheiden müßte. Good (1971) hat errechnet, dass es 46656 Ansätze gibt; die können hier nicht im Einzelnen referriert werden. Bayarri & Berger (2004) liefern eine Diskussion und Übersicht über die Wechselbeziehungen zwischen frequentistischem und Bayesschem Ansatz.

In der Einleitung sind einige Relevanzmaße eingeführt worden, (1), (2), (3), (3) und (8). Die Frage ist, ob sie ein konsistentes Bild der Evidenz in den Daten geben. Fitelson (1999) hat diese Maße, bzw. logarithmierte Varianten dieser Maße, in Bezug auf diese Frage diskutiert. Die von ihm betrachteten Versionen werden hier noch einmal zusammengefasst: zunächst wird (3) in einer etwas anderen Fassung

[21] Null Hypothesis Significance Testing

als allgemeines Relevanzmaß eingeführt:

$$c(H,K) \begin{cases} > 0, & \text{wenn } P(H|E\&K) > P(H|K) \\ < 0 & \text{wenn } P(H|E\&K) < P(H|K) \\ = 0, & \text{wenn } P(H|E\&K) = P(H|K) \end{cases} \tag{244}$$

wobei K wieder Hintergrundwissen repräsentiert. Dann hat man die Spezialfälle

$$d(H,E|K) =_{def} P(H|E\&K) - P(H|K) \tag{245}$$

$$r(H,E|K) =_{def} \log\left[\frac{P(H|E\&K)}{P(H|K)}\right] \tag{246}$$

$$l(H,E|K) =_{def} \log\left[\frac{P(E|H,K)}{P(E|\neg H,K)}\right] \tag{247}$$

$$\begin{aligned} r_c(H,E|K) &=_{def} P(H\&E\&K)P(K) - P(H\&K)P(E\&K) \\ &= P(K)P(E\&K)d(H,E|K). \end{aligned} \tag{248}$$

r_c ist auch als Carnaps Maß bekannt, der es in Carnap (1962), § 67 einführte; wie Fitelson (1999) ausführt, kann man Carnaps Maß als eine Art Kovarianzmaß interpretieren: wenn K tautologisch ist, ergibt sich

$$\begin{aligned} r_c(H,E|K) &= P(H\&E\&K)P(K) - P(H\&K)P(E\&K) \\ &= P(H\&E) + P(H)P(E) = \mathrm{Kov}(H,E). \end{aligned}$$

Die Maße sind Maße für die *inkrementelle Bestätigung* (incremental confirmation) von H durch E, gegeben K. Ein Argument $\mathcal{A}$ heißt *maßabhängig*, wenn die Validitäteinschätzung von $\mathcal{A}$ mit der Wahl eines der Maße (245) bis (248) variiert. Gilt $\mathcal{A}$ als valide unabhängig von der Wahl eines der Maße, so heißt $\mathcal{A}$ *maßunabhängig* oder einfach auch *robust*. Fitelson (1999) zeigt dann, dass viele Argumente bezüglich der Bayesschen Bestätigungstheorie maßabhängig sind; diese Tatsache stellt ein Problem für den allgemeinen Bayesschen Ansatz dar.

Weiter läßt sich zeigen, dass keines der möglichen Paare der Maße (245) bis (248) allgemein äquivalent ist, d.h. verwendet man die verschiedenen Maße der Reihe nach zur Beurteilung einer gegebenen Reihe von Hypothesen, so ergeben sich *verschiedene Rangordnungen* der Hypothesen hinsichtlich ihres Grades der Bestätigung.

Damit hat man sicherlich ein Problem, sofern man an einer allgemeinen Bayesschen Epistemologie interessiert ist, d.h. an einer Theorie der Entwicklung der Wissenschaft auf der Basis des Bayesschen Ansatzes. Forster (1995) in seiner Besprechung des Buches von Earman (1992) stellt eine Reihe von Fragen bezüglich dieser Epistemologie heraus, von denen er meint, sie seien nicht adäquat von Earman behandelt worden. Allerdings stellt er dann fest:

> "... a million of Bayesian statisticians can't all be wrong. Why is Bayesianism so popular? The short answer is that Bayesianism works in *statistical* inference, as opposed to scientific inference more generally." (Forster (1995), p. 402)

Wie Forster ausführt, funktioniere der Bayessche Ansatz gut, wenn es um die Schätzung von Parametern im Rahmen eines engen Kontextes K gehe; innerhalb dieses Rahmens verlieren die allgemeinen philosophischen, d.h. wissenschaftstheoretischen Einwände gegen den Bayesschen Ansatz ihre argumentative Kraft. Die

allgemeine wissenschaftliche Inferenz umfasse aber auch die Auswahl von K, generell der betrachteten Theorien. Das Verständnis dieser Prozesse stelle ein größeres und viel schwierigeres Problem dar.

Damit mag Forster Recht haben. Ob ein praktizierender Wissenschaftler allerdings eine solche allgemeine Theorie braucht, ist wiederum eine ganz andere Frage, – bisher ist die Wissenschaft ganz gut ohne eine solche allgemeine Theorie zurecht gekommen.

7 Anhang

7.1 Neyman & Pearson - Tests

Definition 7.1 *Es sei φ ein Hypothesentest. Dann wird* $\sup_{\vartheta\in\Theta_0} G_\varphi(\vartheta)$ *der* Umfang *des Tests φ genannt. Wird ein Wert $\alpha \in [0,1]$ vorgegeben und ist* $\sup_{\vartheta\in\Theta_0} G_\varphi(\vartheta) \leq \alpha$*, so heißt φ ein* Test zum Niveau α *(kurz* Niveau-α-Test*).*

Es sei $\mathcal{T}$ eine Menge von Tests für alle Niveaus α für die Hypothesen H_0: $\vartheta \in \theta_0$ versus H_1: $\vartheta \in \theta_1$. Gegeben sei ein Test $\varphi^* \in \mathcal{T}$ aus $\mathcal{T}$, für dessen Power-Funktion G^* die Bedingung

$$G^*(\vartheta) \geq G(\vartheta) \in \mathcal{T}) \tag{249}$$

gilt, wobei G die Power-Funktion eines Tests $\varphi \in \mathcal{T}$ ist. Dann heißt φ^* *gleichmäßig bester Test* (uniformly most powerful – UMP). Ist H_0 eine einfache Hypothese, d.h gilt H_0: $\vartheta = \vartheta_0$, so heißt der Test *bester Test zum Niveauα* (most powerful level α test).

Das Ziel ist, möglichst einen UMP-Test zu finden. Die Existenz eines solchen Tests kann, wie im folgenden Abschnitt gezeigt wird, für einfache Hypothesen stets gefunden werden.

7.1.1 Einfache Hypothesen

Es wird zunächst eine allgemeine Definition von Neyman-Pearson-Tests gegeben (Pruscha (2000), p. 222)

Definition 7.2 *Ein (randomisierter) Test φ heißt* Neyman-Pearson-Test (NP-Test) *für eine einfache Hypothese H_0 gegen eine einfache Alternativhypothese H_1, falls es eine Konstante κ gibt derart, dass*

$$\varphi(\boldsymbol{x}) \equiv \varphi^*(\boldsymbol{x}) = \begin{cases} 1, & \textit{falls } f_1(\boldsymbol{x}) > \kappa f_0(\boldsymbol{x}) \\ \gamma, & \textit{falls } f_1(\boldsymbol{x}) = \kappa f_0(\boldsymbol{x}), \\ 0, & \textit{falls } f_1(\boldsymbol{x}) < \kappa f_0(\boldsymbol{x}). \end{cases} \tag{250}$$

γ ist die Wahrscheinlichkeit, mit der nach dem Zufall für bzw. gegen H_0 entschieden wird.

Offenbar entspricht der Bedingung $\varphi_\kappa(\mathbf{x}) = f_1(\mathbf{x}) > \kappa f_0(\mathbf{x})$ die Aussage $\Lambda(\mathbf{x}) = f_1(\mathbf{x})/f_0(\mathbf{x}) > \kappa$, und analog für die anderen Aussagen, die also als Aussagen über den Likelihood-Quotienten gelesen werden können. Ebenso wie $\Lambda(\mathbf{x})$ ist $\varphi(\vartheta)$ eine

zufällige Veränderliche. Der Umfang eines Tests ergibt sich als Erwartungswert von φ, gegeben H_0:

$$\begin{aligned} \text{Umfang von } \varphi &= 1P(P(f_1 + \kappa f_0) + 0P(f_1 \leq \kappa f_0) + \gamma P(f_1 = \kappa f_0) \\ &= \mathbb{E}_0(\varphi_\kappa) = P(f_1 > \kappa f_0) + \gamma P(f_1 = \kappa f_0). \end{aligned} \tag{251}$$

Ein NP-Test φ_κ ist bester Test zum Niveau des eigenen Umfangs $\mathbb{E}_0(\varphi_\kappa)$:

Satz 7.1 *Es sei $\kappa \in [0, \infty)$ und $\gamma \in [0, 1]$ und φ_κ sei ein NP-Test der Art(250). Dann ist φ_κ bester Test zum Niveau α.*

Beweis: S. Pruscha (2000), p. 223.

Man hat dann das *Fundamentallemma* von Neyman & Pearson

Satz 7.2 *Es gelten die Aussagen*

1. *Zu vorgegebenem $\alpha \in (0, 1)$ existiert ein NP-Test φ_κ zum Umfang α,*
2. *Ist φ ebenfalls bester Test zum Niveau α, dann gilt für φ_κ aus 1 mit $D = \{\boldsymbol{x} | f_1(\boldsymbol{x}) \neq \kappa f_0(\boldsymbol{x})\}$*

$$\varphi(\boldsymbol{x}) = \varphi_\kappa(\boldsymbol{x}) \text{ für } \mu\text{-fast alle } \boldsymbol{x} \in D. \tag{252}$$

Beweis: S. Pruscha (2000), p. 224. Es genügt hier, eine Plausibilitätsbetrachtung durchzuführen, wie man sie etwa in Kendall & Stuart (1973) findet.

Das Ziel ist, einen *kritischen Bereich* (*critical region*) (Neyman et al., 1933) C zu finden derart, dass für vorgegenen Wert von α die Wahrscheinlichkeit der Akzeptanz von H_1, wenn H_1 korrekt ist, maximal ist. Die Wahrscheinlichkeit, dass H_1 akzeptiert wird, ist

$$1 - \beta = \int_C f_1(\mathbf{x}) d\mathbf{x} = \int_C L(\mathbf{x}|H_1) d\mathbf{x}. \tag{253}$$

f_1 ist die Dichte von $\mathbf{x}$, wenn H_1 gilt; diese Dichte ist gerade gleich der Likelihood $L(\mathbf{x}|H_1)$ von $\mathbf{x}$ unter der Bedingung H_1. Die Wahrscheinlichkeit α eines Fehlers erster Art ist

$$\alpha = \int_C f_o(\mathbf{x}) d\mathbf{x} = \int_C L(\mathbf{x}|H_0) d\mathbf{x}. \tag{254}$$

f_0 ist die Dichte von $\mathbf{x}$, wenn H_0 gilt, und ist gleich der Likelihood $L(\mathbf{x}|H_0)$ von $\mathbf{x}$. (253) kann sicher in der Form

$$1 - \beta = \int_C \frac{L(\mathbf{x}|H_1)}{L(\mathbf{x}|H_0)} L(\mathbf{x}|H_0) d\mathbf{x} \tag{255}$$

geschrieben werden. Der Quotient

$$\Lambda(\mathbf{x}) = \frac{L(\mathbf{x}|H_1)}{L(\mathbf{x}|H_0)} \tag{256}$$

ist aber gerade der Likelihoodquotient. Nach (255) ist $1 - \beta$ gerade der Erwartungswert $\mathbb{E}_C(\Lambda)$ von $\Lambda(\mathbf{x})$ auf C. Damit $1 - \beta$ maximal wird, muß $\Lambda(\mathbf{x})$ auf C

maximale Werte annehmen. Man kann diesen Sachverhalt so ausdrücken, dass man sagt, dass auf C die Ungleichung

$$\Lambda(\mathbf{x}) = \frac{L(\mathbf{x}|H_1)}{L(\mathbf{x}|H_0)} \geq \kappa_\alpha \tag{257}$$

gelten muß, wobei κ_α eine von α abhängige Konstante ist, d.h. es muß

$$L(\mathbf{x}|H_1) \geq \kappa_\alpha L(\mathbf{x}|H_0), \quad \mathbf{x} \in C \tag{258}$$

gelten. Die Abhängigkeit κ von α wird durch (254) hahegelegt, wo ja α, $L(\mathbf{x}|H_0)$ und C zueinander in Beziehung gesetzt werden. Die folgenden Betrachtungen machen diese Abhängigkeit explizit.

Das folgende Beispiel illustriert u. A. die Notwendigkeit der Randomisierung im Falle diskreter Variablen.

Beispiel 7.1 Es sei $f(x, \lambda) = \lambda^x \exp(-\lambda)/x!$, d.h. $x = 0, 1, 2, \ldots$. Betrachtet werden die Hypothesen

$$H_0\text{: } \lambda = \lambda_0, \quad H_1\text{: } \lambda = \lambda_1, \quad 0 < \lambda_0 < \lambda_1.$$

Dann ist

$$\Lambda(x) = \frac{f(x, \lambda_1)}{f_0(x, \lambda_0)} = \left(\frac{\lambda_1}{\lambda_0}\right)^x e^{-(\lambda_1 - \lambda_0)}, \text{ mit } x \in \mathbb{N}.$$

$\Lambda(x)$ wächst streng monoton mit x. Dann existiert eine Konstante κ mit $\Lambda(x) > \kappa$ und $\Lambda(x) = \kappa$ umformuliert werden können in $x > k$ bzw $x = k$ mit

$$\varphi_\kappa(x) = \begin{cases} 1, & x > k \\ \gamma, & x = k \\ 0, & x < k \end{cases}.$$

Für $\alpha \in (0, 1)$ ergeben sich die Konstanten k aus

$$P(x > k|H_0) + \gamma P(x = k) = \alpha,$$

mit

$$P(x > k|H_0) = 1 - e^{-\lambda_0} \sum_{i=0}^{k} \frac{\lambda_0^i}{i!}, \quad P(x = k|H_0) = e^{-\lambda_0} \frac{\lambda_0^k}{k!}.$$

Man wählt das kleinste α, das $P(x > k|H_0) \leq \alpha$ erfüllt und setzt $\gamma = (\alpha - P(x > k|H_0)/P(x = k|H_0))$. □

Beispiel 7.2 (Nach Kendall & Stuart, 1972, p. 175) Die Dichte von x sei durch die Gauß-Dichte gegeben, also

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x - \mu)^2}{2\sigma^2}\right]. \tag{259}$$

Betrachtet werden die Hypothesen

$$H_0\text{: } \mu = \mu_0, \text{ versus } H_1\text{: } \mu = \mu_1 \neq \mu_0$$

Gegeben sei die Stichprobe $\mathbf{x} = (x_1, \ldots, x_n)$; die x_i seien stochastisch unabhängig. Weiter sei $\sigma^2 = 1$. Die Likelihoods sind dann durch

$$\begin{aligned} L(\mathbf{x}|H_i) = \prod_{j=1}^{n} f_i(x_j) &= (2\pi)^{-n/2} \exp\left[-\frac{1}{2}\sum_{j=1}^{n}(x_j - \mu_i)^2\right] \\ &= (2\pi)^{-n/2} \exp\left[-\frac{n}{2}s^2 + (\bar{x} - \mu_i)^2\right], \quad i = 0, 1 \end{aligned} \tag{260}$$

denn, wie aus der Varianzanalyse bekannt,

$$\sum_{j=1}^{n}(x_j - \mu_i)^2 = \sum_{j=1}^{n}(x_j - \bar{x} + \bar{x} - \mu_i)^2 = \sum_{j=1}^{n}(x_j - \bar{x})^2 + \sum_{j=1}^{n}(\bar{x} - \mu)^2,$$

und mit $s^2 = \sum_{j=1}^{n}(x_j - \bar{x})^2/n$ hat man $\sum_{j=1}^{n}(x_j - \bar{x})^2 = ns^2$. Dann folgt, dass auf C die Ungleichung

$$\begin{aligned} \frac{L(\mathbf{x}|H_1)}{L(\mathbf{x}|H_0)} &= \exp\left[\frac{n}{2}((\bar{x} - \mu_1)^2 - (\bar{x} - \mu_0)^2)\right] \\ &= \exp\left[\frac{n}{2}((\mu_0 - \mu_1)2\bar{x} + (\mu_1^2 - \mu_0^2))\right] \geq \kappa_\alpha, \end{aligned} \tag{261}$$

gilt, woraus

$$\frac{n}{2}((\mu_0 - \mu_1)2\bar{x} + (\mu_1^2 - \mu_0^2)) \geq \log \kappa_\alpha$$

und also

$$(\mu_0 - \mu_1)\bar{x} \geq \frac{1}{2}(\mu_0^2 - \mu_1^2) + \frac{1}{n}\log \kappa_\alpha \tag{262}$$

folgt. Die Daten gehen hier nur noch über den Mittelwert $\bar{x}$ ein; offenbar ist die Statistik, über die die Entscheidung über H_0 und H_1 getroffen wird, durch $T(\mathbf{x}) = \bar{x}$ gegeben. Man hat dann, wegen $\mu_0^2 - \mu_1^2 = (\mu_0 + \mu_1)(\mu_0 - \mu_1)$, für $\mu_0 > \mu_1$

$$\bar{x} \geq \frac{1}{2}(\mu_0 + \mu_1) + \frac{\log \kappa_\alpha}{n(\mu_0 - \mu_1)} = K, \tag{263}$$

und für $\mu_0 < \mu_1$ ergibt sich[22]

$$\bar{x} \leq \frac{1}{2}(\mu_0 + \mu_1) - \frac{\log \kappa_\alpha}{n(\mu_1 - \mu_0)} = K, \tag{264}$$

Es fehlt noch die explizite Charakterisierung von C. Es muß $C = [\bar{x}_\alpha, \infty)$ sein, so dass $\bar{x}_\alpha$ bestimmt werden muß. Es wird der Fall $\mu_1 > \mu_0$ betrachtet. Da die x_j normalverteilt sind, ist auch $\bar{x}$ normalverteilt. Gilt $\mu = \mu_0$, so muß

$$\alpha = \int_K^\infty \sqrt{\frac{n}{2\pi}} \exp\left[-\frac{n}{2}(\bar{x} - \mu_0)^2\right] d\bar{x} \tag{265}$$

dies ist die Wahrscheinlichkeit, dass bei Gültigkeit von H_0 ein $\bar{x}$-Wert größer als K auftritt, H_0 also fälschlich zugunsten von H_1 verworfen wird. α soll einen bestimmten Wert haben, etwa $\alpha = .1$ oder $\alpha = .05$. Einem gegebenen α-Wert entspricht ein bestimmter z-Wert $z_\alpha = (\bar{x}_\alpha - \mu_0)/(1/\sqrt{n}) = \sqrt{n}(\bar{x}_\alpha - \mu_0)$, so dass

$$\alpha = 1 - \Phi(z_\alpha),$$

[22] Vorzeichen beim Umgang mit Ungleichungen etc

Φ die Standardnormalverteilung. Für $\alpha = .05$ hat man $z_\alpha = 1.645$, und $\bar{x}_\alpha = z_\alpha\sqrt{n} + \mu_0$. Für die *power* des Tests ergibt sich

$$1 - \beta = \int_K^\infty \sqrt{\frac{n}{2\pi}} \exp\left[-\frac{n}{2}(\bar{x} - \mu_1)^2\right] d\bar{x}. \tag{266}$$

Hier korrespondiert K zum z-Wert $\sqrt{n}(K - \mu_1)$. Substituiert man für K den Wert $z_\alpha\sqrt{n} + \mu_0$, so hat man nun

$$1 - \beta = \Phi[(n^{1/2}(\mu_1 - \mu_0) - z_\alpha].$$

Offenbar folgt $1 - \beta \to 1$ für $n \to \infty$, d.h. die Power steigt mit wachsendem Stichprobenumfang, ebenso mit größer werdender Differenz $\mu_1 - \mu_0$. Es wird insbesondere $\mu_1 > \mu_0$ getestet. Für $\mu_1 \to -\infty$ geht die Power gegen Null. □

Beispiel 7.3 (Nach Lindgren, 1962) Es werde wieder eine Hypothese bezüglich des Erwartungswertes einer Gauß-verteilten Variablen mit der Varianz $\sigma^2 = 1$ betrachtet: es sei H_0: $\mu = 0$ versus H_1: $\mu = 1$. Dann ist

$$f_0(\mathbf{x}) = L(\mathbf{x}|H_0) = \frac{1}{(8\pi)^{n/2}} \exp\left[-\frac{1}{8}\sum_j x_j^2\right] \tag{267}$$

$$f_1(\mathbf{x}) = L(\mathbf{x}|H_1) = \frac{1}{(8\pi)^{n/2}} \exp\left[-\frac{1}{8}\sum_j (x_j - 1)^2\right] \tag{268}$$

Der Likelihood-Quotient ist dann durch

$$\begin{aligned} \Lambda(\mathbf{x}) = \frac{L(\mathbf{x}|H_1)}{L(\mathbf{x}|H_0)} &= \exp\left[-\frac{1}{8}(\sum_j (x_j - 1)^2 - \sum_j x_j^2)\right] \\ &= \exp\left[\frac{n}{8}(2\bar{x} - 1)\right] \end{aligned} \tag{269}$$

gegeben. Aus $\Lambda(\mathbf{x}) \leq \lambda$ (Indikation für H_0) folgt dann

$$\log\lambda \leq \frac{n}{8}(1 - 2\bar{x}),$$

d.h. nach Auflösung dieser Ungleichung nach $\bar{x}$

$$\frac{1}{2} - \frac{4}{n}\log\lambda = K \leq \bar{x}. \tag{270}$$

Für $\bar{x} \leq K$ würde man sich also für H_0 entscheiden. Es ist dann

$$\alpha = P(\bar{x} > K|H_0) = 1 - P(\bar{x} \leq K|H_0) = 1 - \Phi[(K - 0)/(1/\sqrt{n}) = 1 - \Phi[\sqrt{n}K], \tag{271}$$

Φ die Standardnormalverteilung. Für gegebenen Wert von α findet man den zugehörigen z-Wert; etwa für $\alpha = .05$ hat man $z = 1.64$, so dass $1.64 = \sqrt{n}K$ oder $K = 1.64/\sqrt{n}$. Für größer werdenden Stichprobenumfang n wird K kleiner, d.h. $\bar{x}$ muß einen kleineren K-Wert unterschreiten, damit H_0 akzeptiert wird, bzw. einen kleineren K-Wert überschreiten, damit H_1 akzeptiert wird. □

Während also beim Fisherschen p-Wert die Wahrscheinlichkeit p, dass ein Wert größer als der beobachtete $\bar{x}$- oder $\Delta\bar{x}$-Wert gefunden wird, wird beim Neyman-Pearson-Test ein kritischer K-Wert bestimmt, der einerseits vom μ-Wert unter H_0 und H_1 abhängt, – diese Werte gehen in den λ-Wert ein (vergl. (270), und der andererseits vom gewählten α-Niveau abhängt, vergl. (271). Dieser K-Wert garantiert, dass H_1 mit maximaler Wahrscheinlichkeit akzeptiert wird, wenn H_1 korrekt ist. Der Bereich C ist durch $C = [K, \infty)$ definiert.

Die Frage ist, welche der Hypothesen $\mu = \mu_0$ und $\mu = \mu_1$ man mit der Nullhypothese identifizieren soll. Die Entscheidungen für oder gegen eine Hypothese können von dieser Wahl abhängen. Als allgemeine Strategie für die Wahl von H_0 und damit von H_1 kann man sagen, dass man mit H_1 diejenige Hypothese auszeichen sollte, die mit maximaler Wahrscheinlichkeit akzeptiert werden soll, wenn sie korrekt ist. Hir gehen eventuell allgemeine Kostenerwägungen ein, die im Rahmen der Statistik zunächst nicht entschieden werden können. Tests der betrachteten Art heißen *beste Tests zum Niveau α.*

Beste einseitige Tests: Man kann auch beste einseitige Tests betrachten, etwa

$$H_0\colon \vartheta \leq \vartheta, \quad H_1\colon \vartheta > \vartheta_0, \tag{272}$$

für vorgegebenes ϑ_0. Der Test hat die Form

$$\varphi^*_\kappa(\mathbf{x}) = \begin{cases} 1, & \Lambda(\mathbf{x}) > \kappa \\ \gamma, & \Lambda(\mathbf{x}) = \kappa, \\ 0, & \Lambda(\mathbf{x}) < \kappa. \end{cases} \tag{273}$$

Es läßt sich nun der folgende Satz beweisen (Pruscha (2000), p. 229)

Satz 7.3 *Der Test sei durch $\varphi^*_\kappa(\boldsymbol{x})$ in (273) definiert, und $\Lambda(\boldsymbol{x})$ sei monoton. Dann gilt (a) für $\vartheta_0 \in \Theta$ und $\alpha \equiv \mathbb{E}_0(\varphi^*_\kappa) > 0$ ist φ^*_κ bester Test zum Niveau α für H_0 versus H_1, (b) $\vartheta_0 \in \Theta$ und $\alpha \in (0,1)$ existiert ein $\kappa \in \mathbb{R}$ und $\gamma \in [0,1]$, so dass φ^*_κ ein Test zum Umfang α ist, (c) die Gütefunktion $G_{\varphi_{\kappa^*}}(\vartheta) = \mathbb{E}_\vartheta(\varphi_{\kappa^*})$ ist streng monoton wachsend.*

Beispiel 7.4 Einseitiger Gauß-Test Es sei $\mathbf{x} = \{x_1, \ldots, x_n\}$, $\bar{x} = \sum_j x_j/n$, und die x_j seien i.i.d. $N(\mu, \sigma^2)$-verteilt; σ^2 sei bekannt. Weiter sei

$$T(\mathbf{x}) = \frac{\sqrt{n}(\bar{x} - \mu_0)}{\sigma}. \tag{274}$$

Der Test

$$T(\mathbf{x}) = \begin{cases} 1, & T(\mathbf{x}) > \kappa^* \\ \gamma, & T(\mathbf{x}) = \kappa^* \\ 0, & T(\mathbf{x}) < \kappa^* \end{cases}$$

mit $\kappa^* = u_{1-\alpha}$ bester Test für $H_0\colon \mu \leq \mu_0$ versus $H_1\colon \mu > \mu_0$ zum Niveau α. □

7.1.2 Zusammengesetzte Hypothesen; Nuisance-Parameter

Die Hypothesen H_0 und H_1 sind *einfach*, wenn die jeweiligen Wahrscheinlichkeitsverteilungen vollständig spezifiziert sind, andernfalls sind sie *zusammengesetzt*

(*composite*). So ist im Beispiel 7.4 der Parameter σ^2 als bekannt vorausgesetzt worden. Im allgemeinen sind aber weder μ noch σ^2 bekannt. Will man Hypothesen bezüglich μ testen, so ist der unbekannte Parameter σ^2 ein *Störparameter* (*nuisance parameter*), die entsprechende Hypothese ist zusammengesetzt. Ist $H_0\colon\ \mu = \mu_0$, $H_1\colon\ \mu > \mu_0$, so ist H_1 auch im Falle bekannter Varianz zusammengesetzt, da ja die Verteilung unter H_1 nicht genau spezifiziert ist. Zu jedem bestimmten Wert $\mu_1 > \mu_0$ korrespondiert eine spezielle einfache Hypothese, zhu $\mu > \mu_0$ korrespondiert also eine Menge einfacher Hypothesen. Ist für jede spezielle Hypothese der Test *most powerful*, so ist er in Bezug auf die zusammengesetzte Hypothese $\mu > \mu_1$ der *gleichförmig beste Test* (*uniform most powerful* (UMP)).

Will man im Rahmen der Neyman-Pearson-Theorie den kritischen Bereich für eine zusammengesetzte Hypothese bestimmen, so ist das im Neyman-Pearson-Lemma beschriebene Verfahren nicht anwendbar, weil ja der Parameter ϑ_1 für eine spezifische Hypothese nicht festgelegt ist. Man kann nun versuchen, den kritischen Bereich dennoch auf einen minimalen Bereich festzulegen, $\alpha(\vartheta_1) \leq \alpha$. Gilt insbesondere $\alpha(\vartheta_1) = \alpha$, so heißt der kritische Bereich dem gesamten Stichprobenraum *ähnlich*.

Hypothesen der Form $H_0\colon\ \mu = \mu_0$ und $H_1\colon\ \mu = \mu_1$ stehen im Allgemeinen nicht zur Diskussion; oft hat man zwar eine bestimmte Aussage, die mit H_0 assoziiert werden kann (zwei Therapien oder allgemein Experimentalbedingungen unterscheiden sich nicht), aber die Alternativhypothese ist nicht weiter spezifiziert, – die Bedingungen unterscheiden sich, aber man weiß nicht, um wie viel. H_1 ist dann eine zusammengesetzte Hypothese. Auch H_0 kann zusammengesetzt sein: etwa $H_0\colon\ \vartheta < \vartheta_0$ versus $H_1\colon\ \vartheta \geq \vartheta_0$. Außerdem: es ist möglich, dass nicht nur ein Parameter unbekannt ist, sondern mehr als einer: nicht nur μ, sondern auch σ^2. Die Frage ist, wie nun ein Test zu entwickeln ist.

Es sei ϑ die Menge der möglichen Werte, die der in Frage stehende Parameter ϑ annehmen kann. Insbesondere sei ϑ_0 ein vorgegebener Wert, und es werde das Testproblem

$$H_0\colon\ \vartheta = \vartheta_0 \ \text{ versus } \ \vartheta \neq \vartheta_0, \ \ \vartheta \in \vartheta \subseteq \mathbb{R} \tag{275}$$

betrachtet. Für diese Fragestellung kann es keinen besten Testzum Niveau α geben(vergl. Pruscha (2000), p. 231). Der Begriff des besten Tests kann aber eingeschränkt werden auf den des *unverfälschten Tests*.

7.1.3 Stichprobenumfang und Effektgröße

Ist H_1 nicht spezifiziert, kann der β-Fehler nicht berechnet werden. Einen Ausweg aus dieser Schwierigkeit liefert der Begriff der Effektgröße. Ist etwa der Erwartungswert der Statistik unter H_0 gleich μ_0 und gleich μ_1 unter H_1 und ist σ die Streuung[23], so ist

$$\varepsilon = \frac{\mu_1 - \mu_0}{\sigma} \tag{276}$$

die *Effektgröße*. Man kann sich nun "kleine" und "große" Effektstärken ansehen und entscheiden, für welche Werte von ε sich überhaupt eine Untersuchung lohnt, zumal man für eine hinreichend große Stichprobe stets einen signifikanten p-Wert erhalten kann. Insbesondere kann man für einen vorgegebenen ε-Wert die Stichprobengröße berechnen, die benötigt wird, um für vorgegebenen α- und β-Wert eine Signifikanz zu erhalten, wenn H_1 gilt.

[23] Geschätzt als *pooled estimate*.

Es sei nun $T(\mathbf{x}) = \bar{x}$ bei einem Stichprobenumfang von n, und s^2 sei die Stichprobenvarianz. Insbesondere sei $\bar{x}_c$ ein kritischer T-Wert: $P(\bar{x} \leq \bar{x}_c|H_0) = 1 - \alpha$, $P(\bar{x} \leq \bar{x}_c|H_1) = \beta$. Die Varianz der Mittelwerte ist $s_{\bar{x}}^2 = s^2/n$, und $\bar{x}$ korrespondiert unter H_0 bzw. H_1 zu den z-Werten

$$z_{1-\alpha} = \frac{\bar{x}_c - \mu_0}{s_{\bar{x}}}, \quad z_\beta = \frac{\bar{x}_c - \mu_1}{s_{\bar{x}}}.$$

Hieraus folgt

$$\bar{x}_c = s_{\bar{x}} z_{1-\alpha} + \mu_0, \quad \bar{x}_c = s_{\bar{x}} z_\beta + \mu_1,$$

und

$$s_{\bar{x}} z_\beta + \mu_1 - s_{\bar{x}} z_{1-\alpha} - \mu_0 = 0,$$

so dass

$$z_{1-\alpha} - z_\beta = \frac{\sqrt{n}(\mu_1 - \mu_0)}{s_{\bar{x}}} = \hat{\varepsilon}\sqrt{n}, \tag{277}$$

und schließlich

$$n = \frac{(z_{1-\alpha} - z_\beta)^2}{\hat{\varepsilon}^2} \tag{278}$$

folgt. Für vorgebenen α- und β-Werte und angenommenen ε-Wert erhält man also den gewünschten n-Wert.

Der Punkt ist, dass man über einen angenommenen ε-Wert gewissermaßen die Bedingungen des Neyman-Pearson-Lemmas wieder herstellt.

power calculation in R: $http://www.ats.ucla.edu/stat/R/dae/t_test_power2.htm$
kleines t und p zeigen vorangehenden unterstrich an

applet: http://www.cs.uiowa.edu/ rlenth/Power/

See Hung, O'Neill, Bauer, Köhne 1997 Behavior of p-value when h1 ist true — for <4p<4-value as a random variable. Further Literature

Weitere Beispiele für die ε-Technik.

7.2 Score-Funktion und Fisher-Information

Es sei X eine zufällige Veränderliche, deren Verteilung von einem Parameter θ abhänge. Weiter sei $L(\theta; X)$ die Likelihood-Funktion. Die *Score-Funktion* $V(\theta)$ ist die erste Ableitung von $\log L(\theta; X)$, also der Gradient des Logarithmus der Likelihood-Funktion; nach der Kettenregel erhält man

$$V(\theta) = \frac{\partial \log L(\theta; X)}{\partial \theta} = \frac{1}{L(\theta; X)} \frac{\partial L(\theta; X)}{\partial \theta}. \tag{279}$$

DA $\log L$ eine monotone Funktion von L ist, nimmt $\log L$ ein Maximum an, wenn L ein Maximum annimmt, was natürlich mit (279) übereinstimmt: Der Gradient V ist gleich Null, wenn

$$\partial L/\partial\theta|_{\theta=\hat{\theta}} = 0,$$

wenn also $\hat{\theta}$ die Maximum-Likelihood-Schätzung von θ ist.

Da X zufällig ist, ist auch V zufällig. Es gilt

$$\mathbb{E}(V|\theta) = 0. \tag{280}$$

Denn

$$\begin{aligned}\mathbb{E}(V|\theta) &= \int_{-\infty}^{\infty} \frac{\partial \log L(\theta;X)}{\partial\theta} \\ &= \int_{-\infty}^{\infty} \frac{1}{L(\theta;X)} \frac{\partial L(\theta;X)}{\partial\theta} f(X|\theta)\, dX \\ &= \frac{\partial L}{\partial\theta} \int_{-\infty}^{\infty} f(X|\theta)\, dX = \frac{\partial 1}{\partial\theta} = 0, \end{aligned} \tag{281}$$

(unter der Voraussetzung der Vertauschbarkeit von Integration und Differentiation). Für die Varianz $\mathbb{V}(V|\theta)$ findet man, wegen $\mathbb{V}(V) = \mathbb{E}(V^2) - \mathbb{E}^2(V)$ und $\mathbb{E}(V) = 0$

$$\mathbb{V}(V|\theta) = \mathbb{E}\left[\left(\frac{\partial \log L}{\partial\theta}\right)^2\right] = \mathcal{I}(\theta). \tag{282}$$

$\mathcal{I}(\theta)$, also die Varianz von V, ist die *Fisher-Information*. Bevor der Begriff der (Fisher-) Information an Beispielen erläutert wird, wird noch der folgende Satz bewiesen:

Satz 7.4 *Es gilt*

$$\mathcal{I}(\theta) = \mathbb{E}\left[\frac{\partial^2 \log L}{\partial\theta^2}\right]. \tag{283}$$

Beweis: Aus der Definition von L folgt

$$\int L(\theta;\mathbf{x}) d\mathbf{x} = 1 \quad \text{für alle } \theta$$

und da $\partial 1/\partial\theta = 0$ folgt weiter

$$\int \frac{\partial L}{\partial\theta} d\mathbf{x} = \int \frac{\partial^2 L}{\partial\theta^2} = 0. \tag{284}$$

Dann hat man

$$\int \frac{\partial L}{\partial\theta} d\mathbf{x} = \int \frac{1}{L}\frac{\partial L}{\partial\theta} L\, d\mathbf{x} = \int \frac{\partial \log L}{\partial\theta} L\, d\mathbf{x} = \mathbb{E}\left[\frac{\partial \log L}{\partial\theta}\right] = 0;$$

dies ist eine alternative Ableitung von (280) und (281). Nun ist

$$\frac{\partial^2 \log L}{\partial\theta^2} = \frac{\partial}{\partial\theta}\left(\frac{1}{L}\frac{\partial L}{\partial\theta}\right) = -\frac{1}{L^2}\left(\frac{\partial L}{\partial\theta}\right)^2 + \frac{1}{L}\frac{\partial^2 L}{\partial\theta^2}$$

und

$$\mathbb{E}\left[\frac{\partial^2 \log L}{\partial\theta^2}\right] = -\int \frac{1}{L^2}\left(\frac{\partial L}{\partial\theta}\right)^2 L\, d\mathbf{x} + \int \frac{1}{L}\frac{\partial^2 L}{\partial\theta^2} L\, d\mathbf{x}.$$

Das Integral rechts ist aber wegen (284) gleich Null (L kürzt sich heraus), und das erste Integral rechts ist gerade gleich $\mathbb{E}[(\partial L/\partial\theta)^2]$. Damit ist (283) wegen (282) gezeigt. □

Beispiel 7.5 X sei binomialverteilt, so dass

$$P(X = k|\theta, n) = \binom{n}{k}\theta^k(1-\theta)^{n-k}.$$

Sind k "Erfolge" gegeben, so ist gleichzeitig

$$L(\theta|k) = P(X = k|\theta, n),$$

und

$$\log L = k \log \theta + (n - k) \log(1 - \theta),$$

so dass

$$V(\theta) = \frac{k}{\theta} - \frac{n-k}{1-\theta}.$$

Wegen $\mathbb{E}(k) = n\theta$ hat man

$$\mathbb{E}(V) = \frac{n\theta}{\theta} - \frac{n}{1-\theta} + \frac{n\theta}{1-\theta} = \frac{n}{1-\theta} - \frac{n}{1-.\theta} = 0.$$

Dann ist

$$= \mathbb{V}(V) = \mathbb{V}\left[\frac{k - n\theta}{\theta(1-\theta)}\right] = \frac{n}{\theta(1-\theta)} = \mathcal{I}(\theta).$$

Offenbar ist in diesem Fall

$$\mathcal{I}(\theta) = \frac{1}{\mathbb{V}(X)},$$

d.h. die Fisher-Information ist der Reziprokwert der Varianz von X. Je kleiner die Varianz von X, desto größer die Information über θ, – ein unmittelbar einleuchtendes Resultat, wenn man an die Bedeutung des Ausdrucks 'Information' denkt. □

Beispiel 7.6 Es sei $X \sim N(\mu, \sigma^2)$, σ^2 sei bekannt, und $\theta = \mu$. Dann ist die Likelihood durch

$$L(\mu) = \frac{1}{\sigma^n (2\pi^{n/2}} \exp\left[-\frac{1}{2}\sum_i (x_i - \mu)^2\right]$$

gegeben, und die Score-Funktion durch

$$V = \frac{\partial}{\partial \mu}\left[\frac{1}{2\sigma^2}\sum_i (x_i - \mu)^2 - \log(\sigma^n (2\pi)^{n/2}\right] = -\frac{1}{\sigma^2}\sum_i (x_i - \mu).$$

Dass $\mathbb{E}(V) = 0$, ist sofort einsichtig. Die Fisher-Information ist dann durch

$$\mathcal{I}(\theta) = \mathbb{V}(V) = \mathbb{E}\left[\left(\frac{\partial \log L}{\partial \mu}\right)^2\right] = \mathbb{E}\left[\frac{\partial^2 \log L}{\partial \mu^2}\right]$$

gegeben. Es ist

$$\mathbb{E}\left[\left(\frac{\partial \log L}{\partial \mu}\right)^2\right] = \frac{1}{\sigma^4}\mathbb{E}\left[\left(\sum_i (x_i - \mu)\right)^2\right],$$

und

$$\left(\sum_i (x_i - \mu)\right)^2 = \sum_i (x_i - \mu)^2,$$

wegen der Unshängigkeit der x_i, und mithin

$$\frac{1}{\sigma^4}\mathbb{E}\left[\left(\sum_i (x_i-\mu)\right)^2\right] = \frac{1}{\sigma^4}\mathbb{E}\left[\sum_i (x_i-\mu)^2\right] = \frac{1}{\sigma^4}\sum_i \mathbb{E}(x_i-\mu)^2 = \frac{n\sigma^2}{\sigma^4},$$

d.h.

$$\mathcal{I}(\theta) = \frac{n}{\sigma^2}. \tag{285}$$

Ebenso folgt

$$\mathbb{E}\left[\frac{\partial^2 \log L}{\partial \theta^2}\right] = \mathbb{E}\left[\frac{\partial}{\partial \mu}\left(-\frac{1}{\sigma^2}\sum_i x_i + \frac{n\mu}{\sigma^2}\right)\right] = \mathbb{E}\left[\frac{n}{\sigma^2}\right] = \frac{n}{\sigma^2},$$

also natürlich das gleiche Resultat wie (285). Die Information in der Stichprobe bezüglich des Parameters μ ist proportional zum Stichprobenumfang n und umgekehrt proportional zur Varianz σ^2 der Daten. □

7.3 Kullback-Leibler-Distanz

Die Kullback-Leibler-Distanz – auch Kullback-Leible-Divergenz genannt – wurde von Kullback & Leibler (1951) eingeführt. Diese Distanz ist ein Maß für die Verschiedenheit zweier Wahrscheinlichkeitsverteilungen auf der Basis des Informations- bzw. Entropiebegriffs. Information entspricht der Ungewißheit, die beseitigt wird, wenn ein bestimmtes zufälliges Ereignis aus der Menge $\{A_1, A_2, \ldots, A_k\}$ mit den entsprechenden $p = (p_1, \ldots, p_k)$ eingetreten ist. Shannon führte dafür das Maß

$$H_p = -\sum_{j=1}^{j} p_j \log p_j, \quad p_j > 0 \tag{286}$$

ein. Für stetige Variable X mit der Dichte f läßt sich diese Definition erweitern zu

$$H_f = -\int f(x) \log f(x) dx. \tag{287}$$

Formal entspricht die Definition von H dem Erwartungswert von $\log p$ bzw. $\log f$ bezüglich der Verteilung $p_1, \ldots, p_k$ bzw. f, d.h. H repräsentiert die durchschnittliche Ungewißheit , die durch ein Experiment beseitigt wird und damit den durchschnittlichen Informationsgewinn. Es sei $q = (q_1, \ldots, q_k)$ eine zweite Verteilung. $H_p - H_q$ ist die Differenz zwischen den Informationsgewinnen. Kullback & Leibler betrachteten stattdessen aber den Abstand bzw. die "Divergenz"

$$I(p,q) = \sum_{j=1}^{n} p_j \log \frac{p_j}{q_j}; \tag{288}$$

dies ist die *Kullback-Leibler-Distanz* (KL-Distanz) oder die *Kullback-Leibler-Divergenz* . Sie ist der Erwartungwert von $\log p_j/q_j$ bezüglich der Verteilung p. Hat man allgemein zwei Verteilungen (Dichten) f und g für eine zufällige Veränderliche X, so kann man die KL-Distanz auch durch

$$I(f;g) = \mathbb{E}\left[\log \frac{f(X)}{g(X)}\right] \tag{289}$$

definieren, d.h. eben als Erwartungwsert von $\log f(X)/g(X)$ bezüglich der Verteilung f.

Allgemein gilt für irgendzwei Verteilungen/Dichten f und g (i) $I(f,g) \geq 0$, (ii) $I(f,g) = 0$ nur dann, wenn $f \equiv g$. Deswegen ist $I(f,g)$ keine Distanz im üblichen Sinne, da die Reflexivitätsforderung für eine Metrik mit der Distanz I nicht erfüllt ist, wie man sich leicht anhand von

$$\sum_{j=1}^{n} p_j \log \frac{p_j}{q_j} \neq \sum_{j=1}^{n} q_j \log \frac{q_j}{p_j} \text{ bzw. } \mathbb{E}\left[\frac{f(X)}{g(X)}\right] \neq \mathbb{E}\left[\frac{g(X)}{f(X)}\right]$$

für $p \neq q$ bzw. $f \neq g$, klarmacht.

Beispiel 7.7 Gegeben seien zwei Binomialverteilungen f und g mit den Parametern θ_1 und θ_2, so dass

$$P(X = k|\theta_i) = \binom{n}{k} \theta_i^k (1-\theta_i)^{n-k}, \;\; i = 1, 2.$$

Dann ist

$$I(f,g) = \sum_{k=1}^{n} P(X = k|\theta_1) \log \frac{P(X = k|\theta_1)}{P(X = k|\theta_2)},$$

d.h.

$$I(f,g) = \sum_{k=1}^{n} \binom{n}{k} \theta_1^k (1-\theta_1)^{n-k} \log \frac{\theta_1^k (1-\theta_1)^{n-k}}{\theta_2^k (1-\theta_2)^{n-k}}$$

Nun ist

$$\begin{aligned} \log \frac{\theta_1^k (1-\theta_1)^{n-k}}{\theta_2^k (1-\theta_2)^{n-k}} &= k \log \theta_1 + (n-k) \log(1-\theta_1) - k \log \theta_2 - (n-k) \log(1-\theta_2) \\ &= k \log \frac{\theta_1}{\theta_2} + (n-k) \log \frac{1-\theta_1}{1-\theta_2}. \end{aligned}$$

Wegen

$$\sum_{k=1}^{n} \binom{n}{k} \theta_1^k (1-\theta_1)^{n-k} = 1, \quad \sum_{k=1}^{n} k \binom{n}{k} \theta_1^k (1-\theta_1)^{n-k} = n\theta_1$$

erhält man

$$I(f,g) = n \left(\theta_1 \log \frac{\theta_1}{\theta_2} + (1-\theta_1) \log \frac{1-\theta_1}{1-\theta_2} \right). \tag{290}$$

□

Beispiel 7.8 Gegeben seien zwei Normalverteilungen $f_1 = N(\mu_1, \sigma_1^2)$ und $f_2 = N(\mu_2, \sigma_2^2)$; gesucht ist die KL-Distanz zwischen f_1 und f_2. Man hat

$$\begin{aligned} I(f_1, f_2) &= \int f_1(x) \log \frac{f_1(x)}{f_2(x)} dx \\ &= \frac{1}{\sigma_1 \sqrt{2\pi}} \int \exp\left(-\frac{(x-\mu_1)^2}{2\sigma_1^2} \right) \log \left(\frac{\sigma_2}{\sigma_1} \frac{\exp\left(-(x-\mu_1)^2/2\sigma_1^2\right)}{\exp(-(x-\mu_2)^2/2\sigma_2^2)} \right) dx \\ &= \frac{1}{\sigma_1 \sqrt{2\pi}} \int \exp\left(-\frac{(x-\mu_1)^2}{2\sigma_1^2} \right) \left(\log \frac{\sigma_2}{\sigma_1} + \frac{1}{2} \left(\frac{(x-\mu_2)^2}{\sigma_2^2} - \frac{(x-\mu_1)^2}{\sigma_1^2} \right) \right) dx \\ &= \frac{1}{2} \left(\log \frac{\sigma_2^2}{\sigma_1^2} + \frac{\sigma_1^2}{\sigma_2^2} + \frac{(\mu_1-\mu_2)^2}{\sigma_2^2} - 1 \right), \end{aligned} \tag{291}$$

wie man durch Nachrechnen leicht bestätigt (man beachte, dass $\int x^2 f_1(x)dx = \sigma_1^2 + \mu_1^2$, etc.).

Es sei insbesondere $\sigma_1 = \sigma_2 = \sigma$. Für diesen Fall vereinfacht sich $I(f,g)$ zu

$$I(f,g) = \frac{1}{2}\frac{(\mu_1 - \mu_2)^2}{\sigma^2} = I(g,f), \tag{292}$$

d.h. $2\sqrt{I(f,g)}$ ist gerade 2-mal die Wurzel aus der Effektgröße, die im Allgemeinen Linearen Modell eine zentrale Rolle spielt.

Im allgemeinen Fall findet man für 2 d-dimensionale Gauss-Dichten mit den Varianz-Kovarianz-Matrizen Σ_f und Σ_g und Erwartungswertvektoren μ_f, μ_g

$$I(f,g) = \frac{1}{2}\left[\log\frac{|\Sigma_g|}{|\Sigma_g|} + sp(\Sigma_g^{-1}\Sigma_f](\mu_f - \mu_g)^T\Sigma_g^{-1}(\mu_f - \mu_g) - d\right]. \tag{293}$$

(291) ist natürlich ein Spezialfall dieses allgemeinen Ausdrucks. Für zwei beliebige Dichten ist $I(f,g)$ nicht notwendig in geschlossener Form darstellbar. □

7.4 Wein und Wasser

Eine mögliche Auflösung des Paradoxons: Es könnte aber so sein, dass das Indifferenzprinzips nur auf bestimmte Größen, aber nicht auf alle anwendbar ist; problematisch sind Größen, die aus anderen in bestimmter Weise zusammengesetzt sind. Ist X z. B. ein Quotient, über dessen Wert man nichts weiß, so weiß man natürlich nichts über die möglichen Werte von $Y = 1/X$. Aber die Annahme der Gleichverteilung für X impliziert, dass Y nicht gleichverteilt sein kann, und *vice versa*. Gibt es kein Argument, dass etwa auf X zu fokussieren sei, so ist die Wahl von X oder $Y = 1/X$ beliebig. Daraus folgt noch nicht, dass das Indifferenzprinzip für das Mischungsverhältnis grundsätzlich nicht gilt. Denn die Mischung muß nicht als Verhältnis dargestellt werden; man kann sie auch als Summe darstellen. Die Gesamtmenge der Flüssigkeit wird ja als konstant angenommen. Dann kann man sagen, dass die Gesamtmenge durch n Teilchen (Moleküle) gebildet wird, etwa n_1 zum Wein gehörende Moleküle[24] und n_2 Wassermoleküle bzw. zum Wasser gehörende Moleküle. Also hat man $n = n_1 + n_2$ Teilchen insgesamt. Um die Beziehung zum Quotienten $X = n_1/n_2$ herzustellen, ist es nützlich, zu Anteilen überzugehen, womit auch die Frage nach der Summation von Äpfeln und Birnen umgangen wird. Also hat man

$$b = \frac{n_1}{n}, \quad 1 - b = \frac{n_2}{n}, \quad X = \frac{n_1}{n_2}. \tag{294}$$

Indifferenz bezüglich der Mischung ist nun Indifferenz bezüglich b; nimmt man eine Gleichverteilung von b an, so ist auch $1 - b$ gleichverteilt (Symmetrie).

Um die Auflösung des Wein/Wasser-Paradoxons zu diskutieren, müssen die Intervallgrenzen für b sowie die Beziehung zwischen b und dem Quotienten X hergestellt werden. Aus (294) folgt

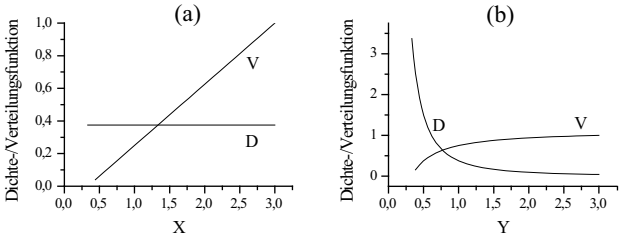
$$\frac{b}{n_2} = \frac{n_1}{n_2 n} = \frac{X}{n}, \text{ d.h. } \frac{bn}{n_2} = \frac{b}{1-b} = X \text{ und } b = \frac{X}{1+X}. \tag{295}$$

Für $X = 1/3$ folgt dann $b = 1/4$, und für $X = 3$ erhält man $b = 3/4$, also

$$\frac{1}{4} \leq b \leq \frac{3}{4}. \tag{296}$$

[24] Man muß ja alle Substanzen, die im Wein sind, zählen.

Abbildung 16: (a) Dichte-(D-) und Verteilungsfunktion (V) einer auf $(1/3, 3)$ gleichverteilten zufälligen Veränderlichen X, (b) Dichte und Verteilungsfunktion der korrespondierenden Veränderlichen $Y = 1/X$.



Auf b angewendet bedeutet das Indifferenzprinzip, dass b auf $[1/4, 3/4]$ gleichverteilt ist. Es werden, wie im Wein/Wasser-Paradoxon, die Fälle $X \leq 2$ und $Y = 1/X \geq 1/2$ betrachtet. Aus (295) erhält man

$$X \leq 2 \Rightarrow \frac{b}{1-b} \leq 2 \Rightarrow b \leq \frac{2}{3},$$

und

$$Y \geq \frac{1}{2} \Rightarrow \frac{1}{X} = \frac{1-b}{b} \geq \frac{1}{2} \Rightarrow b \leq \frac{2}{3},$$

d.h. in Bezug auf b haben $X \leq 2$ und $Y \geq 1/2$ die gleichen Konsequenzen, d.h. die Fragestellung nach der Mischung ist nun symmetrisch. Man erhält, wenn man von einer Gleichverteilung für b ausgeht,

$$P(X \leq 2) = P(Y \geq 1/2) = P(b \leq 2/3) = \frac{2/3 - 1/4}{3/4 - 1/4} = \frac{5}{6}. \tag{297}$$

Der wesentliche Unterschied zur Betrachtung des Verhältnisses $X = n_1/n_2$ ist, dass die Frage nach der Indifferenz bezüglich b symmetrisch zu der nach der Indifferenz bezüglich $1-b$ ist. Dass das Volumen in insgesamt n Teile aufgeteilt gedacht wurde, macht in Bezug auf Flüssigkeiten sicherlich Sinn, stellt aber darüber hinaus keine Einschränkung für einen allgemeineren Fall dar, d.h. b darf auch irrationale Werte annehmen.

Einen im Prinzip identischen Ansatz hat Mikkelson (2004) gemacht. Burock (2005) zeigt, dass der hier präsentierte Ansatz ein Spezialfall einer Klasse von Auflösungen des Wein/Wasser-Paradoxones ist.

Ist etwa X gleichverteilt auf dem Intervall $[a, b]$, so gilt für die Dichte $f_X(x) = 1/(b-a)$ und die Verteilungsfunktion ist durch

$$P(X \leq x) = \int_a^x \frac{dx}{b-a} = \frac{x-a}{b-a} \tag{298}$$

gegeben. Es folgt sofort

$$P(X > x) = 1 - \frac{x-a}{b-a} = \frac{b-x}{b-a} \tag{299}$$

Nun sei $Y = 1/X$; für die Verteilungsfunktion von Y erhält man

$$P(Y \leq y) = P(1/X \leq y) = P(1/y \leq X),$$

und wegen (299) hat man

$$P(Y \leq y) = P(X \geq 1/y) = \frac{b - 1/y}{b - a}. \tag{300}$$

Literatur

[1] Aitkin, M.(1986) Statistical modelling: the likelihood approach. *The Statistician* 35, 103–113

[2] Akaike, H. (1974) A new look at the statistical model identification. *IEEE Transactionss on Automatic Control*, 19(6), 716 – 723

[3] Barnard, G. A. (1947) The meaning of the significance level. *Biometrika*, 34 (1/2), 179–182

[4] Barnard, G. A. (1949) Statistical inference. *Journal of the Royal Statistical Society* 11(29) 115 – 149

[5] Barnard, G. A. (1967) The use of the likelihood function in statistical practise. *Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability*. Vol. 1, 27 –40

[6] Basu, D.(1975) Statistical Information and Likelihood [with Discussion], *Sankhya: The Indian Journal of Statistics, Series A* , 37 (1), 1– 71

[7] Bayarri, J. M., Berger, J.O. (2000) P values for composite null models. *Journal of the Amerian Statistical Association*, 95, 1127– 1142

[8] Bayarri, M. J., Berger, J. O.(2004) The Interplay of Bayesian and Frequentist Analysis, *Statistical Science*, 19 (1), 58 – 80

[9] Berger, J. O. (2003) Could Fisher, Jeffreys and Neyman have agreed on testing? *Statistical Science*, 18 (1), 1 – 32

[10] Berger, J.O., Sellke, T. (1987) Testing a point null hypothesis: the irreconcilability of p values and evidence. *Journal of the American Statistical Assiciation* 82, No. 397, Theory and Methods, 112 – 122

[11] Berger, J., Delampady, M. (1987) Testing precise hypotheses. *Statistical Science*, 2 (3), 317 – 335

[12] Berger, J.O. (2006) The Case for Objective Bayesian Analysis. *Bayesian Analysis*, 1 (3), 385– 402

[13] Bernardo, J.M. (1979) Reference Posterior Distributions for Bayesian Inference *Journal of the Royal Statistical Society. Series B (Methodological)*, 41(2), 113 – 147

[14] Bernardo, J. M., Rueda, R. (2002) Bayesian Hypothesis Testing: a Reference Approach *International Statistical Review*, (2002), 70, 3, 351 – 372

[15] Birnbaum, A. (1962) On the foundations of statistical inference. *Journal of the American Statistical Association*, 67, 269 – 306

[16] Birnbaum, A. (1977) The Neyman-Pearson theory as decision theory, and as inference theory; with a criticism of the Lindley-Savage argument for Bayesian theory. *Synthese*, 36, 19 – 49

[17] Birnbaum, M. H. (1983) Base rates in Bayesian inference:Signal detection analysis of the cab problem. *American Journal of Psychology*, 96 (1), 85 – 94

[18] Blackwell, D., Dubins, L. (1962) Merging of opinions with increasing information. *Annals of Mathematical Statistics*, 33, 882 – 886

[19] Blyth, C. R., Staudte, R. G. (1995) Estimating statistical hypotheses. *Statistics & Probability Letters*, 23, 45 – 52

[20] Box, G. E.P., Tiao, G.C.: Bayesian inference in statistical analysis. Addison-Wesley Publication Company, Reading (Mass.) 1973

[21] Carnap,R.: Logical Foundations of Probability. Chicago: UNiversity of Chicago Press 1950

[22] Casella, G., Berger, R. L. (1987) Reconciling Bayesian and Frequentist Evidence in the One-Sided Testing Problem. *Journal of the American Statistical Association*, 82, 106 – 111

[23] Christensen, R. (2005) Testing Fisher, Neyman, Pearson, and Bayes. *The American Statistician*, 59 (2), 121 –126

[24] Cohen, J.: Statistical power analysis for the bahavioral sciences. New York 1969

[25] Cornfield, J.(1966) Sequential trials, sequential analysis, and the likelihood-principle. *The American Statistician*, 20 (2), 18 – 23

[26] Cox, D.R. (1958) Some problems connected with statistical inference. *The Annals of Mathematical Statistics*, 29 (2), 357 – 372

[27] DeGroot, M. H. (1973) Doing what comes naturally: Interpreting a tail area as aposterior proability or as a likelihood ratio. *Journal of the American Statistical Association*, 68 (344), 966 – 969

[28] Dickey, J. M. (1977) Is the Tail Area Useful as an Approximate Bayes Factor? *Journal of the American Statistical Association*, 72, 138 – 142

[29] Dorling, J. (1979) Bayesian personalism, the methodology of scientific research programmes, and Duhem's problem. *Studies in History and Philosophy of Science Part A*, 10 (3), 177 – 187

[30] Durbin, J.(1979) On Birnbaum's Theorem on the Relation Between Sufficiency, Conditionality and Likelihood. *Journal of the American Statistical Association*, 65, 395 – 398

[31] Earman, J.: Bayes or bust? A critical examination of Bayesian confirmation theory. The MIT Press, Cambridge Mass., London 1992

[32] Edwards, w., Lindemann, H., Savage, 'L.J. (1963) Bayesian statistical inference in psychological research. *Psychological Research*, 70 (3), 193 – 242

[33] Fienberg, S. E. (2006) When did Bayesian inference become "Bayesian"? *Bayesian Analysis*, 1 (1), 1 – 40

[34] Fisher, R.A. (1922) On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London, Series A* , 222, 309 — 369

[35] R. A. Fisher: Statistical Methods for Research Workers, Edinburgh: Oliver and Boyd 1925

[36] Fisher, R. A.: Statistical methods and scientific inference. New York 1973

[37] Fitelson, B. (1999) The plurality of Bayesian measures of confirmation and the problem of measure sensitivity. *Philosophy of Science*, 66, 362 – 378

[38] Forster, M. R. (1995) Review: Bayes and Bust: Simplicity as a prolem for a probabilist's approach to confirmation. *The British Journal for the Philosophy of Science*, 46 (3), 399 – 424

[39] Fraser, D.A.S. (1963) On the sufficiency and likelihood principles. *Journal of the American Statistical Association*, 58 (3), 641 – 647

[40] Gaifman, H., Snir, M. (1982) Probabilities over rich languages. *Journal of Symbolic Logic*, 47, 495 – 548

[41] Giere, R.N.(1977) Allan Birnbaum's conception of statistical evidence. *Synthese*, 36, 5 – 13

[42] Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J., , and Krüger, L.: The Empire of Chance: How Probability Changed Everyday Life. Cambridge University Press, 1989

[43] Gigerenzer, G.: Über den mechanischen Umgang mit statistischen Methoden. In: Roth, E., Holling, H.(eds) Sozialwissenschaftliche Methoden. Lehr- und Handbuch für Forschung und Praxis. Oldenbourg Verlag, München 1993a

[44] Gigerenzer, G.: The superego, the ego, and the id in statistical reasoning. In: Keren, G.: A handbook for data analysis in the behavioral sciences 1, 2. Hillsdale, NJ: Erlbaum, 1993b (pp. 311-339)

[45] Gillies, D.A. (1971) A falsifying rule for probability statements. *British Journal of the Philosophy of Science*, 22, 231 – 261

[46] Gillies, D. A.: Philosophical Theories of Probability. London/New York – Routledge, 2000a

[47] Gillies, D.A. (2000b) Varieties of propensity. *British Journal of the Philosophy of Science*, 51, 807 – 835

[48] Glover, S., Dixon, P.(2004) Likelihood Ratios: a simple and flexible statistics for empirical psychologists. *Psychological Bulletin & Review* 11 (59; 795 – 806

[49] Good, I.J. (1971), 46656 varieties of Bayesians. In: Hamdan, M.A., Pratt, J.W., Gottlieb, P. Good, I.J., Hamdan, A., Carmer,S. G., Walker,W. M., Valentine,T. J., D'Agostino, R.B., Kshirsagar, A. M., Gwyn Evans, I., Kabe, D. G., Cureton, E. E., Rutherford, J. R., Sharma, J. K., Harvey, J. R.: Letters to the editor. *The American Statistician*, 25 (5), 56 – 63

[50] Goodman, S. N., Royall, R. (1988) Evidence and scientific research. *American Journal of Public Health*, 78 (12), 1568 – 1574

[51] Goodman, S. N. (1992) A comment on replication, p-values and evidence. *Statistics in Medicine*, 11, 875 – 879

[52] Goodman, S.N. (1993) p- values, hypothesis tests, and likelihood: Implicatgions for epidemiology of a neglected historical debate. *American Journal of Epidemiology* 137 (5), 485 – 495

[53] Goodman, S.N. (1999a) Towards evidence-based medical statistics. 1: The P value fallacy. *Annals of Internal Medicine*, 130 (12), 995 – 1004

[54] Goodman, S.N. (1999b) Towards evidence-based medical statistics. 2: The Bayes Factor. *Annals of Internal Medicine*, 130 (12), 1005 – 1013

[55] Hacking, I.: Logic of statistical inference. Cambridge, Cambridge University Press 1965/2009

[56] Hardin, C. L. (1966) The scientific work of reverend John Michell. *Annals of Science*, 22 (1), 27 - 47

[57] Held, L: Methoden der statistischen Inferenz – Likelihood und Bayes. Heidelberg, Spektrum Akademischer Verlag 2008

[58] Howson, C. (1987) Popper, prior probabilities, and inductive inference. *The British Journal for the Philosophy of Science*, 38 (2), 207 – 224

[59] Howson, C. (1997) A logic of induction. *Philosophy of Science*, 64, 268 – 290

[60] Howson, C.: Hume's Problem: Induction and the justification of belief. Oxford 2000

[61] Howson, C., Urbach, P.: Scientific Reasoning – The Bayesian Approach. Open Court, La Salle 1989

[62] Huberty, C. J. (1993) Historical origin of statistical testing practises: The treatment of Fisher versus Neyman-Pearson views in textbooks. *Journal of Experimental Education*, Summer, 317 – 333

[63] Hubbard, R., Bayarri, M. J.(2003) Confusion over measures of evidence (p's) versus errors (α's) in classical statistial testing. *The American Statistician* 57 (3), 171 – 182

[64] Hung, M.H., O'Neil, R.T., Bauer, P., Köhne, K. (1997) The behavior of the p-value when the alternative hypothesis is true. *Biometrics*, 53, 11 – 22

[65] Jaynes,E. T. (1968) Prior Probabilities. *IEEE Transactions on Systems, Science, and Cybernetics*, Vol 4 sec 4, 3, 227–241

[66] Jaynes, E. T.: Probability Theory. The Logic of Science. Cambridge University Press, Cambridge 2003

[67] Jefferys, W., Berger, J.O. (1991) Sharpening Ockham's razor by a Baysian strop. Technical Report, Department of Statistics, Purdue University, August 1991.

[68] Jeffrey, R. C. (1984) The impossibility of inductive probability. *Nature*,310 (2), 433

[69] Jeffreys, H.: Theory of probability. Oxford 1961/2003

[70] Joyce, J. (2003) Bayes' Theorem. *Stanfor Encyclopedia of Philosophy*, http://plato.stanford.edu/entries/Bayes-theorem

[71] Kass. R.E., Rafferty, A.E. (1995) Bayes Factors. *Journal of the American Statistical Association*,, 90 (430), 773 – 795

[72] Kass, R. E., Wasserman, L. (1996) The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, 91 (435), 1343 – 1370

[73] Kelly, T. (2006) Evidence. *The Stanford Encyclopedia of Philosophy*

[74] Kendall, M.G., Stuart, A.: The advanced theory of statistics, Vol. 21. London 1973

[75] Keynes, J. M.: A treatise on probability. 1921 – Digitized by Watchmaker Publishing 2008

[76] Krantz, D. H. (1999) The null hypothesis testing controversy in psychology. *Journal of the American Statistical Association*, 44, 1372 – 1381

[77] Kullback, S., Leibler, R. A. (1951) On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22 (1), 79 – 86

[78] Lambert, D., Hall, W.J. (1982) Asymptotic lognormality of P values. *The Annals of Statistics*, 10 (1), 44 – 64

[79] Laplace, P. S. (1774) Mémoire sur la Probabilité des Causes par les Événements. *Mémoires des Mathématiques et de Physique Presentés à l'Académie Royale des Sciences,Par Divers Savans,& Lûs dans ses Assemblées*, 6: 621 – 656

[80] Lehmann, E.L.: Testing statistical hypotheses. Second Edition, Chapman & Hall, New York London 1994

[81] Lenhard, J. (2006) Models and statistical inference: the controversy between Fisher and Neyman-Pearson. *The British Journal of Philosophy of Science*, 57, 69–91

[82] Levi, I. (1984) The impossibility of inductive probability. *Nature*, 310 (2), 433

[83] Lindley, D. V. (1956) On a Measure of the Information Provided by an Experiment. *The Annals of Mathematical Statistics*, 27(4), 986 – 1005

[84] Lindgren, B. W.: Statistical Theory. New York 1962

[85] Marbe, K.: Naturphilosophische Untersuchungen zur Wahrscheinlichkeitslehre. Leipzig 1899

[86] Marbe, K.: Die Gleoichförmigkeit in der Welt. München 1916

[87] Michell, J. (1767) An Inquiry into the probable Parallax, and Magnitude, of the Fixed Stars, from the Quantity if Lightthey afford us, and the particular Circumstances of the Situation. *Philosophical Transactions*, v 57, 234 – 264

[88] Moosbrugger, H., Brandl, Y.: Signifikanztest, Effektgrößenbestimmung und optimale Stichprobenumfänge. Arbeiten aus dem Institut für Psychologie der Johann Wolfgang Goethe-Universität,Heft 1, 2002

[89] Nagel, E. (1936). The Meaning of Probability (with discussion). *Journal of the American Statistical Association*, 31, 10 – 30

[90] Neyman, J., Pearson, E. S. (1928) On the use and interpretation of certain test criteria for purposes of statistical inference: Part I *Biometrika*, 20A, 1/2, 175– 240

[91] Neyman, J., Pearson, E. S. (1933) On the problem of the most efficient test. *Biometrika*, 231, 289–337

[92] Neyman, J.: First course in probability and statistics. New York Henry Holt and Company 1959

[93] Oakes, M.: Statistical inference: a commentary for the social and behavioral sciences. New York 1986

[94] Okasha, S. (2001) What did Hume really show about induction? *The Philsophical Quarterly*, 51, 307 – 327

[95] Ostmann, A., Wutke, J. (1994) Statistische Entscheidung. In: Herrmann, T., Tack, W. H. (Hrsg.) Enzyklopädie der Psychologie, Bd. I Methodische Grundlagen derPsychologie, Göttingen 1994, p. 694 – 737

[96] Passon, O. (2006) What you always wanted to know about Bohmian mechanics but were afraid to ask. *Physics and Philosophy*, ISSN: 1863–7388

[97] Peto, R., Pike, M.C., Armitage, P., Breslow, N.E., Cox, D. R., Howard, S. V., Mantel, N., McPherson, K., Peto, J., Smith, P.G. (1976) Design and analysis of randomized clinical trials requiring prolonged observation of each patient, I: Introduction and Design. *British Journal of Cancer*, 34, 585–612

[98] Popper, K. R.: Logik der Forschung. Wien 1934/Tübingen 1994

[99] Popper, K.R.: The propensity interpretation of the calculus of probability and the quantum theory. I; S. Körner (ed.) Observation and Interpretation. Proceedings of the Ninth Symposium of the Colston Research Society, University of Bristol, 65-79, 88-89

[100] Popper, K., Miller, D. (1983) On the impossibility of inductive probability. *Nature*, 302, 687 – 688

[101] Popper, K. R., Miller, D. W. (1987) Why probabilistic support is not inductive. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 321 (1562), 569 – 591

[102] Pratt, J. W. (1977) 'Decisions' as statistical evidence and Birnbaum's 'Confidence Concept'. *Synthese* 36, 59 – 69

[103] Pratt, J. W. (1961) Review of 'Testing statistical hypotheses' by E. L. Lehmann. *Journal of the American Statistical Association*, 56, 163 – 167

[104] Pruscha, H.: Vorlesungen über mathematische Statistik. Stuttgart 2000

[105] Redhead, M. (1985) On the impossibility of inductive probability. *British Journal for the Philosophy of Science*, 36, 185 – 191

[106] Reichenbach, H. (1935a) Wahrscheinlichkeitslehre. Leiden 1935

[107] Reichenbach, H. (1935b) Bemerkungen zu Karl Marbes statistischen Untersuchungen zur Wahrscheinlichchkeitsrechnung. *Erkenntnis*, 5, 305 – 322

[108] Robins, J.M., van der Waart, A., Ventura, V. (2000) Asymptotic distribution of P values in composite null models. *Journal of the American Statistical Association*, 95, 1143 – 1156

[109] Roth, E. (Hg.): Sozialwissenschaftliche Methoden. Lehr- und Handbuch für Forschung und Praxis. München, Wien 1993

[110] Royall, R. M. (1986) The effect of sample size on gthe meaning of significance tests. *The American Statistician*, 40 (4), 313–315

[111] Royall, Richard. M.: Statistical evidence: a likelihood paradigm. Chapman & Hall, London New York 1997

[112] Rozeboom, W.W. (1960) The fallacy ofthe null-hypothesis significance test. *Psychological Bulletin*, 57 (5), 416 – 428

[113] Salmon, W., (1981) Rational prediction. *British Journal of the *hilosophy of Science*, 32, 115 – 125

[114] Savage, L. J. (1961) The foundations of statistics reconsidered. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and PRobability*, 1, 583

[115] Savage, L.J.: The foundations of statistical inference. Methuen'Monographs on Applied Probability and Statistics. London 1962

[116] Schervish, M. J. (1996) P values: What they are and what they are not. *The American Statstician*, 50 (3), 203 – 206

[117] Sedlmeier, P. (1996) Jenseits des Signifikanztest-Rituals: Ergänzungen und Alternativen. *Methods of Psychological Research Online* 1 (4), 41 – 63

[118] Seidenfeld, T. (1979) Why I am not an objective Bayesian; some reflections prompted by Rosenkrantz. *Theorie and Decision*, 11, 413 – 440

[119] Sellke, T., Bayarri, M.J., Berger, J.O. (2001) Calibration of p-values for testing precise null hypotheses. *The American Statistician*, 55, 62 – 71

[120] Sober,E. (1994) Contrastive empriricism. In: From a biological point of view: Essays in evolutionary philosophy (pp. 114– 135), Cambridge, Cambridge University Press

[121] Sober, E. (2005) Is drift a serious alternative to natural selection as an explanation of complex adaptive traits? in: A. O'Hear (Ed.) Philosophy, biology, and life. Cambridge, Cambridge University Press

[122] Spielman, S. (1973) A Refutation of the Neyman-Pearson Theory of Testing. *The British Journal for the Philosophy of Science*, 24 (3), 201 – 222

[123] Stegmüller W.: Erklärung, Begündung, Kausalität. Springer-Verlag, Berlin, Heidelberg, New York 1983

[124] Sterling, T. D. (1959) Publication Decisions and their possible effects on inferences drawn from tests of significance – or vice versa. *Journal of the American Statistical Association*, 54, 30 – 34

[125] Stern, H. S. (2000) Asymptotic distribution of P values in composite null models: comment. *Journal of the American Statistical Association*, 95, 1157 – 1159

[126] Skékely, G. J.: Paradoxa – klassische und neue Überraschungen aus Wahrscheinlichkeitsrechung und mathematischer Statistik. Verlag Harri Deutsch, Thun, Frankfurt/Main 1990

[127] Talbot, W. (2008) Bayesian epistemology. *Stanfortd Encyclopedia of Philosophy*, http://plato.stanford.edu/entries/epistemology-bayesian/

[128] Tukey, J.W. (1960) Conclusions vs Decisions. *Technometrics*, 2(4), 423 – 433

[129] van Fraassen, B.C.: Laws and symmetry. Oxford 1989

[130] Vickers, J. (2010) The problem of induction. *Stanfortd Encyclopedia of Philosophy*, http://plato.stanford.edu/entries/induction-problem/

[131] Wagenmakers, E.J. (2007) A theoretical solution to the pervasive problems of p-vlalues, *Psychonomic Bulletin & Review*, 14 (5), 779 – 804

[132] Wald, A. (1939) Contributions to the theory of statistical estimation and testing hypotheses. *Annals of Mathematical Statistics*, 10, 299–362

[133] Wald, A.: Statistical Decision Functions. New York, Wiley 1950

[134] Weintraub, R. (1995) What was Hume's contribution to the problem of induction? *The Philosophical Quarterly*, 45 (181), 460 – 470

[135] Wittgenstein, L.: Tractatus logico-philosophicus. Logisch-philosophische Abhandlung. Edition Suhrkamp, Frankfurt 1963

[136] Zabell, S. L. (1989) The rule of succession. *Erkenntnis*, 31(2/3), 283 – 321

[137] Zhang, Z. (2009) A law of likelihood for composite hypotheses. *http://arxiv.org/abs/0901.0463*

Index